Universidade Federal de São Carlos
Centro de Ciências Exatas e Tecnologia
Programa de pós-graduação em Estatística

Título

Comparação do desempenho de Modelos Lineares Generalizados (MLG) e Modelos Aditivos Generalizados (MAG) na predição de dados financeiros em *credit score*

Lorene Guirado

São Carlos - SP
2010

Universidade Federal de São Carlos
Centro de Ciências Exatas e Tecnologia
Programa de pós-graduação em Estatística

Comparação do desempenho de Modelos Lineares Generalizados (MLG) e Modelos Aditivos Generalizados (MAG) na predição de dados financeiros em *credit score*

Lorene Guirado
Orientador: Prof. Dr. Francisco Louzada Neto
Co-Orientadora: Prof. Dra. Gleici da Silva Castro Perdoná

Dissertação apresentada ao programa de pós-graduação em Estatística da Ufscar como parte dos requisitos para a obtenção do título de Mestre em Estatística, área de concentração: Modelagem Estatística.

São Carlos - SP
2010

Universidade Federal de São Carlos
Centro de Ciências Exatas e Tecnologia
Programa de pós-graduação em Estatística

Comparison of Generalized Linear Models (GLM) and Generalized Additive Models (GAM) performances on financial credit scores predictions.

Lorene Guirado
Advisor: Prof. Dr. Francisco Louzada Neto
Co-advisor: Prof. Dra. Gleici da Silva Castro Perdoná

Dissertation submitted to the UFSCar' Statistics Post Graduation programme as a partial fulfillment to Statistics Master Title, area studies: Statistics Modelling.

São Carlos - SP
2010

Lorene Guirado

Comparação do desempenho de Modelos Lineares Generalizados (MLG) e Modelos Aditivos Generalizados (MAG) na predição de dados financeiros em *credit score*

Dissertação apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovada em 07 de outubro de 2010.

BANCA EXAMINADORA

Presidente _____
Prof. Dr. Francisco Louzada Neto (DEs-UFSCar/Orientador)

1° Examinador _____
Profa. Dra. Gleici da Silva Castro Perdoná (FMRP-USP/Co-Orientadora)

2° Examinador _____
Prof. Dr. Carlos Alberto Ribeiro Diniz (DEs-UFSCar)

3° Examinador _____
Prof. Dr. Ronaldo Dias (Imecc-UNICAMP)

Resumo: Esse trabalho teve como objetivo apresentar e comparar o desempenho de duas diferentes metodologias de modelagem estatística para dados financeiros com resposta dicotômica, especificamente exemplificadas pelos modelos de *credit score*, bem como metodologias para validação e análise de desempenho desses modelos. Uma das medidas que utilizamos nessa análise é o *lift*, muito utilizado no marketing, mas ainda pouco utilizado na área financeira, essa medida também é utilizada como técnica descritiva para categorização de variáveis.

As técnicas aqui apresentadas são os Modelos Lineares Generalizados (MLG), metodologia mais usual, e os Modelos Aditivos Generalizados (MAG), ainda pouco usual na área financeira por tratar-se de um modelo semi-paramétrico ou não-paramétrico, gerando ainda certa dificuldade de interpretação pelo fato de não apresentar parâmetros. As capacidades preditivas das duas técnicas são comparadas em uma aplicação em dados reais e em um estudo de simulação.

Abstract: This study aimed to present and compare the performance of two different methodologies for statistical modeling of financial data with dichotomous response, specifically exemplified by models of *credit score* as well as methodologies for validation and performance analysis of these models. One of the measures used in this analysis is the *lift*, often used in marketing, but little used in the financial area, this measure is also used as a descriptive technique for categorizing variables.

The techniques presented here are the Generalized Linear Models (GLM), the most usual method, and Generalized Additive Models (GAM), unusual in finance because it is a semi-parametric or nonparametric model, generating even some difficulty in interpretation because it does not present parameters. The predictive capabilities of the two techniques are compared in an application on real data and in a simulation study.

Lista de Abreviaturas

Y: Variável aleatória de interesse para a modelagem estatística

n: Número de observações

i: Indicador para a observação, i=1,2,...,n

p: Número de covariáveis do modelo

j: Indicador para a covariável, j=1,2,...,p

X: Variável aleatória independente

θ: Parâmetro de localização

ϕ: Parâmetro de dispersão

μ: Média de Y

σ: Desvio padrão de Y

m: Número de ensaios na distribuição Binomial no Capítulo 2

m: Número de replicações da amostra no Capítulo 4

π_i: Probabilidade de sucesso do i-ésimo indivíduo no caso da distribuição Binomial

ϕ^{-1}: Inverso da integral da distribuição Normal padrão

Se: Sensibilidade

Es: Especificidadde

Pc: Ponto de corte

K: Número de segmentos no Capítulo4

K: Número de nós no Capítulo 3

k: Indicador de variável no Capítulo 2

k: Indicador do nó no Capítulo 3

k: Indicador para o segmento, k=0,1,...,K-1 no Capítulo 4

s_i: Escore do i-ésimo indivíduo

B: Função Base B-*Spline*

ξ: Nó do polinômio por partes

r: Indicador para a covariavel, r=1,...,p

W: Matriz diagonal de pesos

s(.): Função *spline*

f(.): Função suave

g(.): Função de ligação

η(.): Preditor

Sumário

Capítulo 1

Introdução

Nas instituições financeiras os modelos de crédito têm sido utilizados para quantificar o risco que determinado cliente ou transação oferece durante as diferentes etapas do ciclo de crédito. Em geral, essas etapas são definidas pelo ato da solicitação, pela análise do comportamento e pelo estudo de cobrança, a cada uma dessas etapas esta relacionado um modelo de *scoring* de crédito que são, respectivamente, *credit score*, *behavior score* e *collection score* (Sabato, 2009). As técnicas de modelagem de *scoring* de crédito, baseiam-se na determinação de uma pontuação (*score*) para cada cliente, a qual pode ser interpretada como a probabilidade da ocorrência da inadimplência, e têm como uma de suas grandes vantagens a capacidade de auxiliar rapidamente a tomada de decisões e, muitas vezes, automatizá-las.

Os modelos de crédito consideram a variável binária que representa a ocorrência ou não da inadimplência (aqui utilizamos 1 para indicar a inadimplência e 0 para caso contrário) como variável resposta, ou variável de interesse, e outras variáveis independentes de acordo com a etapa do ciclo de crédito. Nos modelos de solicitação (*credit score*) são utilizadas informações cadastrais, que em geral são informações sócio-demográficas, obtidas no ato da solicitação como, por exemplo, sexo, idade, renda, situação residencial, entre outras. Já nos modelos de *behavior score* e de *collection score*, dado que os dados do cliente já estão na base da instituição há algum tempo, é possível utilizar informações comportamentais (referentes as suas transações) e de relacionamento. Estas informações adicionais trazem ao modelo uma melhor capacidade preditiva em relação aos modelos de

solicitação.

Nesta dissertação consideramos os modelos aplicados à concessão de crédito que apesar de ser uma das principais fontes de receita dos bancos e instituições financeiras em geral (Abreu, 2005, p. 1), ainda apresentam modelos com baixa capacidade de previsão se comparada à análise comportamental e ao estudo de cobrança (Sabato, 2009). Devido a esta relativa baixa capacidade preditiva, diversos modelos de previsão já foram aplicados a problemas de concessão de crédito visando sempre a comparação da capacidade de previsão dos mesmos. Dois exemplos são: Gonçalves (2005) que compara os modelos de Regressão Logística, Redes Neurais e Algoritmo Genético e Mendonça (2008) que compara Regressão Logística Clássica, Regressão Logística Bayesiana e Redes Neurais.

Segundo Sabato (2009), no final da década de 50 a análise de crédito baseava-se na análise univariada, na década de 60 houve a introdução da análise discriminante multivariada neste campo pelos trabalhos seminais de Beaver (1967), seguido por Altman (1968), que se tornou a metodologia estatística mais popular na modelagem de crédito. Em 1980, Ohlson aplicou o modelo logit condicional ao estudo da previsão de inadimplência. Desde então diversas técnicas foram desenvolvidas visando aumentar o poder de previsão dos modelos de crédito tais como, Regressão Probit, Regressão Logística Bayesiana e Redes Neurais. No entanto, segundo Sabato (2009), os resultados empíricos jamais demonstraram benefícios realmente significativos, considerando ainda a Regressão Logística o método mais popular. Segundo Liu (2007), apesar de suas vantagens a Regressão Logística recebe críticas quanto a falta de suporte às estruturas não lineares do efeito dos preditores sobre a variável dependente. Dessa forma, quando o efeito de uma variável preditora contínua não apresenta estrutura linear sobre a variável resposta, usamos de categorizá-la. Gruenstein (1998) menciona que a transformação das variáveis contínuas em discretas, principalmente nos casos de estruturas não lineares de relacionamento, pode trazer ganhos no poder preditivo do modelo, no entanto a utilização da categorização de variáveis contínuas pode implicar numa perda considerável de informação.

A Regressão Logística é um caso particular dos Modelos Lineares Generalizados (Nelder e Wedderburn, 1972), estes modelos apresentam uma estrutura unificada para todos os modelos cuja variável resposta seja membro da família exponencial, tais como a Gaussiana, Binomial, Gama e Poisson, estes modelos assumem que a relação entre a

resposta e o efeito dos preditores é linear, o que nem sempre ocorre.

Em 1990, Hastie e Tibshirani, propuseram o Modelo Aditivo Generalizado (MAG) que "relaxa a suposição de linearidade do Modelo Linear Generalizado e assume que a variável resposta é dependente dos preditores univariados suavizados e não dos preditores em si " (Liu, 2007, p. 2). Este modelo, por ser uma extensão dos Modelos Lineares Generalizado (MLG), é aplicável a todos os casos de modelagem nos quais queremos fazer previsões para uma variável cuja distribuição pertença a família exponencial.

Os Modelos Aditivos Generalizados ajustam funções suaves a cada variável do modelo e nesta dissertação as funções suaves utilizadas foram as *splines* cúbicas, que são uma forma de ajuste dada por uma série de polinômios por partes definidos em sub-intervalos cujos extremos são denominados nós. Mais detalhes abordaremos na seção 3.3.

Este trabalho teve como principal objetivo comparar o desempenho dos Modelos Lineares Generalizados com os Modelos Aditivos Generalizados para o caso da distribuição binomial. A comparação do desempenho dos modelos foi feita em termos de medidas de precisão da predição, robustez e em concordância entre os valores preditos nos diferentes métodos de análise utilizados (MLG e MAG).

Várias são as técnicas utilizadas para garantir que um modelo de regressão apresenta um bom desempenho em uma base de dados que não tenha sido utilizada para construir o modelo, caso isto ocorra, temos o indicativo de que o modelo pode ser aplicado a novos exemplos do mesmo domínio de dados. Dessa forma é usual na modelagem estatística, separar aleatoriamente a base de dados disponível em duas partes mutuamente excludentes, uma para a construção e outra para a validação do modelo (Picard, 1990).

Para tanto nesta dissertação, primeiramente utilizamos algumas medidas conhecidas, como a estatística de Kolmogorov-Smirnov (seção 4.1) e a área sob a curva ROC (seção 4.2), e algumas técnicas não tão usuais na área financeira, como a análise de Decis (seção 4.3) e a Curva *Lift* (*Lift Chart*) (seção 4.3.1). A Análise de Decis baseia-se na segmentação da base de dados ordenada pelos valores preditos pelo modelo (*scores*) em 10 partes iguais denominadas decis ou faixas de *score*, a análise é feita em cada um desses segmentos e baseia-se numa tabela denominada Tabela de Análise de Decis, que contém para cada uma das 10 faixas de *score*, o *score* médio e a proporção ou taxa real de eventos.

O *lift*, medida que gera a *Curva Lift* (citada anteriormente), será introduzida no contexto de *credit score*. Essa medida é muito usual no marketing (S.Coppock (2002)), mas ainda é pouco usual na área financeira, que em geral utiliza medidas como o KS, Sensibilidade, Especificidade, entre outras que discutiremos ao longo do texto. Veremos que o *lift* pode ser utilizado tanto na análise de desempenho, como pode ser uma técnica descritiva para a categorização e recategorização de variáveis.

Segundo a Wikipedia, a enciclopédia livre, o *lift* mede o desempenho de um modelo de segmentação de uma população, aqui a segmentação foi definida pelos decis, sendo que o *lift* de um subconjunto da população é dado pela razão entre a taxa de resposta (evento) real do subconjunto em questão em relação a taxa de resposta real de toda a população. O *Lift Chart* é um gráfico de *Lift* $\times$ Segmento, no nosso caso o decil, este gráfico nos permite dizer se o modelo esta sendo capaz de ordenar o evento de interesse e também o quanto esta capacidade é superior à abordagem aleatória (a não utilização de qualquer ferramenta preditiva).

Para analisar a capacidade preditiva do modelo de classificação, os modelos estatísticos após a aplicação do ponto de corte buscaram avaliar a qualidade da classificação binária utilizando as medidas de desempenho (seção 4.4), as quais se baseiam uma tabela contendo as freqüências cruzadas entre a variável de resultados reais e a variável correspondente aos resultados do modelo de classificação, esta tabela é conhecida como matriz de confusão. Utilizamos as medidas de desempenho mais conhecidas, como a Sensibilidade, a Especificidade, Valores Preditivos (positivo e negativo) e a Capacidade de Acerto Total, além uma medida não tão conhecida, que é o Coeficiente de Correlação de Matthew, que segundo a enciclopédia livre, Wikipedia, geralmente é considerada a medida que melhor consegue descrever a matriz de confusão através de um único número.

Para avaliar a robustez do modelo comparamos os resultados, de cada uma das técnicas utilizadas para avaliar a capacidade preditiva do modelo de regressão e do modelo de classificação, da amostra desenvolvimento com os resultados da amostra validação, e fizemos a análise dos intervalos de confiança para a medida *lift* (ver seção 4.3.2).

Portanto nesta dissertação, apresentamos a técnica de modelagem de dados binários, os Modelos Lineares Generalizados (Nelder & Wedderburn, 1972) cuja Regressão Logística é um caso particular (Capítulo 2). No Capítulo 3, apresentamos os Modelos

Aditivos Generalizados. Já no Capítulo 4 apresentamos as técnicas utilizadas para a avaliação e validação de modelos binários: Estatística KS, Curva Roc, Análise de Decis, *lift* e Medidas de Desempenho em Modelos de Classificação.

No Capítulo 5, foi realizado um estudo de simulação, onde foram geradas, segundo Breiman (1998), amostras contendo duas variáveis preditoras de origem contínua e uma variável resposta, binária. Foram simuladas 50 amostras para cada uma das proporções de evento na variável resposta: 1%, 10%, 25% e 50%, somando um total de 200 amostras. Foram ajustados modelos de Regressão Logísticas e Modelos Aditivos Generalizados em todos os cenários e posteriormente os resultados foram comparados. Todo o estudo foi feito utilizando o software SAS.

No Capítulo 6, apresentamos uma aplicação real de uma base disponibilizada por uma instituição financeira que atua no mercado varejista brasileiro a mais de vinte anos. Esta base corresponde a fase de concessão no ciclo de crédito, tratando-se assim de um modelo de *credit score*. Foram propostas duas metodologias para o ajuste desse modelo: a Regressão Logística (MLG) e o MAG para dados binários, todo o estudo foi feito utilizando o software SAS. E finalizamos no Capítulo 7 com conclusões e propostas futuras.

Capítulo 2

Modelos Lineares Generalizados (MLG)

2.1 Introdução

Modelos de regressão buscam elaborar uma função que relacione, em termos médios, a variação de variável dependente (Y) com a variação de variáveis independentes (X), sendo que a parcela de variação da variável dependente não explicada pela variação das variáveis independentes é atribuída a uma variável aleatória denominada erro. Dada esta função temos que a partir de valores das variáveis independentes, ou covariáveis, podemos estimar o valor da variável de interesse, ou variável resposta. As formas mais simples de modelos de regressão envolvem uma dependência linear entre a resposta e as covariáveis e ainda supõe que o erro tenha distribuição normal em torno de zero, que são conhecidos como Modelos de Regressão Linear.

No caso dos Modelos de Regressão Linear, a variável resposta é dada por uma combinação linear das covariáveis adicionada por um erro cuja distribuição supomos ser normal, temos então que tal modelo requer que a variável resposta também siga uma distribuição normal, no entanto, na prática isso muitas vezes não ocorre. Em 1972 uma extensão do modelo de regressão linear foi proposta por Nelder e Wedderburn, os Modelos Lineares Generalizados (MLG), que admitem a não normalidade nos erros e, portanto, na variável resposta. Estes modelos podem ser utilizados quando a distribuição da variável de interesse é qualquer distribuição da família exponencial tais como a Gaussiana, Binomial, Gama e Poisson. Neste Capítulo apresentamos esses modelos de uma forma geral e

também o caso particular de resposta binária.

Na seção 2.2 apresentamos o Modelo Linear Generalizado e suas componentes. Na Seção 2.3 introduzimos a família de distribuições considerada nos Modelos Lineares Generalizados, a família exponencial. Na Seção 2.4, a estimação dos parâmetros do Modelo Linear Generalizado e na Seção 2.5 os testes de hipótese mais utilizados para tal modelo.

Posteriormente, na Seção 2.6, o caso de resposta binária e as funções de ligação, que é a função que faz a ligação entre a média da variável resposta e o preditor linear (vide Seção 2.2), mais utilizadas para esse caso. Para o caso da função de ligação canônica (logito, para o caso binário) apresentamos o processo de estimação e a interpretação dos parâmetros.

2.2 Definição do modelo

Suponha que temos uma única variável aleatória Y com distribuição de probabilidade $f(.)$ a qual acreditamos estar associada um conjunto de variáveis explicativas $X_1, X_2, ..., X_p$, suponha também que temos uma amostra contendo n observações sendo que para cada observação temos $(y_i, \mathbf{x}_i)$, onde $\mathbf{x}_i$ é o vetor coluna contendo as variáveis explicativas $\mathbf{x}_i = (x_{1i}, x_{2i}, ..., x_{pi})^T$. Em sua forma mais simples temos que os modelos de regressão linear são dados por:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \epsilon_i \tag{2.1}$$

onde y_i vem de uma distribuição normal, ϵ_i é o erro da i-ésima observação sendo uma ocorrência da distribuição normal com média zero e variância σ^2 e os erros são independentes das covariáveis $X_j's$.

A generalização desde modelo, proposta em 1990 por Hastie e Tibshirani é baseada na extensão da distribuição associada ao modelo para toda distribuição da familia exponencial. O Modelo Linear Generalizado envolve três componentes:

- Componente Aleatório: Um conjunto das variáveis aleatórias $Y_1, Y_2, ..., Y_n$ as quais pertencem a uma mesma distribuição da família exponencial, cujas médias são $\mu_1, \mu_2, ...\mu_n$.

- Componente Sistemático: As variáveis explicativas entram no modelo como uma soma linear de seus efeitos através de um preditor linear (η)

$$\eta_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}. \tag{2.2}$$

- Função de Ligação: Para relacionar a variável resposta com a função linear dos parâmetros, é uma função que liga o componente sistemático ao componente aleatório, aqui denominada $g(\mu_i)$

$$\eta_i = g(\mu_i), \tag{2.3}$$

onde $g(.)$ deve ser uma função monótona derivável. Esta função define a forma com que os efeitos sistemáticos $x_1, x_2, ..., x_p$ são transmitidos para a média da variável resposta. Caso a função de ligação escolhida resulte em $g(\mu_i) = \theta_i$, onde θ_i é o parâmetro de localização da família exponencial também denominado parâmetro canônico (vide seção 2.3), esta função é chamada de função de ligação canônica. Neste caso o preditor linear modela diretamente o parâmetro canônico, o que frequentemente resulta em uma interpretação prática para os parâmetros da regressão bem como numa maior facilidade computacional.

Vemos aqui que o parâmetro θ_i da distribuição pertencente a família exponencial não são de interesse direto, mas sim os parâmetros $\beta_0, \beta_1, ...\beta_p$ tais que uma combinação linear desses β's seja uma função da esperança de Y_i.

Segundo Demétrio (2002), a escolha da distribuição da variável resposta, a matriz do modelo (covariáveis) e a função de ligação são as escolhas mais importantes na utilização do Modelo Linear Generalizado.

2.3 Família Exponencial

Seja Y uma variável aleatória, então Y pertence a família exponencial de distribuições se a distribuição de Y puder ser escrita da seguinte forma, pela função

$$f(y; \theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}, \tag{2.4}$$

onde $a(\phi), b(\theta)$ e $c(y, \phi)$ são funções específicas para cada distribuição e serão apresentadas adiante. O parâmetro θ é o parâmetro de localização, também denominado parâmetro canônico, e ϕ é um parâmetro de dispersão, $\phi > 0$ e conhecido, geralmente denominado σ^2. De acordo com a função de densidade probabilidade (2.4) a média e a variância da variável resposta Y são dadas respectivamente por,

$$E(Y) = \mu = \frac{db(\theta)}{d\theta} \tag{2.5}$$

$$Var(Y) = \frac{d^2b(\theta)}{d\theta^2}a(\phi) = a(\phi)V(\mu_i). \tag{2.6}$$

Alguns exemplos de distribuições que podem ser expressas na forma (2.4) são as distribuições Normal, Poisson e Binomial. Como a distribuição Normal faz parte das distribuições da família exponencial, os MLG's são uma extensão dos Modelo de Regressão Linear usual. Apresentamos abaixo as funções específicas, em ordem, de cada uma destas distribuições citadas acima:

- Distribuição normal

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{-(y-\mu)^2}{2\sigma^2}\right] \tag{2.7}$$

$$= exp\left[\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\right]. \tag{2.8}$$

Desta forma temos que $\theta = \mu; b(\theta) = \frac{\mu^2}{2}, a(\phi) = \phi; \phi = \sigma^2$ e $c(y, \phi) = -\frac{1}{2}\left[ln(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right]$.

A média e a variância de y são:

$$E(y) = \frac{db(\theta)}{d\theta} = \mu; \qquad Var(y) = \frac{d^2b(\theta)}{d\theta^2}a(\phi) = \sigma^2. \tag{2.9}$$

- Distribuição poisson

$$f(y; \mu) = \frac{\mu^y e^{-\mu}}{y!} \tag{2.10}$$

$$= exp\left[yln(\mu) - \mu - ln(y!)\right]. \tag{2.11}$$

Desta forma $\theta = ln\mu; b(\theta) = -\mu; \alpha(\phi) = \phi; \phi = 1$ e $c(y, \phi) = -ln(y!)$.

A média e a variância de y são:

$$E(y) = \frac{db(\theta)}{d\theta} = \mu; \qquad Var(y) = \frac{d^2b(\theta)}{d\theta^2}a(\phi) = \mu. \tag{2.12}$$

- Distribuição binomial

$$\begin{aligned} f(y;\mu,m) &= \binom{m}{y}\mu^y(1-\mu)^{m-y} && (2.13)\\ &= \binom{m}{y}exp\left[y\ln(\mu)+(m-y)\ln(1-\mu)\right] && (2.14)\\ &= \binom{m}{y}exp\left[y\ln(\mu)+m\ln(1-\mu)-y\ln(1-\mu)\right] && (2.15)\\ &= \binom{m}{y}exp\left[y\ln\left(\frac{\mu}{1-\mu}\right)+m\ln(1-\mu)\right] && (2.16)\\ &= exp\left[y\ln\left(\frac{\mu}{1-\mu}\right)+m\ln(1-\mu)+\ln\binom{m}{y}\right]. && (2.17) \end{aligned}$$

Desta forma $\theta = \ln\left(\frac{\mu}{1-\mu}\right)$, como $\mu = \frac{e^\theta}{1+e^\theta}, b(\theta) = -m\ln(1-\mu) = m\ln(1+e^\theta)$, $\alpha(\phi)=\phi; \phi = 1$ e $c(y,\phi) = \ln\binom{m}{y}$.

A média e a variância de y são:

$$E(y) = \frac{db(\theta)}{d\theta} = \mu; \qquad Var(y) = \frac{d^2b(\theta)}{d\theta^2}a(\phi) = \mu(1-\mu). \tag{2.18}$$

2.4 Estimação dos parâmetros

Seja $Y = (y_1, y_2, ..., y_n)$, $L(Y;\theta,\phi)$ a função de verossimilhança, a distribuição conjunta dos y's, dada por

$$L(Y;\theta,\phi) = \prod_{i=1}^{n} f(y_i;\theta,\phi). \tag{2.19}$$

Então a estimação dos parâmetros é feita através da maximização do logaritmo da função de verossimilhança

$$l(Y;\theta,\phi) = \ln L(y;\theta,\phi) = \sum_{i=1}^{n}\ln f(y_i;\theta,\phi) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right] \tag{2.20}$$

derivando o logaritmo da função de verossimilhança em relação ao parâmetro β_j,para $j = 1, ..., p$ temos

$$\frac{dl(\cdot)}{d\beta_j} = \sum_{i=1}^{n}\frac{dl}{d\theta_i}\frac{d\theta_i}{d\eta_i}\frac{d\eta_i}{d\beta_j}. \tag{2.21}$$

Como

$$\frac{dl(\cdot)}{d\theta_i} = \sum_{i=1}^{n}\frac{1}{a(\phi)}\left[y_i - \frac{db(\theta_i)}{d\theta_i}\right] = \sum_{i=1}^{n}\frac{1}{a(\phi)}\left[y_i - \mu_i\right] \tag{2.22}$$

e $\frac{d\eta_i}{d\beta_j} = x_{ij}$ e $a(\phi)$ constante, então

$$\frac{dl(\cdot)}{d\beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} \frac{d\theta_i}{d\eta_i} x_{ij}. \tag{2.23}$$

Igualando a equação acima a zero temos as equações *score*

$$\sum_{i=1}^{n} (y_i - \mu_i) \frac{d\theta_i}{d\eta_i} x_{ij} = 0. \tag{2.24}$$

Resolvendo as equações escore obtemos as estimativas dos parâmetros no entanto, como estas equações são não-lineares, há a necessidade da aplicação de métodos iterativos para a obtenção das estimativas como, por exemplo, o método iterativo de Newton-Rapson. Vale ressaltar que uma propriedade importante da família exponencial é que o máximo global do log da verossimilhaça é dado unicamente pela solução do sistema de equações $U_\theta = \frac{dl(\cdot)}{d\theta} = 0$. McCullagh e Nelder (1989) mostram que a solução para tais equações podem ser obtidas usando o procedimento iterativo de Mínimos Quadrados Ponderados.

Seja U_θ a solução do sistema de equações $U_\beta = \frac{dl(\cdot)}{d\beta} = 0$ utilizando a aproximação de Taylor para a função $U(\beta)$ nas vizinhanças do ponto β_0, temos

$$U(\beta) = U(\beta_0) + (\beta_0 - \beta)U'(\beta_0) = 0 \tag{2.25}$$

resultando em

$$\begin{aligned} \beta &= \beta_0 - \frac{U(\beta_0)}{U'(\beta_0)} && (2.26) \\ &= \beta_0 - U(\beta_0)[U'(\beta_0)]^{-1}. && (2.27) \end{aligned}$$

Seja m o m-ésimmo passo para a obtenção de $\hat{\beta}$ temos

$$\begin{aligned} \beta^{(m+1)} &= \beta^{(m)} - U(\beta^{(m)})[U'(\beta^{(m)})]^{-1} && (2.28) \\ \beta^{(m+1)} &= \beta^{(m)} + U(\beta^{(m)})[-U'(\beta^{(m)})]^{-1}. && (2.29) \end{aligned}$$

Como $U_\beta = \frac{dl}{d\beta} = 0$

$$\begin{aligned} \beta^{(m+1)} &= \beta^{(m)} - U(\beta^{(m)})[U'(\beta^{(m)})]^{-1} && (2.30) \\ \beta^{(m+1)} &= \beta^{(m)} + \left[\frac{dl}{d\beta}\right]_{\beta_m} \left[\frac{d^2l}{d\beta^2}\right]^{-1}_{\beta_m}. && (2.31) \end{aligned}$$

Cosideremos $k = 1, ..., p$, para o caso multivariado temos

$$\beta^{(m+1)} = \beta^{(m)} + \left[\frac{dl}{d\beta}\right]_{\beta_m} \left[-\frac{d^2l}{d\beta_j d\beta_k}\right]^{-1}_{\beta_m} \tag{2.32}$$

$$\beta^{(m+1)} = \beta^{(m)} + \left[\frac{dl}{d\beta}\right]_{\beta_m} [I]^{-1}_{\beta_m} \tag{2.33}$$

$$\tag{2.34}$$

onde I é denominada matriz de informação observada.

Como nem sempre a solução de $\left[\frac{d^2l}{d\beta_j d\beta_k}\right]_{\beta_m}$ é simples, algumas vezes as derivadas de segunda ordem não são obtidas facilmente, substitui-se a matriz de derivadas parciais de 2^a ordem pela matriz de valores esperados das derivadas parciais, matriz de informação esperada de Fisher ($\Im$), cujos elementos são dados por $\Im_{jk} = E\left[-\frac{d^2l}{d\beta_j d\beta_k}\right]$, que é a matriz de covariâncias dos $U'_j s$. Temos então

$$\beta^{(m+1)} = \beta^{(m)} + \left[\frac{dl}{d\beta}\right]_{\beta_m} E\left[-\frac{d^2l}{d\beta_j d\beta_k}\right]^{-1}_{\beta_m} \tag{2.35}$$

$$\beta^{(m+1)} = \beta^{(m)} + \left[\frac{dl}{d\beta}\right]_{\beta_m} \left[\Im^{(m)}\right]^{-1}. \tag{2.36}$$

Multiplicando $\Im_{jk}$ na equação temos

$$\Im_{jk}\beta^{(m+1)} = \Im_{jk}\beta^{(m)} + U_j^{(m)}. \tag{2.37}$$

Pode ser provado que

$$\Im_{jk} = \frac{1}{\phi} X^T W X \tag{2.38}$$

onde, $W = diag\{W_1, W_2, ..., W_n\}$, $W_i = \frac{w_i}{V(\mu_i)}\left[\frac{d\mu_i}{d\eta_i}\right]^2$ e $w_i = \frac{\phi}{a_i(\phi)}$.

E também que

$$U_j = \frac{1}{\phi} X^T W \Delta (y - \mu) \tag{2.39}$$

onde $\Delta = diag\left\{\frac{d\eta_1}{d\mu_1}, \frac{d\eta_2}{d\mu_2}, ..., \frac{d\eta_n}{\mu_n}\right\} = diag\{g'(\mu_1), g'(\mu_2), ..., g'(\mu_n)\}$.

Substituindo na equação 2.37 temos

$$X^T W^{(m)} X \beta^{(m+1)} = X^T W^{(m)} Z^{(m)} \tag{2.40}$$

onde $Z = \eta^{(m)} + \Delta(y - \mu)^{(m)}$, ou seja

$$\beta^{(m+1)} = \left(X^T W^{(m)} X\right)^{-1} X^T W^{(m)} Z^{(m)} \tag{2.41}$$

Note que o valor estimado para β será dado pela estimativa de Mínimos Quadrados Ponderados Iterativamente a partir de uma estimativa inicial $\beta^{(0)}$ que atualiza os valores de $\beta^{(m+1)}$ até que a convergência seja obtida e, portanto obtemos $\hat{\beta} = \beta^{(m+1)}$. Geralmente o vetor inicial $\beta^{(0)}$ é calculado utilizando-se $\hat{\eta}$ como variável dependente e X a matriz do modelo, onde $\hat{\eta}_i = g(\hat{\mu}_i) = g(y_i)$, considerando cada observação como a estimativa de seu valor médio $\hat{\mu}_i = y_i$.

O algoritmo de estimação de Mínimos Quadrados Ponderados Iterativamente é dado por:

1. Obter as estimativas $\eta_i^{(m)} = \sum_{j=1}^{p} x_{ij}\beta_j^{(m)}$ e $\mu_i^{(m)} = g^{-1}(\eta_i^{(m)})$;
2. Obter a variável dependente ajustada $z_i^{(m)} = \eta_i^{(m)} + (y_i - \mu_i^{(m)})g'(\mu_i^{(m)})$ e os pesos $W_i^{(m)} = \frac{w_i}{V(\mu_i^{(m)})}\left[g'(\mu_i^{(m)})\right]^2$;
3. Calcular $\beta^{(m+1)} = \left(X^T W(m) X\right)^{-1} X^T W^{(m)} z^{(m)}$, voltar ao passo (1) com $\beta^{(m)} = \beta^{(m+1)}$ até que $\beta^{(m+1)}$ pare por um critério de convergência. Existem muitos critérios para a verificação da convergência, no entanto como os mesmos vão além do objetivo deste projeto não serão abordados aqui, detalhes sobre a convergencia podem ser vistos em Demétrio (2002).

2.5 Teste de significância dos parâmetros

Para testar a significância dos coeficientes, McCullagh e Nelder (1989) recomendam usar a função *deviance* (desvio), que mede, para a base de dados em questão, a discrepância entre o modelo saturado e o modelo em estudo. O modelo saturado é o modelo para o qual os valores ajustados $\hat{\mu}_i$ são iguais às respostas observadas, o que ocorre quando o número de parâmetros do modelo é o número de observações da base (ou seja, $p = n$). A *deviance* é definida por,

$$D(Y;\hat{\mu}) = -2\ln\left[\frac{L(Y,\hat{\mu})}{L(Y,Y)}\right] = -2\left[\ln L(Y,\hat{\mu}) - \ln L(Y,Y)\right] \tag{2.42}$$

onde $L(Y,\hat{\mu})$ é a função de máxima verossimilhança do modelo em estudo e $L(Y,Y)$ é a função de máxima verossimilhança do modelo saturado. Note que a função desvio é uma distância entre o logaritmo do máximo da função de verossimilhança do modelo saturado, contendo n parâmetros, e do modelo sob investigação, contendo p parâmetros. Um valor pequeno para a função desvio indica que, para um número menor de parâmetros ($p < n$), obtém-se um ajuste tão bom quanto o ajuste com o modelo saturado.

Observe que no caso da distribuição normal a *deviance* é a soma de quadrados dos resíduos. De acordo com Lindsey (1997) $D(Y,\hat{\mu})$ segue, assintoticamente, uma distribuição χ^2 com $(n-p)$ graus de liberdade. O teste de significância (teste da razão de verossimilhança) é baseado na diferença de *deviances* de dois modelos aninhados, um contendo p parâmetros e outro contendo $(p+q)$ parâmetros. Suponha o modelo 1 sendo o modelo saturado, o modelo 2 o modelo contendo $(p+q)$ parâmetros e o modelo 3 sendo o modelo aninhado ao modelo 2 contendo p parâmetros, temos,

$$\begin{aligned} D_3(Y,\hat{\mu}) - D_2(Y,\hat{\mu}) &= -2\left[\ln L_3(Y,\hat{\mu})\right] + 2\left[\ln L_2(Y,\hat{\mu})\right] && (2.43)\\ &= -2\ln\left[\frac{L_3(Y,\hat{\mu})}{L_2(Y,\hat{\mu})}\right] && (2.44)\end{aligned}$$

Temos que sob a hipótese de que o modelo 3 é correto a diferença acima segue distribuição χ^2_q.

2.6 Resposta Dicotômica

Quando a variável de interesse é binária, como por exemplo, ocorrência ou não de inadimplência, assumimos que a variável tem distribuição de probabilidade Binomial (2.13) com parâmetros (μ, m). Considerando que cada observação y_i é o resultado de um ensaio de Bernoulli, temos que $(m_i = 1)$, neste caso μ_i é uma probabilidade a qual denominaremos π_i.

2.6.1 Funções de Ligação

As funções de ligação mais comumente utilizada nos casos de resposta dicotômica são a função de ligação logito, probito e complemento log-log. Seja π_i a probabilidade de sucesso

(ocorrência do evento) para o i-ésimo indivíduo, então as funções de ligação citadas acima são dadas por:

- Função logito

$$g_1(\pi_i) = ln\left[\frac{\pi_i}{1-\pi_i}\right], \tag{2.45}$$

- Função probito

$$g_2(\pi_i) = \Phi^{-1}(\pi_i), \tag{2.46}$$

- Função complemento log-log

$$g_3(\pi_i) = ln\left[-ln(1-\pi_i)\right]. \tag{2.47}$$

onde $\Phi^{-1}(\pi_i)$ é o inverso da integral da normal padrão, a função logística é o inverso de uma função de distribuição acumulada simétrica e a função complemento log-log é a inversa da distribuição acumulada assimétrica. A função complemento log-log é limitada a situações onde a probabilidade de sucesso é assimétrica. Já as transformações logito e probito são semelhantes, sendo que a função logito é mais conveniente do ponto de vista computacional.

Temos que a função logística é a função de ligação canônica para a distribuição binomial, $\eta_i = \theta_i = log\left(\frac{\pi_i}{1-\pi_i}\right)$ e devido as facilidades das funções de ligação canônicas e do bom desempenho da função logística, esta função será escolhida como função de ligação padrão para dados binários neste trabalho e o MLG para este caso pode ser denominado Modelo de Regressão Logística.

Segundo Myers e Montgomery (2002) a utilização da função de ligação canônica implica em algumas propriedades interessantes. A sua escolha é conveniente porque não só simplifica as estimativas de máxima verossimilhança dos parâmetros do modelo, mas também o cálculo do intervalo de confiança para a média da resposta.

2.6.2 Estimação dos Parâmetros - Ligação Canônica

Para a função de ligação logística temos o logaritmo da função de verossimilhança:

$$L(Y,\pi) = \ln l(Y;\pi) = \sum_{i=1}^{n}\left[y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\right] \tag{2.48}$$

Temos que as equações escore são dadas por:

$$\sum_{i=1}^{n}(y_i - \mu_i)\frac{d\theta_i}{d\eta_i}x_i = 0. \tag{2.49}$$

Como $\eta_i = \theta_i$, $\frac{d\theta_i}{d\eta_i} = 1$ então as equações escore serão dadas por:

$$\sum_{i=1}^{n}(y_i - \pi_i)x_i = 0. \tag{2.50}$$

Podemos reescrever montando os parâmetros de forma iterativa.

2.6.3 Interpretação dos parâmetros - Ligação Canônica

Considerando (2.2) e a função de ligação logito para o caso binário temos o seguinte modelo de regressão $ln\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$, neste caso o coeficiente de uma variável independente representa a taxa de mudança ou a inclinação da função logito para cada incremento unitário na unidade de medida de tal variável, considerando fixas as demais covariáveis do modelo. Por exemplo, se $\beta_1 = 3,04$ temos que a cada unidade de medida incrementada em X_1 temos um acréscimo de 3,04 na inclinação do logito da média.

Para variáveis categóricas, a forma mais utilizada para quantificar a influência das variáveis independentes na variável resposta é a razão de chances (*Odds Ratio*), que nos permite dizer quantas vezes a ocorrência do evento é mais provável para uma categoria da variável x_1, por exemplo, do que para outra. Suponha que uma variável X tenha duas possíveis categorias c1 e c2, então a razão de chances é calculada da seguinte forma:

$$\psi = \frac{\frac{\pi_i(c1)}{1-\pi_i(c1)}}{\frac{\pi_i(c2)}{1-\pi_i(c2)}} \tag{2.51}$$

Neste caso o valor ψ nos indicaria quantas vezes é mais provável a ocorrência do evento na ocorrência da categoria 1 da covariável X_1, em relação à categoria 2 da covariável X_1. Caso a razão de chances seja igual a 1, então a variável em questão não contribui na explicação da variável resposta.

2.7 Comentários Finais

No Capítulo 2 foi introduzida um resumo da teoria necessária para o conhecimento geral dos Modelos Lineares Generalizados e particularmente os Modelos de Regressão Logística, o que será fundamental para o entendimento do conteúdo desta dissertação. Esses modelos são freqüentemente utilizados em diversas áreas do conhecimento e em diversas instituições. A Regressão Logística é a metodologia mais usual em todos os modelos do ciclo de crédito (Sabato 2009), porém quando trabalhamos com variáveis preditoras de natureza contínua, temos que supor uma estrutura de relação linear entre essas variáveis e a resposta (2.1). Quando essa estrutura é não ocorre, os Modelos Aditivos Generalizados, particularmente o Modelo Logístico Aditivo, surgem como alternativa.

Capítulo 3

Modelos Aditivos Generalizados (MAG)

3.1 Introdução

Os Modelos Lineares Generalizados assumem que a estrutura funcional entre e variável resposta e as variáveis explicativas é linear, no entanto $f(X)$ pode ser não linear em X, segundo Hastie representar essa estrutura por um modelo linear "é geralmente conveniente e as vezes uma aproximação necessária"(Hastie, 2002, p. 115). Nos casos em que esta estrutura é desconhecida, costuma adicionar-se termos como o quadrado e o logaritmo da covariável, no entanto é difícil fixar qual a forma funcional mais apropriada apenas pela análise gráfica dos dados (Hastie, 1990, p. 1).

A estrutura de relação entre a variável resposta e uma variável preditora pode ser avaliada através de técnicas de alisamento ou regressão não paramétrica. A utilização individual de técnicas de alisamento e regressão não paramétrica como forma de modelagem não tem bom desempenho para grandes números de covariáveis no modelo, isto ocorre porque a grande quantidade de variáveis independentes no modelo causa rápida inflação na variância das estimativas. Um exemplo pode ser visto em Hastie e Tibshirani (1990, p. 83-84).

A questão da rápida inflação na variância de acordo com o aumento da dimensionalidade do modelo é conhecida como a "maldição da dimensionalidade". Em 1985

Stone propôs os Modelos Aditivos (MA), que utilizam as técnicas de alisamento separadamente para cada uma das covariáveis do modelo considerando a distribuição Normal. Desta forma evita-se o problema da dimensionalidade ao preço de que a aproximação não poderá ser feita conjuntamente. Outro ganho que se tem ao utilizar os alisadores univariados é que a interpretabilidade do efeito de uma covariável dadas as demais constantes, que é uma importante característica dos modelos lineares, é garantida sendo as funções univariadas análogas aos coeficientes da regressão linear. Em 1990 Hastie e Tibshirani propuseram uma forma mais geral para os MA que vieram a ser chamados Modelos Aditivos Generalizados (MAG), estendendo os Modelos Aditivos a todas as distribuições da família exponencial.

A função de alisamento univariada de cada uma das j covariáveis, $j = 1, ..., p$ será representada por $f_j(.)$, em analogia aos MLG's temos a substituição do preditor linear $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$ por um preditor aditivo $\eta_i = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + ... + f_p(x_{pi})$, onde $f_j(.)$ são funções preditoras. Apesar dessas funções $f_j(.)$ serem funções suaves, elas podem assumir uma forma paramétrica mais rígida, expandindo esta classe de modelos ao caso paramétrico e semi-paramétrico. Os Modelos Aditivos Generalizados são definidos como uma extensão aos Modelos Lineares Generalizados, basta considerar $f_j(x_j) = \beta_j x_j$ (Hastie, 1990, p. 82). Em relação à interpretabilidade temos que $f_j(.)$ é análoga ao coeficiente β_j no caso dos modelos lineares, já que a estimativa $f_j(x_j)$ explica como a resposta muda de acordo com x_j (Liu, 2007). Ainda mais, por não assumir uma forma rígida, esta função aproxima a real estrutura de relação entre a covariável em questão e a variável de interesse e, por ser uma função univariada apenas em x_j, descreve a relação isolada entre covariável e a resposta.

Para termos não lineares a função $f_j(.)$ pode ser estimada por qualquer método de alisamento (Liu, 2007), dentre eles podemos citar Médias Móveis, *Loess* e *Splines*. Neste trabalho, em particular, as funções utilizadas para a suavização são as *splines*, polinômios por partes.

Para os casos com resposta binária consideramos a distribuição binomial e as funções de ligação podem ser as mesmas utilizadas nos MLG, por exemplo, logito, complemento loglog e probito. Neste estudo a função de ligação escolhida foi a função logito já que ela é de fácil interpretação, de baixo custo computacional (Liu, 2007) e vem apre-

sentando bons resultados no caso dos MLG através do uso da Regressão Logística que, segundo Liu (2007), é o modelo estatístico mais utilizado na indústria de *credit scoring*.

3.2 Definição do modelo

Os Modelos Aditivos (MA) são uma extensão dos Modelos Lineares (ML), que assim como os ML são aditivos nos efeitos dos preditores (Hastie, 1990, p. 86). Sejam $f_j(x_j)$ funções suaves, temos que um Modelo Aditivo é dado por

$$y_i = \alpha + f_1(x_{1i}) + f_1(x_{1i}) + ... + f_p(x_{pi}) + \epsilon_i, \tag{3.1}$$

onde y_i vem de uma distribuição normal, ϵ_i é o erro da i-ésima observação sendo uma ocorrência da distribuição normal com média zero e variância σ^2 e os erros são independentes das covariáveis $X_j's$, implicitamente temos que $E\{f_j(X_j)\} = 0$. As $f_j's$ são funções univariadas, no entanto, segundo Hastie (1990, p. 86), apesar de ser conveniente pensar nestas funções como alisadores univariados esta propriedade não é necessária. No entanto esta discussão e vertente vão além dos objetivos desta dissertação.

A relação entre a variável resposta e as variáveis preditoras é dada por um preditor através de uma função de ligação, neste caso o preditor linear $\eta = \sum \beta_j X_j$ dos MLG, é substituído por um preditor aditivo $\sum \alpha + f_j(X_j)$ que é uma soma de funções suaves. Utilizamos então uma função de ligação para relacionar a variável resposta aos preditores aditivos, temos então o seguinte modelo

$$g(\mu_i) = \alpha + f_1(x_1) + f_2(x_2) + ... + f_p(x_p) + \epsilon_i. \tag{3.2}$$

Neste caso há a necessidade da utilização do método de Modelos Aditivos Ponderados (seção 3.4) adicionado de algumas modificações as quais seguem a mesma linha de modificação na forma de estimação dos ML para os MLG. Na seção 3.4 apresentamos a forma de estimação utilizada.

3.3 Splines

As *splines* são curvas dadas por uma série de polinômios por partes definidos

em sub-intervalos cujos extremos são denominados nós, costuma-se unir os polinômios suavemente nos nós. Uma definição de polinômios por partes é dada por Hastie (2002): "Uma função polinomial por partes $f(X)$ é obtida dividindo-se o domínio X em intervalos contíguos e representando f por polinômios separados em cada intervalo" (Hastie, 2002, p. 117), segundo ele "Uma *splines* de ordem M com nós ξ_j, $j = 1, ..., K$ é polinomial por partes de ordem M, e tem derivadas contínuas até a ordem $M-2$" (Hastie, 2002, p. 120).

Uma escolha popular consiste em utilizar polinômios cúbicos (M = 4) por partes os quais são forçados a serem contínuos e terem até a segunda derivada contínua nos nós. Segundo Hastie (2002, p. 120), não há uma boa razão para ir além das *splines* cúbicas já que nossos olhos só conseguem notar descontinuidade até a segunda derivada.

Suponha que desejamos encontrar uma função suavizadora em X que melhor explique a variável de interesse Y, um suavizador intuitivamente eficiente seria a curva que minimizasse a soma de quadrados $\sum_{i=1}^{n}(y_i - f(x_i))^2$. No entanto, de acordo com Hastie e Tibshirani (1990), o resultado seria uma curva de interpolação nem um pouco suave. Uma forma de controlar esse problema de excesso de rugosidade é a inclusão de um termo que penalize a rugosidade da função obtida. Na estimação por *splines* esta medida de rugosidade é dada por $\int [f^{M-2}(u)]^2 du$ em que f^{M-2} refere-se a derivada da função de ordem $M-2$.

Buscamos então uma função $f(.)$ que minimize a seguinte soma de quadrados de resíduos penalizada

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int_a^b [f''(u)]^2 du, \tag{3.3}$$

onde λ é um parâmetro desconhecido e positivo de suavização, sendo que quanto maior o valor de λ maior a suavidade da função estimada. Nos extremos, quando $\lambda = 0$ tem-se a curva de interpolação, quando $\lambda \to \infty$ temos que a suavização predomina forçando com que $f''(x) = 0$, o que resulta na solução da reta de mínimos quadrados. Os valores a e b são arbitrários contanto que o intervalo contenha os dados, $a \leq x_1 \leq x_2 \leq ... \leq x_n \leq b$.

De acordo com Hastie (1990, p. 27) pode ser mostrado que, de todas as funções $f(x)$ com duas derivadas contínuas, a que minimiza a soma de quadrados de resíduos penalizada (3.3), é a *spline* cúbica natural com nós nos valores únicos x_i.

As *splines* cúbicas naturais são uma variante das *splines* cúbicas, onde as funções

são lineares após os dados extremos. Suponha uma seqüência de nós dada por $a \leq \xi_1 < \xi_2 < ... < \xi_K \leq b$, onde $\xi_1, ...\xi_K$ são números reais, suponha também a adição de dois nós nos extremos aos dados, sejam ξ_0 e ξ_{K+1} os nós extremos, basta garantir que $f''' = f'' = 0$ (Hastie, 1990, p. 24) nos pontos ξ_0 e ξ_{K+1} para que tenhamos linearidade nos extremos. Teremos então para a solução da equação (3.3) $(n-2)$ nós interiores e 2 nós extremos.

Segundo Hastie (2002, p. 120) os problemas de irregularidade (valores grandes para a variância) do ajuste polinomial para os dados localizados próximos aos extremos é exagerado no caso das *splines*, a adição da restrição de linearidade nos extremos é uma solução para este problema, o que também acaba por liberar 4 graus de liberdade (são duas restrições em cada banda). O preço pago por estes ganhos é o vício nos valores extremos, mas segundo ele esta linearidade perto das bandas, onde temos menos informação de qualquer forma, é razoável.

Seja n o número de observações da base de dados e considerando que a solução de (3.3) é uma *spline* cúbica natural $S(x)$, podemos representá-la em termos de uma combinação linear de N bases B de seu espaço.

$$S(x) = \sum_{k}^{N} \gamma_k B_k(x), \tag{3.4}$$

onde γ_k são coeficientes desconhecidos e B_k são as funções bases.

A idéia básica em utilizar a combinação linear de bases para reescrever a função alisadora reside em obter um modelo que seja linear nestas bases (novas variáveis) e basta calcularmos B nos valores observados de x.

De acordo com Hastie (1990 p. 24) uma escolha simples seriam as funções conhecidas como "*truncated power series basis*"que derivam da expressão paramétrica do alisador, no entanto, essas funções não são atraentes numericamente e uma alternativa é a utilização das bases B-*Splines* que apresenta maior eficiência computacional, mais detalhes sobre as bases B-*Splines* ver Apêndice.

Podemos contar o número de bases necessárias para representar uma *spline*. Consideremos K nós e, portanto $(K+1)$ regiões nas quais ajustamos os polinômios por partes, de acordo com Hastie (2002), cada região necessita de M bases para representar o polinômio de grau M, temos então a necessidade de $(K+1) \times M$ bases. No entanto, como dito anteriormente, os polinômios são unidos suavemente nos nós, o que nos leva a restrições,

temos que o número de restrições em cada nó é dado por $(M-1)$. Considerando K nós teremos o ganho de $(M-1)\times K$ parâmetros, teremos então que o número de bases é dado por

$$\begin{aligned} N_{bases} &= (K+1)\times M-(M-1)\times K && (3.5)\\ &= KM+M-KM+K && (3.6)\\ &= M+K && (3.7)\end{aligned}$$

No caso das *splines* cúbicas basta substituir M por 4 e no caso das *splines* cúbicas naturais, as quais apresentam duas restrições extras em cada banda, temos que o número de bases necessárias é de $4+K-4=K$, ou seja, o número de bases é o número de nós da *spline* (Hastie, 2002, p. 121), considerando as *splines* cúbicas naturais temos que o número de bases é o número de valores únicos de x, para simplificar vamos supor por enquanto que o não há valores repetidos de x.

Estendendo a soma de quadrados dos resíduos penalizada para o caso de mais de uma covariável temos a seguinte expressão (Hastie, 1990, p. 111)

$$\sum_{i=1}^{n}[y_i-\sum_{j=1}^{p}f_j(x_{ij})]^2+\sum_{j=1}^{p}\lambda_j\int_a^b[f_j^{''}(u)]^2du \tag{3.8}$$

temos que cada função $f_j(.)$ é penalizada por uma constante separada e se todos $\lambda_j's$ forem iguais a zero teremos $\sum_{j=1}^{p}f_j(x_{ij})=y_i$ para $i=1,...,N$ e caso $\lambda_j's$ forem para o infinito teremos a reta de mínimos quadrados (Hastie, 1990, p. 111).

Considerando todas as funções com segunda derivada contínuas, a solução minimizadora de (3.8) é uma *spline* cúbica em cada uma das covariáveis, podemos reescrever (3.8) matricialmente

$$(y-\sum_{j=1}^{p}f_j)^T(y-\sum_{j=1}^{p}f_j)+\sum_{j=1}^{p}\lambda_j f_j^TK_jf_j, \tag{3.9}$$

onde, $K_j's$ são as matrizes de penalização para cada covariável sendo $K_j=B^{-T}\Omega B^{-1}$, onde $\Omega_{ik}=\int B_i''(x)B_k''(x)dx$. Derivando em relação a f_r e igualando a zero para obter o mínimo temos $-2(y-\sum f_j)+2\lambda_rK_rf_r=0$, para mais detalhes ver (Hastie, 1990), e portanto

$$\hat{f}_r=(I+\lambda_rK_r)^{-1}(y-\sum_{j\neq r}\hat{f}_j) \tag{3.10}$$

para $r = 1, ..., p$ (3.8) são as equações estimadoras e $S_r = (I + \lambda_r K_r)^{-1}$ é a matriz de alisamento da r-ésima *spline* cúbica.

No caso de termos valores repetidos para x_j construímos uma nova base de dados contendo m valores únicos ($m < N$) em seguida agrupamos as observações que apresentam o mesmo valor em x_j atribuindo um valor para a variável resposta, que será a resposta do grupo, esta nova resposta é a resposta média ponderada das observações deste grupo (Hastie, 1990 p.74), por fim ajustamos o MAG considerando o peso de cada grupo como a soma dos pesos das observações que o compõe.

3.4 Estimação

A estimação do MA necessita da resolução de um grande sistema de equações ($np \times np$), uma forma muito utilizada para resolver este problema é o algoritmo *back-fitting*, que é um processo de ajuste iterativo. Este algoritmo é motivado pela seguinte esperança condicional:

$$E[Y - \alpha - \sum_{j \neq r} f_j(X_j)|X_r] = f_r(X_r) \tag{3.11}$$

O algoritmo *back-fitting* é baseado nos resíduos parciais, defina o j-ésimo resíduo parcial como:

$$R_j = y - \alpha - \sum_{r \neq j} f_r(x_r) \tag{3.12}$$

desta forma temos que o j-ésimo resíduo parcial elimina de Y o efeito de todas as outras variáveis e assim pode ser utilizado para modelar o efeito da variável x_j. O algoritmo interativo *back-fitting* começa com valores iniciais de $f_0, f_1, ..., f_p$ e a cada interação cada resíduo parcial R_j é recalculado e a interação acaba quando os componentes individuais não mudam mais.

Esquematizando temos:

1. Inicialize: $\alpha = \sum_i \frac{y_i}{n}, f_j = f_j^0, j = 1, ..., p$

2. Atualize: $j = 1, ..., p, 1, ..., p, ...$
   $f_j = S_j(y - \alpha - \sum_{j \neq r} f_r | x_j)$

3. Repita o passo (2) até que as funções individuais não mudem.

onde $S_j(y|x_j)$ é o alisador da resposta y no preditor x. Como queremos que as funções f_j sejam ajustadas simultaneamente, faz sentido ajustar uma função por vez, retirando o efeito de todas as outras variáveis em y.

No caso generalizado utiliza-se a combinação do algoritmo de estimação utilizado nos modelos lineares generalizados (Mínimos Quadrados Ponderados Iterativamente) com o algoritmo de *back-fitting*, sendo que neste caso substituímos a regressão linear ponderada na variável dependente ajustada por um algoritmo que ajusta um modelo Aditivo Ponderado. Os Modelos Aditivos Ponderados são uma variante dos Modelos Aditivos, eles surgem diretamente como solução para soma de quadrados de resíduos penalizados ponderados (Hastie, 1990, p. 124)

$$(y - \sum_{j=1}^{p} f_j)^T W (y - \sum_{j=1}^{p} f_j) + \sum_{j=1}^{p} \lambda_j f_j^T K_j f_j \tag{3.13}$$

onde W é uma matriz diagonal contendo pesos que podem representar a precisão relativa de cada observação ou ser resultado de um outro procedimento interativo anterior a este (Hastie, 1990, p. 124). As equações de estimação para este caso têm a mesma forma que as equações estimadoras do caso sem pesos (3.10) exceto pelo fato de que os alisadores são *splines* cúbicas ponderadas dadas por $S_j = (W + \lambda_j K_j)^{-1} W$. Mapeando o problema de volta ao caso não ponderado utilizamos as seguintes transformações $y' = W^{1/2} y$, $f_j' = W^{1/2} f_j$, $K_j' = W^{-1/2} K_j W^{-1/2}$. Temos que S_j não é simétrica, mas $W^{1/2} S W^{-1/2}$ é simétrica com autovalores entre $[0, 1]$.

A combinação dos Modelos Aditivos Ponderados ao algoritmo *backfitting* resulta no que denominamos *local-scoring algorithm*. Seja a função de ligação $g(\mu)$ e o preditor η, o MAG dado por $g(\mu) = \alpha + \sum_{j=1}^{p} f_j(x_j)$ e V_i^0 a variância de y no ponto μ_i^0 o método do *local-scoring* segue o seguinte algoritmo:

1. Inicialize: $\alpha = g\left(\sum_i^n \frac{y_i}{n}\right)$; $f_1^0 = f_2^0 = ... = f_p^0$.

2. Atualize:

Construa uma variável dependente $z_i = \eta_i^0 + (y_i - \mu_i^0)\left(\frac{d\eta_i}{d\mu_i}\right)_0$ com $\eta_i = \alpha^0 + \sum_{j=1}^{p} f_j^0(x_{ij})$ e $\mu_i^0 = g^{-1}(\eta_i^0)$.

Construa os pesos $w_i = \left(\frac{d\mu_i}{d\eta_i}\right)_0^2 (V_i^0)^{-1}$

Ajuste um Modelo Aditivo Ponderado a z_i para obter as estimativas para as funções f_j^1, η_i^1 e o valor ajustado μ_i^1.

Calcule o critério de convergência
$\Delta(\eta^1, \eta^0) = \frac{\sum_{j=1}^{p} ||f_j^1 - f_j^0||}{\sum_{j=1}^{p} ||f_j^0||}$

3. Repita o passo (2) substituindo η_i^0 por η_i^1 até que $\Delta(\eta^1, \eta^0)$

3.5 Estimação de λ e do grau de liberdade

Para o caso da utilização das *splines* para uma variável preditora nos MAG temos que $f(x) = S_\lambda y$, notemos que a matriz de suavização é muito similar à matriz de projeção da regressão linear. A definição dos graus de liberdade dos alisadores é baseada na analogia à regressão linear, uma forma simples é dada por $gl = tr(S_\lambda)$, neste caso os graus de liberdade são a soma de autovalores de S_λ e nos dão um indicativo do quanto de ajuste S_λ faz. Quanto maior o número de graus de liberdade, mais complexo é o alisador ajustado. Hastie mostra que existem outras formas de definir os graus de liberdade do suavizador que são $n - tr(2S_\lambda - S_\lambda S_\lambda^T)$ e $tr(S_\lambda S_\lambda^T)$, sendo que cada definição é útil para diferentes propósitos.

Dentre as diferentes formas de estimação para λ, Liu (2007) aconselha selecionar λ a partir dos graus de liberdade, fixando seu valor em 4 (gl = 4) o que segundo ele nos permite resguadar o *over-fitting*, que pode ser muito freqüente em casos de ajustes não lineares em x.

3.6 Teste de significância dos parâmetros

Para testar a significância dos *splines* em cada covariável, Hastie (1990) recomenda usar a função *deviance* (desvio), que mede para a base de dados em questão a discrepância entre o modelo saturado e o modelo em estudo. O modelo saturado é o modelo para o qual os valores ajustados $\hat{\mu}_i$ são iguais às respostas observadas, a *deviance* é definida da seguinte forma,

$$D(Y;\hat{\mu}) = -2\ln\left[\frac{L(Y,\hat{\mu})}{L(\mu_{max},Y)}\right] = -2\left[\ln L(Y,\hat{\mu}) - \ln L(\mu_{max},Y)\right], \tag{3.14}$$

onde μ_{max} é o valor do parâmetro que maximiza $L(\mu, y)$, onde $L(Y,\hat{\mu})$ é a função de máxima verossimilhança do modelo em questão e $L(\mu_{max},Y)$ é a função de máxima verossimilhança do modelo saturado.

No caso dos MLG a teoria assintótica para a diferença de *deviances* entre modelos aninhados já é bem conhecida, no entanto para o caso dos MAG's ainda é pouco desenvolvida, mesmo assim, embasado por estudos de simulação, Hastie aconselha o uso informal do teste baseado nesta diferença. Para o caso da distribuição Binomial e Poisson a distribuição assintótica da diferença de *deviances* pode ser aproximada por uma distribuição $X^2_{gl^{erro}}$. Os testes seguem o mesmo formato visto na seção 2.5 e os graus de liberdade devido a variável x_j considerados por Hastie (1990) são dados por,

$$gl_j^{erro} = tr(2S_\lambda - S_\lambda S_\lambda^T) - tr(2S_{(j)\lambda} - S_{(j)\lambda}S_{(j)\lambda}^T). \tag{3.15}$$

3.7 Resposta binária

Quando a estrutura funcional entre as variáveis preditoras e a variável resposta é não linear utiliza-se adicionar termos como a variável preditora ao quadrado no Modelo Linear utilizado. No caso de resposta binária dentro do contexto dos Modelos Lineares Generalizados uma prática muito comum reside em, após a confirmação da não linearidade, fazer uma análise bivariada entre cada preditora e a resposta, segmentando a covariável de forma que cada grupo seja o mais homogêneo possível com relação a proporção de eventos (mais detalhes são abordados na seção 6.2.1).

No entanto, a categorização de uma covariável contínua pode acarretar em perda significativa de informação o que nos motiva a buscar uma alternativa que permita a

utilização da covariável em sua forma original, e aqui propomos como solução o uso dos Modelos Aditivos Generalizados para dados binários, o qual também pode ser denominado Modelo Logístico Aditivo.

$$\begin{aligned} \eta &= g(\pi) && (3.16)\\ &= log\left(\frac{\pi}{1-\pi}\right) && (3.17)\\ &= \alpha + f_1(x_1) + f_2(x_2) + ... + f_p(x_p) && (3.18) \end{aligned}$$

Nesta caso z_i e w_i mencionados no algoritmo *local-scoring* são dados por:

$$z_i = \alpha + \sum_{j=1}^{p} f_j(x_j) + \frac{y_i - \pi_i}{pi(1-\pi_i)}; \quad w_i = \pi_i(1-\pi_i) \tag{3.19}$$

onde f_j são as estimativas atuais e $\pi_i = \frac{exp\{\alpha+\sum_j f_j(x_j)\}}{1+exp\{\alpha+\sum_j f_j(x_j)\}}$, novas funções f_j e α são calculadas ajustando-se um modelo aditivo ponderado em z_i.

3.8 Comentários Finais

No Capítulo 3 foi introduzida a teoria necessária para o conhecimento dos Modelos Aditivos Generalizados, particularmente do Modelo Logístico Aditivo, a partir do uso de funções *splines*, o que será fundamental para o entendimento do conteúdo desta dissertação. Esses modelos são apresentados como alternativa aos Modelos Lineares Generalizados, particularmente ao Modelo de Regressão Logística. A Regressão Logística é a metodologia mais usual em todos os modelos do ciclo de crédito (Sabato 2009) porém, como vimos, o Modelo Aditivo Logístico pode ser uma alternativa quando variáveis preditoras de natureza contínua não apresentam uma estrutura de relação linear com a resposta.

Dadas essas duas metodologias (MLG e MAG), após o ajuste é fundamental a análise da capacidade preditiva desses modelos. Tanto na aplicação em dados reais como no estudo de simulação, ambos abordados nesse trabalho, foram utilizadas as técnicas descritas no Capítulo 4.

Capítulo 4

Análise do desempenho do modelo

Existem dois interesses primordiais nos modelos do ciclo de crédito: a classificação dos clientes como bons ou maus pagadores e a predição da probabilidade de inadimplência de cada cliente. Considerando estes dois interesses devemos validar tanto o modelo de probabilidade como o modelo de classificação por meio de métricas de rank e de classificação.

É de interesse primordial garantir o método de modelagem apresente bom desempenho em uma base de dados que não tenha sido utilizada para construir o modelo, caso isto ocorra temos o indicativo de que o modelo pode ser aplicado a novos exemplos do mesmo domínio de dados (Picarrd, 1990), portanto as métricas apresentadas neste capítulo foram aplicadas às duas bases obtidas a partir da amostra original, a base desenvolvimento e a base validação.

Nesta dissertação a separação foi feita em 70% e 30% da base original, a primeira foi utilizada para a construção do modelo (base desenvolvimento ou treinamento) e a segunda foi utilizada para indicar se o modelo é válido em novos exemplos do mesmo domínio de dados (base teste ou validação).

A mesma base de desenvolvimento foi utilizada para o ajuste de ambos os modelos: os Modelos Lineares Generalizados e os Modelos Aditivos Generalizados para dados binários e as comparações entre eles foram feitas considerando a capacidade de predição e a robustez dos dois diferentes modelos obtidos. As análises da capacidade preditiva e de robustez foram feitas a partir dos valores preditos obtidos pela aplicação do modelo na amostra de desenvolvimento e validação.

Para o caso dos Modelos Lineares Generalizados as variáveis contínuas foram categorizadas (Gruenstein, 1998) e para ambos os casos, Modelos Lineares Generalizados e Modelos Aditivos Generalizados, algumas variáveis categóricas foram recategorizadas, e a recategorização foi a mesma para ambos os modelos, seguindo uma análise pré modelagem (seção 6.2.1). As categorizações e recategorizações foram feitas de acordo com a análise descritiva bivariada entre cada uma das covariáveis e a variável resposta, neste caso a ocorrência ou não da inadimplência.

Para cada uma das amostras (treinamento e teste) a análise da capacidade preditiva do modelo foi feita em duas etapas: a primeira foi feita a partir das probabilidades de inadimplência (*scores*) preditas pelo modelo estatístico e a segunda foi feita após a aplicação do ponto de corte e da classificação de cada cliente como adimplente ou inadimplente, ou seja, primeiramente avaliamos a capacidade de previsão do modelo estatstico e posteriormente a capacidade preditiva do modelo de classificação. Todas as técnicas de análise da capacidade preditiva e da análise de robustez tratadas neste trabalho podem ser aplicadas a outras bases de dados cujo interesse seja o de modelar uma variável binária.

A prática sugere que a avaliação do modelo na amostra de treinamento, utilizada para o seu desenvolvimento, apresenta resultados melhores do que se avaliado na amostra teste, uma vez que o modelo incorpora peculiaridades inerentes da amostra utilizada para a sua construção (Abreu, 2004).

As métricas de análise de desempenho utilizadas para os modelos de *score* foram: a estatística de Kolmogorov-Smirnov (KS), a curva ROC, a análise dos decis (baseada na divisão de dados ordenados pelos valores preditos, em 10 partes iguais) e a medida *lift* (razão entre a taxa de resposta predita do subconjunto em questão em relação a taxa de resposta geral de toda a população) baseada na segmentação por decis.

No caso dos modelos de classificação a análise do desempenho também é chamada de análise da capacidade preditiva do modelo e é caracterizada pela capacidade do mesmo em classificar corretamente sujeitos como susceptíveis ou não à ocorrência de um determinado fenômeno.

Para os modelos de classificação utilizamos as métricas baseadas na matriz de confusão: Sensibilidade, Especificidade, VPP, VPN, Acurácia e MCC.

Avaliado o desempenho do modelo na base utilizada para a construção do modelo aplicamos o modelo na amostra de validação, que pode ser uma amostra aleatória ou uma base futura. Essa etapa, chamada de validação, tem o objetivo de verificar se a qualidade do modelo se replica em outro universo de tempo ou aleatoriedade.

4.1 Estatística de Kolmogorov-Smirnov (KS)

Uma forma usual para avaliar o desempenho de modelos de *score* é a medida da estatística de Kolmogorov-Smirnov (KS). Baseada no teste não-paramétrico de Kolmogorov-Smirnov, em que se deseja a partir de duas amostras retiradas de populações possivelmente distintas, testar se duas funções distribuições associadas às duas populações são idênticas ou não. A estatística de Kolmogorov-Smirnov mede o quanto estão separadas as funções distribuições empíricas dos *scores* dos grupos onde há a ocorrência do evento e onde não há a ocorrência do evento, no nosso caso são os grupos de maus (M) e de bons (B) pagadores, respectivamente. Sendo $F_B(e) = \sum_{x \leq e} F_B(x)$ e $F_M(e) = \sum_{x \leq e} F_M(x)$ a função distribuição empírica do grupo de bons e maus pagadores, respectivamente, então a estatística de Kolmogorov-Smirnov é dada por:

$$KS = \max|\mathrm{F_B(e)} - \mathrm{F_M(e)}| \tag{4.1}$$

onde $F_B(e)$ e $F_M(e)$ corresponde às proporções de clientes dos grupos onde não há a ocorrência do evento e do grupo onde há a ocorrência do evento respectivamente, no nosso caso bons e maus, com *score* menor ou igual a e. A estatística KS é obtida através da distância máxima entre essas duas proporções acumuladas ao longo dos *scores* obtidos pelos modelos, apresentada na figura 4.1.

O valor dessa estatística pode variar de 0 a 1, em que o valor máximo indica uma separação total dos escores dos bons e maus clientes e um valor mínimo de 0 sugere uma sobreposição total das distribuições dos escores dos dois grupos. A representação dessa estatística pode ser vista na Figura 4.2.

Na Figura 4.2 temos no painel superior esquerdo um KS de aproximadamente 0, no painel superior direito um KS entre 0.3 e 0.6 e no painel inferior um KS de aproximadamente 1. Pode-se notar que o KS maior indica a maior distância entre as distribuições dos dois grupos, de ocorrência e não ocorrência do evento, ou no caso de *credit score*, dos

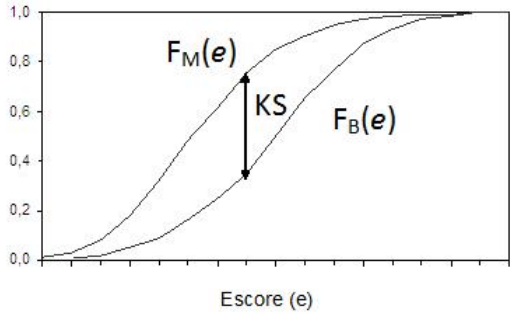

Figura 4.1: Funções de distribuições empíricas para os bons e maus clientes e a estatística KS

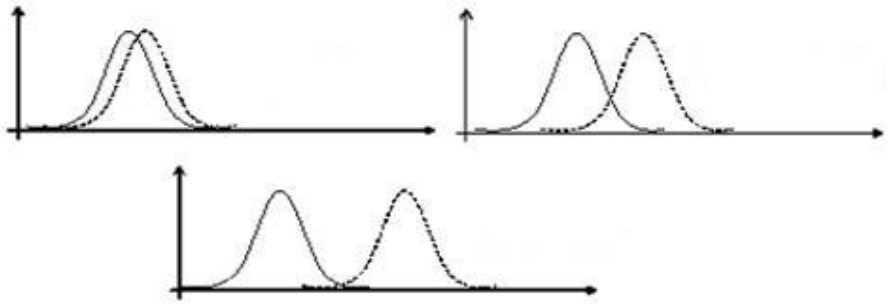

Figura 4.2: Interpretação da estatística KS

maus e dos bons clientes. Para mais informações sobre o teste de Kolmogorov-Smirnov consultar Salvatore, 2002.

4.2 A Curva ROC

A curva ROC (*receiver-operating characteristic*) foi desenvolvida no contexto de detecção de sinais eletrônicos e problemas com radares, no início dos anos 50 (Zweig & Campbell, 1993). O objetivo da técnica era quantificar a habilidade dos radares em distinguir um sinal de um ruído (Reiser & Faraggi, 1997). Na década de 60, curvas ROC foram usadas em psicologia experimental e nos anos 70, a metodologia se disseminou amplamente em vários ramos da pesquisa biomédica.

Esta medida é uma forma gráfica de se avaliar a qualidade de um modelo de Regressão Logística. Além de apresentar um indício sobre a qualidade do modelo, esta curva também é utilizada na determinação de um ponto de corte P_c que aplicado aos valores preditos determinará um modelo de classificação dentre os mais satisfatórios para o

problema. Este ponto de corte, ou limiar, como também é chamado, é o valor limite estimado do *score* que separa os grupos ocorrência e não ocorrência do evento de interesse, no caso dos modelos de crédito, os grupos de inadimplentes e adimplentes, respectivamente. Este ponto deve ser escolhido de forma a minimizar os erros de classificação. Mas é uma medida insensível e ineficiente para precisão preditiva (Van Houwelingen JC, 1990)

Assim como as probabilidades preditas, o ponto de corte deve ser um número entre 0 e 1, sendo que para cada limiar temos um modelo de classificação e a partir deste, podemos calcular as probabilidades condicionais de uma observação que apresentou a ocorrência do evento ser classificada (pelo modelo) corretamente e de uma observação que apresentou a não ocorrência do evento ser classificada (pelo modelo) corretamente, estas probabilidades condicionais são denominadas, respectivamente, sensibilidade (Se) e especificidade (Es), mais detalhes sobre a sensibilidade e especificidade podem ser encontrados na seção 4.4.

Variando-se pontos de corte ao longo da amplitude dos *scores* fornecidos pelo modelo, obtém-se diferentes classificações para as observações. Para cada P_c obtemos os respectivos valores para as medidas de sensibilidade e especificidade, sendo que quanto maior o ponto de corte , maior a especificidade do modelo e menor a sua sensibilidade, e quanto menor o ponto de corte, maior a sensibilidade e menor a especificidade. Assim, a curva ROC é construída tendo no seu eixo horizontal os valores de $(1-Especificidade)$, ou seja, a proporção de observações caracterizadas pela não ocorrência do evento que foram classificadas como portadoras da característica eixo vertical os valores de $Sensibilidade$, isto é, a proporção de observações caracterizados pela ocorrência do evento que foram classificadas como portador da característica.

O modelo de *score* é avaliado através da comparação dos resultados da classificação originada por ele com a classificação aleatória, equivalente a jogar ao alto uma moeda honesta para classificar cada observação segundo a face observada. No gráfico a classificação aleatória equivale a uma reta de 45º, que parte da origem do gráfico até o canto superior direito. Considerando que um modelo de classificação tem no mínimo a mesma qualidade da classificação sem o uso de ferramentas, interpretamos a curva de forma que quanto mais distante da diagonal principal ela estiver, melhor é o desempenho do modelo. Esse fato sugere que quanto maior for a área entre a curva ROC produzida e a diagonal

principal, melhor é o desempenho global do modelo. Uma vantagem da curva ROC está em sua simplicidade já que consiste em uma representação direta do desempenho de um modelo, de acordo com o conjunto de suas possíveis respostas.

A escolha de um ponto de corte ótimo para um modelo de classificação quantitativo depende basicamente do objetivo do estudo. Segundo Schäfer (1989), uma seleção de pontos de corte ótimos poderia ser baseada no valor máximo de uma combinação linear de Se e Es. Tal combinação pode ser, por exemplo, a soma destas medidas, como propõe Linnet & Brandt (1986), sendo esta a situação mais adequada, visto que classificar como portador da característica de interesse um indivíduo que originalmente não apresenta sua ocorrência e classificar como não portador da característica de interesse um indivíduo que originalmente apresenta sua ocorrência têm custos iguais para a empresa (Altman, 1991, p.418).

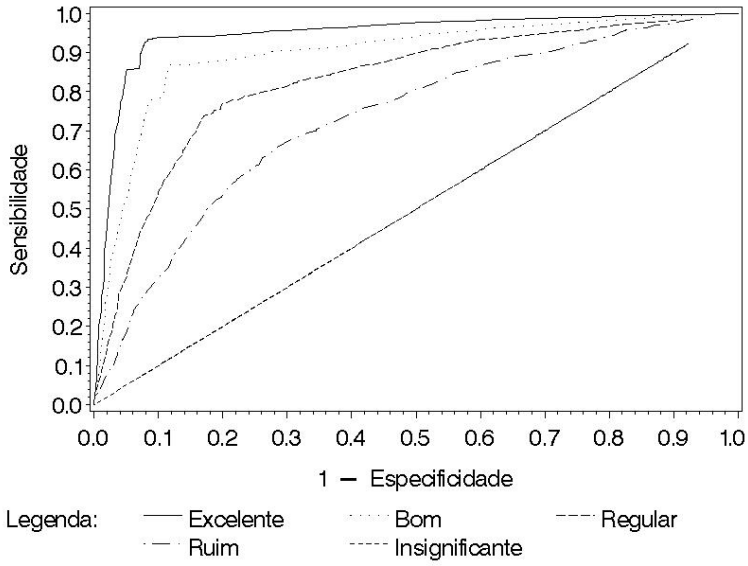

Figura 4.3: Exemplo de curvas ROC

Vemos na Figura 4.3 5 exemplos de curvas ROC, onde o que indica o pior modelo é a curva mais próxima da reta base traçada no plano, e em contrapartida o melhor modelo é o que tem a curva mais distante. Vemos também que o ponto de maior sensibilidade e menor (1-especificidade) é o ponto que mais se aproxima da parte superior esquerda do plano, isto é, o ponto mais afastado da reta base.

Em uma escala comum, gráficos de duas ou mais curvas ROC representando diferentes modelos de classificação contínuos permitem uma imediata comparação de de-

sempenhos. A visualização de estimativas de Se e Es para diferentes pontos de cortes reflete a capacidade do modelo em discriminar observações originalmente portadoras e não portadoras da característica de interesse em várias situações, e, a partir daí, cabe ao investigador a decisão de como melhor utilizar o modelo (ou os modelos) como ferramenta ao seu trabalho de classificar.

4.3 Análise por decis

Outra forma de avaliar os modelos de scoring baseia-se na partição da base em subgrupos, sendo possíveis dois tipos de análise, uma análise é feita em cada segmento e outra é feita através da comparação entre os segmentos. Em cada segmento desejamos verificar se em média os valores preditos pelo modelo se aproximam da condição real das observações como portadora ou não da característica de interesse. Na comparação entre os segmentos desejamos avaliar o quanto, eles diferem com relação a serem originalmente portadores ou não da característica de interesse.

A partição da base deve ser feita de forma que os indivíduos dentro de cada segmento sejam similares quanto à condição real da ocorrência ou não do evento e que os segmentos apresentem o mesmo tamanho. A segmentação é feita a partir das probabilidades preditas pelo modelo $\hat{\pi}_i$, ordenamos as observações através do escore de cada observação e definirmos os segmentos a partir desta ordem. Supondo que o primeiro segmento seja referente aos primeiros indivíduos da base ordenada de forma decrescente pelos valores preditos e o ultimo segmento seja referente aos últimos indivíduos da base ordenada. Esperamos que, se o modelo tem bom desempenho, a média de ocorrência de evento no primeiro segmento será maior do que a média do segundo que será maior do que a média do terceiro, e assim por diante. Esperamos também que para cada segmento a média dos valores preditos seja próxima da média de ocorrência real de evento e que média de ocorrência real de evento do primeiro segmento seja muito maior do que a do último.

Nesta forma de análise é comum trabalhar com 10 segmentos, cada um contendo 10% dos dados, neste caso os segmentos ou faixas de escore são denominados decis. Como os decis são definidos a partir da base ordenada de forma decrescente pelos valores pre-

ditos do modelo, devemos verificar se a média de ocorrência de evento diminui conforme aumentam os decis, dessa forma temos um indicativo de se o modelo está sendo capaz de ordenar o evento de interesse, tendo assim um indicativo sobre sua coerência. É importante verificar também se os escores médios e as taxas reais do evento apresentam valores próximos o que nos indica sobre a aderência. Na comparação entre os segmentos desejamos avaliar o quanto, eles diferem com relação a serem originalmente portadores ou não da característica de interesse. Quanto mais os segmentos diferirem entre si, melhor o modelo esta sendo capaz em discriminar os indivíduos quanto a sua real condição, estas diferenças indicam a eficiência do modelo.

Na análise por decis os escore do modelo são ordenados e alocados em diferentes faixas de escore, em que a faixa 0 conterá os 10% dos maiores escore, a faixa 1 conterá os 10% dos segundos maiores escore, e assim por diante, até a nona faixa, que conterá os 10% dos menores escore. Cada indivíduo, alocado a um grupo de acordo com seu escore ($\hat{\pi}_i$), carregará consigo a variável binária que indica a ocorrência ou não do evento de interesse, de posse desses valores podemos calcular a taxa real de resposta dentro de cada decil.

A construção da Tabela de Análise por Decis segue o seguinte esquema:

1. Ajustamos o modelo logístico e obtemos o *score* de cada indivíduo
2. Ordenamos o *score* de forma decrescente, do mais provável ao menos provável
3. Separamos os 10% primeiros indivíduos → grupo 0: [0-1) decil
4. Separamos os 10% segundos indivíduos → grupo 1: [1-2) decil
5. E assim sucessivamente até chegar no grupo 9: [9-10) decil. E assim criamos 10 faixas de *score*, todas do mesmo tamanho (10% da base)
6. Para cada faixa de *score* k, onde k =0,1,2,...,9 calculamos:

   → O *Score* médio, dado por:

$$\frac{1}{0,1n}\sum_{i=(\frac{k}{10}n)+1}^{(\frac{k+1}{10})n}(\hat{\pi}_i) \tag{4.2}$$

→ Taxa real de resposta

$$\frac{1}{0,1n} \sum_{i=(\frac{k}{10}n)+1}^{(\frac{k+1}{10})n} (y_i)\,, \tag{4.3}$$

onde $\hat{\pi}_i$ é o *score* predito e y_i é o valor da i-ésima variável resposta (0,1)

7. Construir uma tabela cuja primeira coluna indique os Decis, a segunda coluna contenha os *scores* médios e a terceira coluna contenha a Taxa real do evento (inadimplência)

O gráfico de aderência (Figura 4.4) nos permite verificar se taxa real de resposta e o escore médio são próximos em cada faixa de escore e também se a taxa real de resposta decresce monotonicamente conforme aumenta a faixa de escore. Nele temos que o eixo x contém as faixas de escore, o eixo y o escore médio em barras e a taxa de resposta em pontos ligados por uma linha. No painel esquerdo da figura 3.4 temos o exemplo de um modelo aderente, ao contrário do painel direito em que temos um modelo não aderente.

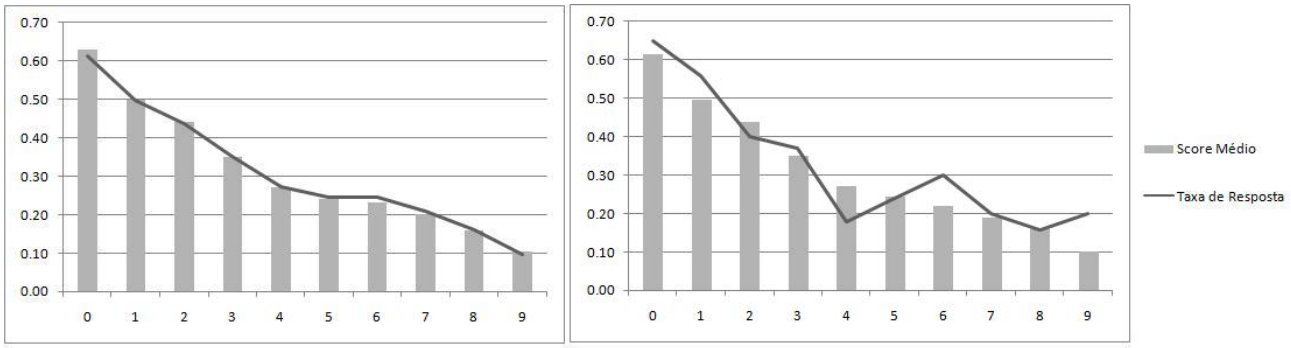

Figura 4.4: Exemplos Gráficos de Aderência

Uma forma simples de avaliar a eficiência do modelo é dividindo a taxa de resposta da faixa 0 pela taxa de resposta da faixa 9, quanto maior for este valor, maior é a eficiência do modelo. Se o resultado for, por exemplo, 8 temos então que a faixa de escore 0 é 8 vezes mais predisposta à resposta do que a faixa 9. Uma forma de avaliar a robustez do modelo é comparando as tabelas geradas pela análise de decis da amostra desenvolvimento com as geradas a partir da amostra validação. Se o comportamento das taxas e do escore médio forem similares e seguirem o mesmo padrão temos um indicativo de um modelo robusto.

A análise de decil nos permite avaliar a capacidade preditiva do modelo, bem como sua aderência e robustez. Utiliza-se também uma tabela mais elaborada, que parte

da análise de decil, denominada Tabela de Ganhos. Esta tabela contém novas medidas e dessa forma nos permite obter novas interpretações.

4.3.1 Tabela de Ganhos

A tabela de ganhos contém algumas medidas chaves que são o *lift* (Rud, 2001) e o percentual do total de eventos a cada faixa de escore. O *lift* mede, a cada faixa, o quanto o modelo se mostra superior à média em detectar o evento de interesse, e a percentagem do total de eventos de cada faixa nos permite avaliar o quanto ganhamos utilizando apenas um grupo de indivíduos. Na tabela de ganhos trabalhamos com estas medidas também em sua forma acumulada, o *cum lift* e o percentual acumulado do total de eventos.

A tabela de Ganhos m sua forma completa é composta por 12 colunas: decil (A), número de observações (B), % da população (C), escore médio (D), taxa real de evento (E), taxa real de evento acumulada(F), número de eventos(G), % do total de eventos (H), número de eventos acumulado(I), % acumulada do total de eventos (J), lift (K) e *lift* acumulado (L).

A descrição de cada coluna é resumida a seguir:

- Coluna A indica o decil, que varia de 0 a 9
- Coluna B mostra cada decil contendo 10% da população
- Coluna C mostra a percentagem acumulada da base, que varia de 10% a 100%
- Coluna D mostra a média das probabilidades preditas de cada indivíduo contido no grupo
- Coluna E mostra a percentagem média da ocorrência do evento em cada grupo. Que representa o número de eventos dividido pelo número de observações no decil (coluna B)
- Coluna F é uma forma acumulada da coluna E, da mesma forma que podemos calcular a percentagem média da ocorrência do evento no decil 0, podemos calcular esta percentagem dado um conjunto que inclui o decil 0 até um decil desejado.

- Coluna G é o número de observação em que houve a ocorrência do evento (coluna B * coluna E)
- Coluna H é a coluna G dividida pela soma da coluna F. O que representa a percentagem da ocorrência de eventos do total de eventos ocorridos na base
- Coluna I é uma forma acumulada da coluna H
- Coluna J é a coluna I dividida pelo número total de eventos
- Coluna K é o *lift* de cada decil, o *lift* é calculado dividindo-se o percentual de eventos pela média do percentual de eventos, ou seja o percentual de eventos da base de dados não segmentada. Mostra assim, em cada decil quantas vezes as observações são mais provaveis de apresentar a ocorrência do evento do que a média geral.
- Coluna L é o *lift* acumulado e é calculado dividindo-se a coluna F pela percentagem geral do evento.

Seja k o indicador do decil, k = 0,1,..,9, n o número de observações da base de dados utilizada na análise, y_i uma variável binária (variável de interesse) que para a i-ésima observação vale 1 se houve a ocorrência do evento e 0 caso contrário o *lift* do k-ésimo segmento também pode ser calculado a partir da fórmula 4.4.

$$lift_k = \frac{\frac{1}{0,1n}\sum_{i=\frac{k}{10}n+1}^{\frac{k+1}{10}n} y_i}{\frac{1}{n}\sum_{i=1}^{n} y_i} \tag{4.4}$$

O *lift* indica quantas vezes do decil em questão é realmente mais propenso do que a abordagem aleatória (a não utilização da modelagem) à ocorrência do evento, ou seja a cada faixa de escore ele mostra o poder do modelo em superar a abordagem aleatória. Suponha uma empresa que deseja vender um de seus produtos, que um modelo logístico já tenha estimado a probabilidade de cada um dos pretendentes em comprar o produto e que a empresa não quer ter gastos desnecessários com os canais de venda. Neste caso seria coerente escolher os pretendentes a partir de sua localização nas faixas de escore, por exemplo, se o primeiro decil tem um *lift* de 3,4 significa que se a empresa fizer a propaganda para os indivíduos deste decil a chance deles comprarem o seu produto é 3,4 vezes maior do que se a empresa escolhesse este mesmo número de pretendentes ao acaso.

Assim como a taxa de resposta, o *lift* pode informar sobre a capacidade de ordenação do modelo. Se os *lifts* decrescerem a medida que aumentamos as faixas de escore temos que o modelo é coerente. Esta análise pode ser feita graficamente através do *lift chart*, onde no eixo x temos os decis e no eixo y temos os *lifts*, se obtivermos uma curva monotonicamente decrescente concluímos que o modelo é capaz de ordenar os indivíduos com relação à ocorrência do evento de interesse. Os primeiros decis são os mais importantes pois eles mostram a capacidade de o modelo em prever a ocorrência do evento de interesse, portanto se houver alguma incoerência, baixa eficiência ou até mesmo erros consideráveis de previsão nos últimos decis o modelo ainda poderá ser considerado relativamente bem ajustado, no entanto alguns analistas mais rigorosos preferem agrupar decis de forma que as novas faixas apresentem o *lift* monotonicamente decrescente. Pode-se ainda, por esse mesmo gráfico, observar o quanto a utilização do modelo é mais eficiente do que uma abordagem aleatória. Para isso, basta compararmos a curva obtida (modelo) com uma reta horizontal (abordagem aleatória).

O *cum lift* é muito útil na análise de eficiência, basta analisar o salto do *cum lift* de um decil para outro, quanto maiores os saltos mais eficiente é o modelo, já que grandes saltos indicam grandes diferenças entre os grupos. Esta análise nos permite verificar se o modelo está conseguindo distinguir os 10 grupos, caso algum salto seja pequeno podemos, por exemplo, agrupar 2 decis formando um grupo só. Se esta providência aumentar a eficiência do modelo é valida.

O desempenho do modelo é avaliado a partir do gráfico de ganhos. Nele temos no eixo x a percentagem da base de dados acumulada a cada faixa de escore, e no eixo y temos a percentagem acumulada do total de eventos. Sobrepomos a este gráfico uma reta de 45 graus, representando a abordagem aleatória, assim verificamos qual a distância entre a abordagem aleatória e o uso do modelo em questão. Quanto maior a distância, mais se ganha ao utilizar o modelo. A figura 4.5 mostra um exemplo do gráfico de ganhos. Nele podemos verificar, por exemplo, que com 30% da base capturamos 55% do evento. Sabemos que, em uma abordagem aleatória, 100% da base representa 100% dos eventos, portanto esperamos que 30% da base representem 30% dos eventos. Utilizando o modelo, como mostra a figura 4.5, podemos captar 83% a mais que na abordagem aleatória, utilizando os mesmos 30% da base. Supondo um problema de *credit score* temos que a recusa do crédito à 30% da base nos permite a eliminação de 55% dos casos

de inadimplência.

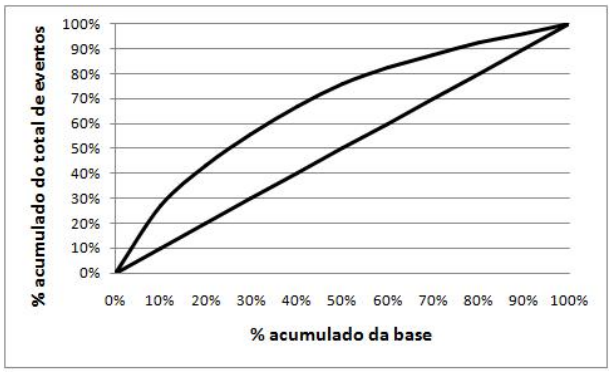

Figura 4.5: Exemplo do gráfico de ganhos

Para avaliar a robustez do modelo, utilizamos intervalos de confiança para o *lift* ou *cum lift*. Quanto menor a amplitude do intervalo de confiança, mais robusto é o modelo. Uma forma mais simples de avaliar a robustez é através da comparação entre os *lifts* da amostra validação e da amostra desenvolvimento. Pode-se também calcular os intervalos de confiança dos *lifts* na base desenvolvimento e verificar se esses valores na base validação estão contidos neste intervalo. Naturalmente haverá uma perda de qualidade do modelo aplicado na base validação em relação ao modelo ajustado na base desenvolvimento, desta forma a comparação deverá ser feita verificando se as diferenças são aceitáveis e se as medidas a cada faixa de *score* seguem o mesmo padrão. Uma forma de visualizar esta comparação é através de um gráfico onde o eixo x contém as faixas de *score* e o eixo y contém os *lifts* da base desenvolvimento e da base validação (*lift chart*). Na figura 4.6 temos no painel direito um exemplo em que os *lifts* da base validação acompanham os padrões dos *lifts* da base desenvolvimento, o oposto do que ocorre no painel direito.

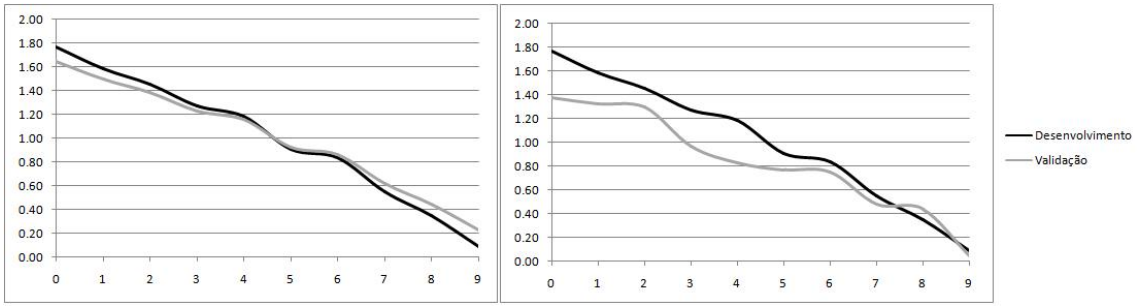

Figura 4.6: Comparação dos *lifts* da base desenvolvimento e validação

A estimação de intervalos de confiança depende do desvio da estimativa, mas como

geralmente temos em mãos apenas uma amostra para o ajuste do modelo, utilizamos uma técnica não paramétrica de reamostragem para a estimação dos desvios dos *lifts* a cada decil. Neste contexto introduz-se o uso da re-amostragem via bootstrap (Rud, 2001) não paramétrico, que permite a obtenção de novas amostras a partir da amostragem com reposição da base original.

4.3.2 Re-amostragem não paramétrica via bootstrap

A reamostragem tenta realizar o que seria desejável realizar na prática, se tal fosse possível: repetir a experiência, ou seja, permite simular outras amostras que poderiam ter sido coletadas da população. A reamostragem via bootstrap não paramétrico consiste em obter uma amostra aleatória completa a partir dos dados originais, considerando que as observações são independentes, sendo que a escolha de cada observação é feita com reposição. É então possível que uma observação seja considerada diversas vezes enquanto outra observação pode não ser considerada nenhuma vez sequer, ou seja, haverá uma ligeira diferença entre as amostras.

Geralmente este procedimento é repetido 25 vezes, gerando assim 25 amostras bootstrap e, a partir delas, pode-se calcular a estimativa de interesse bem como seu intervalo com 95% de confiança baseado na distribuição Normal. Seja *m = 1, 2, ..., 25*, e a estimativa de interesse θ, então o procedimento para a obtenção da estimativa do parâmetro de interesse e de seu intervalo de confiança segundo Rud (2001) é:

1. Obtenha da amostra original 25 amostras independentes e com reposição
2. Calcule a estimativa de cada amostra bootstrap $\rightarrow \theta_{boot(m)}$
3. Calcule a variância $\rightarrow \sigma^2_{boot} = \frac{\sum(\theta_m - \bar{\theta})^2}{25-1}$
4. Encontre a estimativa geral de bootstrap $\rightarrow \theta_{boot} = 2\theta_{original} - \frac{\sum \theta_{boot(m)}}{25}$
5. Encontre o limite superior para a estimativa $\rightarrow \theta_{superior} = \theta_{boot} + Z_{0.025}\sigma_{boot}$
6. Encontre o limite inferior para a estimativa $\rightarrow \theta_{inferior} = \theta_{boot} - Z_{0.025}\sigma_{boot}$

onde, Z_α é o α-*ésimo* percentil da normal padrão, ou seja Z $\sim$ N(0,1).

Desta forma é possível obter uma estimativa para o *lift* de cada decil bem como seu intervalo com 95% de confiança baseados na reamostragem via bootstrap não paramétrico. A idéia é encontrar os intervalos e aplicá-los no *lift chart* para verificar sua amplitude, bem como verificar se as estimativas das amostras estão contidas no intervalo. Como já dissemos, a amplitude do intervalo pode também nos dizer sobre a robustez do modelo ajustado.

Quando tratamos de modelos de classificação, temos interesse em verificar sua capacidade de acerto. Como já vimos, para esses modelos é usual definir um ponto limite que seja capaz de classificar os clientes como, por exemplo, bons e maus pagadores. Algumas medidas são usuais para analisar essa capacidade de classificação, comparando o observado com o classificado pelo modelo. Essas medidas são obtidas a partir de uma matriz conhecida como Matriz de Confusão, e é comum obter-se essas medidas para as amostras de desenvolvimento e validação, sempre comparando seus resultados. Assim, além de verificarmos a capacidade de classificação, temos também uma forma de validar esse modelo.

4.4 Medidas em modelos de classificação

Uma vez construído um modelo de classificação passamos à etapa de avaliação do mesmo, isto é, o quanto o *score* produzido pelo modelo consegue distinguir entre as observações em que há e não há a ocorrência real do evento de interesse, por exemplo, entre maus e bons clientes, uma vez que o objetivo é identificar previamente esses grupos e, no caso da concessão de crédito, tratá-los de forma distinta através de diferentes políticas de relacionamento.

A avaliação do modelo é direcionada pela comparação das previsões feitas por ele com a verdadeira classificação da observação, que é geralmente conhecida e está presente, como informação básica, na amostra de teste ou validação.

Seja D uma variável que indica a classificação original, podendo portanto assumir dois valores, 0 ou 1, T uma variável que indica a classificação resultante da aplicação do ponto de corte nos scores obtidos pelo modelo, podendo portanto assumir dois valores 0 ou 1, $c + d$ o número de observações que não apresentaram a ocorrência do evento de

uma determinada amostra teste e $a + b$ o número de observações que apresentaram a ocorrência do evento de interesse. A partir de um determinado modelo de classificação podemos determinar para cada observação i, um *score* s_i, tal que $0 \leq s_i \leq 1$. Suponha que uma observação seja classificada como não portadora da característica de interesse se $s_i > P_c$ e como portadora da característica de interesse se $s_i \leq P_c$, onde P_c é uma probabilidade denominada ponto de corte (*cut-off*).

Se uma observação que não apresenta a característica de interesse for classificada como não portadora dessa característica ou uma observação que apresenta a característica de interesse for classificada como portadora dessa característica, podemos dizer que ela foi classificada corretamente. Fixando-se um ponto de corte podemos construir a matriz de confusão dada pela Tabela 4.1.

Tabela 4.1: Matriz de confusão.

Resultado	Real		
do modelo	Positivo ($D+$)	Negativo (D-)	Total
Positivo ($T+$)	*a (VP)*	*b (FP)*	$a + b$
Negativo (T-)	*c (FN)*	*d (VN)*	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Na Tabela 4.1, temos que:

- a representa o número de maus clientes, classificados corretamente como maus, isto é, verdadeiros positivos (VP);
- b representa o número de maus clientes, classificados incorretamente como bons, portanto revela o número de resultados falso-positivos (FP) que o método detectou;
- c representa o número de clientes bons, classificados incorretamente como maus, ou seja, o número de resultados falso-negativos (FN);
- d representa o número de clientes bons, classificados corretamente como bons, isto é, o número de resultados verdadeiramente negativos (VN);
- $a + c$ - representa o número total de maus clientes;

- $b + d$ - representa o número total de bons clientes;
- $a + b$ - representa o número de clientes identificados pelo modelo como maus;
- $c + d$ - representa o número de clientes identificados pelo modelo como bons.

A capacidade preditiva um modelo está relacionada com suas medidas de desempenho, que podem ser calculadas a partir da Tabela 4.1, dentre as quais podemos citar: a sensibilidade, a especificidade, os valores de predição positivo e negativo, e a acurácia. Estas medidas são definidas a seguir juntamente com a importante definição de prevalência.

Definição 1 (Prevalência): A prevalência de uma característica é definida como a proporção de observações em uma população que são propensas à mesma. Dessa maneira, ela expressa a probabilidade da observação apresentar a característica de interesse, antes do modelo ser ajustado (abordagem aleatória) e é dada por:

$$p = P(D_+) = \frac{a + c}{a + b + c + d}. \tag{4.5}$$

Definição 2 (Sensibilidade): A sensibilidade é definida como a proporção de verdadeiros positivos entre todos os portadores da característica, e expressa a probabilidade do modelo sob investigação fornecer um resultado positivo dado que a observação tem a característica. Isto é, a sensibilidade corresponde a proporção de observações que apresentam a característica de interesse que são classificadas corretamente através de um modelo qualquer por ter um *score* superior a P_c. Considerando a Tabela 4.1, a sensibilidade é dada por:

$$S = P(T_+|D_+) = \frac{VP}{VP + FN} = \frac{a}{a + c}. \tag{4.6}$$

Definição 3 (Especificidade): A especificidade é definida como a proporção de verdadeiros negativos entre todos os não portadores da característica, e expressa a probabilidade do modelo sob investigação fornecer um resultado negativo dado que a observação está livre da característica. Isto é, a especificidade corresponde a proporção de observações que não apresentam a característica de interesse, classificadas corretamente através de um modelo qualquer por terem *score* menor ou igual à P_c. Considerando a

Tabela 4.1, a especificidade é dada por:

$$E = P(T_-|D_-) = \frac{VN}{VN + FP} = \frac{d}{b + d}. \tag{4.7}$$

Observe que um modelo sensível raramente deixará de diagnosticar observações portadoras da característica de interesse e que um modelo muito específico raramente classificará como portadora da característica uma observação sem a mesma.

Uma outra medida da capacidade preditiva do modelo é dada pela acurácia.

Definição 4 (Acurácia): A acurácia é definida como a proporção de acertos de um modelo, tanto positivos quanto negativos, ou seja, é a proporção de verdadeiros positivos e verdadeiros negativos em relação a todos os resultados possíveis. A acurácia é também denominada capacidade de acerto total do modelo (CAT). Considerando a Tabela 4.1, acurácia é dada por:

$$CAT = \frac{VP + VN}{VP + FP + VN + FN} = \frac{a + d}{a + b + c + d}. \tag{4.8}$$

Observe que a acurácia também pode ser vista como a média ponderada da sensibilidade e especificidade em relação ao número de observações que apresentam e não apresentam a característica de interesse de uma população. Ressaltamos que a acurácia não é a melhor medida para orientar a escolha de um modelo, pois é influenciada pela sensibilidade, especificidade e prevalência. Além disso, dois modelos com sensibilidades e especificidades diferentes podem fornecer valores semelhantes de acurácia se forem aplicados a populações com prevalências muito diferentes. Muitas vezes a escolha de um modelo privilegia grande sensibilidade ou grande especificidade, dependendo dos objetivos de uso do mesmo.

Em relação ao efeito da prevalência sobre a acurácia, suponha por exemplo que a prevalência seja de 10%, então, se um determinado modelo classificar todos os indivíduos com não portadores da característica teremos acerto em 90%, ou seja, uma acurácia alta para um modelo não informativo. Dessa forma devemos tomar muito cuidado com esta medida.

Considerando a Tabela 4.1, quando do resultado de um modelo, observações portadoras de característica são classificadas como não portadoras, os respectivos resul-

tados errôneos deste modelo são denominados de falsos-negativos (FN); quando clientes não portadores da característica são classificados como portadores os resultados errôneos deste modelo são denominados falsos-positivos (FP).

Outra medida de interesse é a probabilidade da presença da característica após o resultado do modelo, ou seja, qual o valor preditivo positivo do modelo ou a probabilidade da observação ser originalmente portadora da caracetrística dado que o modelo assim o classificou. Por outro lado, também de interesse é a probabilidade da não presença da característica após o resultado do modelo, ou seja, qual o valor preditivo negativo do modelo ou a probabilidade da observação ser originalmente não portadora da característica dado que o modelo assim a classificou. Tais medidas exigem certo cuidado na sua interpretação, pois sofrem o efeito de uma terceira medida, a prevalência. Se a amostra obtida para investigar o desempenho do modelo fornece um estimador viciado da prevalência, os valores preditivos positivo e negativo, estimados por estas relações, não podem ser considerados medidas confiáveis. O efeito que a prevalência exerce sobre estas medidas é amplamente discutido por Rosenquist (1989) e Altman (1991, p.411). Por este motivo, as medidas de sensibilidade (SE) e especificidade (ES) são mais utilizadas uma vez que estas medidas não são afetadas pela prevalência (Altman, 1991, p.411).

Em estudos de avaliação de modelos de classificação é sempre importante buscar medidas livres do efeito da prevalência, ou então, sempre identificar como a prevalência pode interferir nas estimativas das medidas encontradas. As conseqüências do uso de estimativas errôneas a priori da prevalência são discutidas por Griner *et. al.* (1981). Estas medidas são definidas a seguir.

Definição 5 (Valor Preditivo Positivo): O valor preditivo positivo (VPP) do modelo é definido como a proporção de verdadeiros positivos entre todas as observações classificadas pelo modelo como positiva para a característica em estudo, ou seja, é a proporção de observações portadoras da característica, dado que o modelo as apontou como portadoras. Considerando a Tabela 4.1 o VPP é dado por:

$$VPP = P(D_+|T_+) = \frac{VP}{VP + FP} = \frac{a}{a+b}. \tag{4.9}$$

Baseado em estimativas confiáveis de sensibilidade e especificidade, o valor preditivo positivo (VPP) pode ser estimado usando o teorema de Bayes e uma estimativa a priori da prevalência (Linnet, 1988):

$$VPP = \frac{Sensibilidade \times Prevalencia}{Sensibilidade \times Prevalencia + (1 - Especificidade) \times (1 - Prevalencia)}. \tag{4.10}$$

Definição 6 (Valor Preditivo Negativo): *o valor preditivo negativo* (VPN) do modelo pode ser visto como a poporção de verdadeiros negativos entre todas as observações classificadas pelo modelo como negativa, ou seja, é definido como a proporção de observações não portadoras da característica, dado que o modelo as apontou como não portadoras para a característica em estudo. Considerando a Tabela 4.1 o VPN é dado por:

$$VPN = P(D_-|T_-) = \frac{VN}{VN + FN} = \frac{d}{c + d}. \tag{4.11}$$

Baseado em estimativas confiáveis de sensibilidade e especificidade, o valor preditivo negativo (VPN) pode ser estimado usando o teorema de Bayes e uma estimativa a priori da prevalência (Linnet, 1988):

$$VPP = \frac{Especificidade \times (1 - Prevalencia)}{Especificidade \times (1 - Prevalencia) + Sensibilidade \times Prevalencia}. \tag{4.12}$$

Quanto mais sensível o modelo, maior seu valor preditivo negativo, isto é, maior é a segurança de que uma observação com resultado negativo não seja não portadora da característica, e quanto mais específico o modelo, maior é o seu valor preditivo positivo, ou seja, maior é a segurança de que uma observação com resultado positivo seja originalmente portadora da característica.

Uma medida de desempenho que pode ser utilizada mesmo no caso de prevalências extremas é o Coeficiente de Correlação de Mattew (Mattew Coeficient Correlation - MCC, Matthew, 1975). Um coeficiente que deriva do Coeficiente de Correlação de Pearson (CCP) (Baldi, 2000), dado por

$$CCP = \sum_i \frac{(d_i - \bar{d})(t_i - \bar{t})}{\sigma_d \sigma_t}. \tag{4.13}$$

Definição 7 (Coeficiente de correlação de Matthew): O Coeficiente de Correlação de Matthew mede o quanto a D e T, após serem padronizadas, tendem a

apresentar o mesmo sinal e magnitude (Baldi, 2000) e como coeficiente de correlação o MCC pode assumir valores entre -1 e 1, sendo que 1 indica completa associação positiva, ou seja perfeita classificação, 0 indica a classificação aleatória e -1 indica classificação completamente inversa. Temos que $\bar{d} = (VP+FN)/(a+b+c+d)$ e $\bar{t} = (VP+FP)/(a+b+c+d)$, desenvolvendo a equação 4.13 algebricamente para o caso de modelos de classificação obtemos a equação 4.14 que depende de quatro medidas contidas na matriz de confusão, os verdadeiros e falsos positivos e negativos. Considerando a Tabela 3.1 o MCC é dado por:

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}}. \tag{4.14}$$

4.5 Comentários Finais

Vimos no Capítulo 4 diferentes métodos e medidas para avaliar e validar modelos de *score* e de classificação. Discutimos a estatística KS, a curva ROC, a análise de decil, o *lift chart*, a tabela de ganhos e, por fim, as medidas de desempenho que são utilizadas nos modelos de classificação. A seguir, é apresentado o estudo de simulação que teve como objetivo comparar o desempenho entre os MLG's e os MAG's para o caso de resposta binária em diferentes cenários de desbalanceamento.

Capítulo 5

Estudo de simulação

Os MAG são ainda relativamente pouco utilizados, e poucos estudos foram desenvolvidos para avaliar seu desempenho como ferramenta de modelagem preditiva e/ou compará-lo com o desempenho de técnicas mais comuns de modelagem. O estudo de simulação que segue tem como objetivo comparar os MLG com os MAG em diferentes cenários de desbalanceamento de resposta binária.

5.1 Especificações Gerais do Estudo de Simulação

Para comparar os modelos, foram geradas 50 amostras com uma variável de interesse (Y) e duas covariáveis (X_1 e X_2) de tamanho 1000 para cada uma das seguintes proporções do evento de interesse: 1%, 10%, 25% e 50%.

Após a definição da variável de interesse (0 e 1) foram gerados valores para as covariáveis através da simulação de uma distribuição bivariada, sendo que para o grupo em que há ocorrência do evento foram gerados valores de uma distribuição $Normal_2(1/\sqrt{2}, I)$, onde I representa a matriz identidade, e para o grupo em que não há ocorrência do evento foram gerados valores de uma distribuição $Normal_2(0, 4I)$. Esta simulação segue o mesmo principio da simulação *ringnorm* utilizada por Breiman (1998) tendo como principal característica o fato de que a superfície de separação da variável de interesse é uma esfera.

No entanto, devido a não convergência em alguns casos utilizando o MAG, causada pelos *outliers* gerados em X_1 e X_2, optou-se por trocar todas as informações abaixo

do percentil de 10% pelo valor P_{10} (percentil 10%) e todas as informações acima do percentil de 90% pelo valor P_{90} (percentil 90%), a fim de resolver o problema da não convergência.

5.2 Análises

Para cada conjunto de dados é retirada uma amostra sem reposição contendo 70% dos dados originais, a amostra desenvolvimento. Os outros 30% restantes serão utilizados para validar o modelo, a amostra validação.

Um modelo aditivo generalizado com 4 graus de liberdade e um modelo linear generalizado, ambos com função de ligação logito, foram aplicados à amostra desenvolvimento, obtendo as estimativas dos parâmetros para cada um dos modelos. Foi feito, então, o agrupamento por decil de acordo com os valores preditos por cada modelo, definindo assim as faixas de *score* em cada caso.

A validação dos modelos foi realizada aplicando-se na amostra validação (30%) o modelo obtido pela amostra desenvolvimento, em cada caso. As faixas de *score* são definidas pelos decis obtidos na amostra desenvolvimento, e então são calculados os valores do *lift* para cada faixa. Estes processos foram realizados para cada uma das 50 amostras de dados simuladas. As comparações finais entre os modelos foram realizadas baseado-se nos KS's e *lifts* médios por decil.

Começamos a análise observando em cada uma das cinqüenta amostras simuladas, para cada uma das quatro proporções de evento estudadas, a quantidade de vezes que as variáveis X_1 e X_2 explicaram significativamente a resposta. A Tabela 5.1 mostra que no modelo logístico (MLG), para a proporção de evento 1%, em nenhuma das cinqüenta bases as variáveis X_1 e X_2 foram significativas, ao contrário do modelo aditivo, que teve significância nas duas variáveis em todas os cinqüenta modelos, das cinqüenta bases simuladas.

Ainda na mesma tabela, observamos que a número de vezes que as variáveis foram significativas nos modelos logísticos aumenta conforme aumenta a proporção de evento, sendo sete vezes na proporção de 10%, trinta na de 25% e quarenta e quatro na de 50%. Já

Tabela 5.1: Número de amostras em que cada variável foi significativa a 5%

Amostras	Modelo	X1	X2
Com 1% sucesso	MLG	0	0
	MAG	50	50
Com 10% sucesso	MLG	7	7
	MAG	50	49
Com 25% sucesso	MLG	30	30
	MAG	50	50
Com 50% sucesso	MLG	44	44
	MAG	50	50

o MAG apresentou significância em todas as variáveis em praticamente todos os modelos para todas as proporções de sucesso analisadas.

Vale ressaltar que para uma variável de natureza contínua ser trabalhada em um modelo logístico, ela deve apresentar uma estrutura de relação que seja linear com a resposta (Liu, 2007). Já nos MAG, são ajustadas funções suaves que expliquem a relação entre essas variáveis, caso não exista essa relação linear. Como vimos, as duas variáveis contínuas simuladas apresentaram significância no MAG em praticamente todas as bases simuladas, ao contrário do MLG, logo é de se esperar que os KS's dos MAG também sejam maiores, como observamos na Tabela 5.2. Para as quatro proporções analisadas, o KS do MAG foi maior que do MLG tanto no desenvolvimento quanto na validação.

Tabela 5.2: Comparação de KS's

Desenvolvimento								
Proporção de Eventos	1%		10%		25%		50%	
Modelo	MLG	MAG	MLG	MAG	MLG	MAG	MLG	MAG
KS	0,46	0,75	0,37	0,56	0,36	0,54	0,36	0,53
Validação								
Proporção de Eventos	1%		10%		25%		50%	
Modelo	MLG	MAG	MLG	MAG	MLG	MAG	MLG	MAG
KS	0,49	0,64	0,39	0,56	0,36	0,54	0,37	0,55

Desenvolvendo a análise de decil e calculando o *lift* médio, para o desenvolvimento

e validação, de cada uma das cinqüenta bases de cada uma das quatro proporções de evento analisadas, observamos nas Tabelas 5.3 e 5.4 que o MAG está sendo capaz de ordenar o evento em praticamente todas as proporções, tanto no desenvolvimento quanto na validação. Por outro lado, o modelo logístico mostra um desempenho inferior, sendo incapaz de ordenar o evento em praticamente todas as proporções de evento analisadas, tanto no desenvolvimento quanto na validação. A análise de desempenho dos modelos de classificação não foi aplicada ao estudo de simulação devido à complexidade de se aplicar o ponto de corte adequado a cada amostra em cada modelo de regressão.

Tabela 5.3: Média dos *Lifs* - Amostra Desenvolvimento

	1%		10%		25%		50%	
Decil	MAG	MLG	MAG	MLG	MAG	MLG	MAG	MLG
0	641,99	18,81	340,17	13,01	245,20	20,89	165,26	43,94
1	205,22	149,42	232,03	153,36	211,70	151,14	159,49	137,70
2	84,52	264,75	180,73	214,40	174,33	188,04	149,39	145,82
3	48,35	242,35	125,75	212,57	149,07	185,28	142,51	146,89
4	4,08	153,86	73,66	176,22	112,79	168,45	132,30	142,63
5	4,59	108,43	33,30	133,78	65,86	135,46	111,86	132,28
6	0,00	44,58	15,39	71,20	28,18	99,19	80,49	118,95
7	0,00	16,03	3,52	22,17	14,17	42,80	43,60	95,94
8	0,00	2,04	0,00	3,17	1,86	8,72	16,83	35,56
9	8,40	0,00	0,30	0,29	0,34	0,34	0,63	0,90

Por fim, vimos através desse estudo de simulação, que o MAG, no caso de variáveis preditoras de natureza contínua e resposta binária, se comporta melhor do que o MLG. As covariáveis entram com uma significância maior e em geral o desempenho se mostra melhor. Além disso, esses modelos mostraram robustos ao aumento no desbalanceamento da variável de interesse mantendo bom desempenho até mesmo nos caso de taxa de evento de apenas 1%.

Tabela 5.4: Média dos *Lifs* - Amostra Validação

	1%		10%		25%		50%	
Decil	MAG	MLG	MAG	MLG	MAG	MLG	MAG	MLG
0	196,58	11,09	299,09	14,83	232,63	25,34	162,71	37,65
1	247,68	119,78	236,33	135,43	208,05	144,54	158,84	136,17
2	207,63	206,60	198,14	236,87	176,55	181,84	148,02	147,49
3	208,51	161,69	135,34	203,39	157,09	191,32	142,26	145,85
4	70,09	152,45	73,22	171,57	108,61	162,37	130,73	141,09
5	32,27	231,60	39,64	147,84	65,26	133,82	110,64	132,58
6	58,91	94,51	17,87	79,08	29,29	103,18	77,01	122,81
7	0,00	32,01	7,50	21,30	18,67	47,42	44,36	92,77
8	34,80	5,54	2,25	3,97	3,98	9,46	19,11	40,70
9	0,00	0,00	0,00	0,00	0,24	0,24	2,24	1,39

Capítulo 6

Estudo de um problema de *credit score*

Este capítulo tem como objetivo aplicar e comparar as técnicas apresentadas em um conjunto de dados reais para um problema de *credit score*, a partir de uma base de uma instituição financeira que atua no mercado varejista brasileiro há mais de vinte anos.

A base de dados utilizada neste trabalho é constituída de informações de uma instituição financeira onde os clientes adquiriram um determinado produto de crédito. Trata-se da etapa de concessão de crédito, o que nos impossibilita de utilizar variáveis históricas e comportamentais, já que o futuro cliente ainda não está nas bases da instituição. Devido ao baixo número e qualidade das informações, esses modelos apresentam desempenho preditivos inferiores aos modelos de *behavior*, por exemplo. Para a instituição, o objetivo desses modelos é a previsão da ocorrência ou não da inadimplência para cada cliente, auxiliando assim na tomada de decisão. Parte-se de uma base aleatória de clientes com um determinado produto de crédito e, a partir do momento da concessão, é observada em um determinado período de tempo a ocorrência ou não da inadimplência, caracterizando então o evento de interesse. Assim, definimos um modelo de *credit score*.

O modelo foi ajustado via Regressão Logística (MLG) e via Modelos Aditivos (MAG). Conforme vimos no Capítulo 3, o MAG tem como vantagem expressar a relação de uma covariável contínua com a resposta, mesmo que essa relação não seja linear. Sendo assim, para que pudéssemos ter uma comparação justa, em um primeiro momento ajustamos ambos os modelos (MLG e MAG) sem categorizar as covariáveis contínuas.

Após essa comparação, veremos na seção 6.2.1 todos os passos e técnicas utilizadas

no tratamento de covariáveis contínuas para o desenvolvimento de modelos com respostas binárias, e por fim esses resultados também são comparados.

Apresentamos, na seção 6.1, uma descrição dos dados e a aplicação dos modelos (MLG e MAG) sem a categorização de variáveis contínuas. Na seção 6.2 descrevemos as etapas necessárias para o processo de modelagem, como a análise bivariada, a identificação e eliminação de multicolinearidade, a análise de desempenho e validação de modelos. Na mesma seção ajustamos o modelo Logístico (MLG) a partir das técnicas apresentadas e, por fim, ajustamos o Modelo Logístico Aditivo (MAG) na seção 6.3, a partir das mesmas técnicas descritas em 6.2.

6.1 Descrição dos dados

As variáveis disponíveis na base de dados foram coletadas no momento da solicitação de crédito, após 12 meses de observação de comportamento do cliente marcamos a variável de interesse a qual indica se houve inadimplência nos 12 meses seguintes ao momento da concessão. A base possui 7318 clientes, sendo que, em aproximadamente 30% dos casos, foi observada a ocorrência da inadimplência. A base de dados de desenvolvimento possui 5123 clientes e a base de validação possui 2195, ambos com aproximadamente a mesma proporção de inadimplentes, 30% e 31% respectivamente. As variáveis que compõem a base de dados estão descritas na Tabela 6.1.

Tabela 6.1: Descrição das variáveis.

Variável	Descrição
Tpclient	Indica se é cliente há mais de um ano (dicotômica)
Tpempreg	Tempo que o cliente está no atual emprego (em meses)
Sexo	Sexo (Feminino, Masculino)
Estcivil	Estado civil (Casado, Solteiro, Divorciado, Outros)
Sitresid	Condição da residência: própria ou alugada
Limite	Valor do crédito concedido (em reais)
Tempores	Tempo de residência na casa atual (em anos)
CEP	Os dois primeiros dígitos do CEP
Idade	Idade do cliente (em anos completos no dia 31/12/2002)
Regiao	Região de residência (varia de 1 a 18)
Localiza	Código de localização da residencia
Mau	**Indicador de mau pagador (0 - adimplente, 1 - inadimplente)**

A variável "Mau" é a variável resposta binária que será considerada neste estudo e as outras estarão como covariável na análise. Na Tabela 6.2 apresentamos os principais quantis referentes a variável limite e na Figura 6.1 o histograma da variável. A partir da Tabela 6.2 podemos perceber que 1% da população em estudo apresenta limite acima de R$574,00, sendo que o limite máximo é de R$99000,00, tais pontos discrepantes justificam o histograma (Figura 6.1) o qual apresenta apenas uma barra contendo quase 100% centrada no ponto zero. Esses *outliers* podem vir a prejudicar a estimação do modelo caso a relação entre a covariável e a resposta não se mantenha para tais valores, neste caso esses pontos são considerados influentes (Hosmer & Lemeshow, 1989). Para simplificar a análise optamos por substituir os *outliers* por 574, que é o valor do 99º pecentil da variável limite, a distribuição da nova variável, limite99, pode ser vista no histograma localizado no painel superior direito da Figura 6.2.

Tabela 6.2: Quantis da variável limite

Quantil	100%	99%	95%	90%	75%	50%	25%	10%	5%	1%	0%
Valor	99865	574	276	202	118	66	31	5	1	0	0

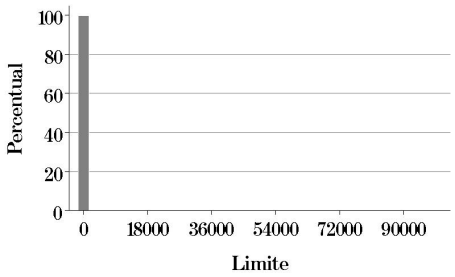

Figura 6.1: Histograma do Limite

Os histogramas referentes as covariáveis contínuas de entrada na etapa de definição dos modelos (MLG logístico e MAG logístico) são apresentados na Figura 6.2.

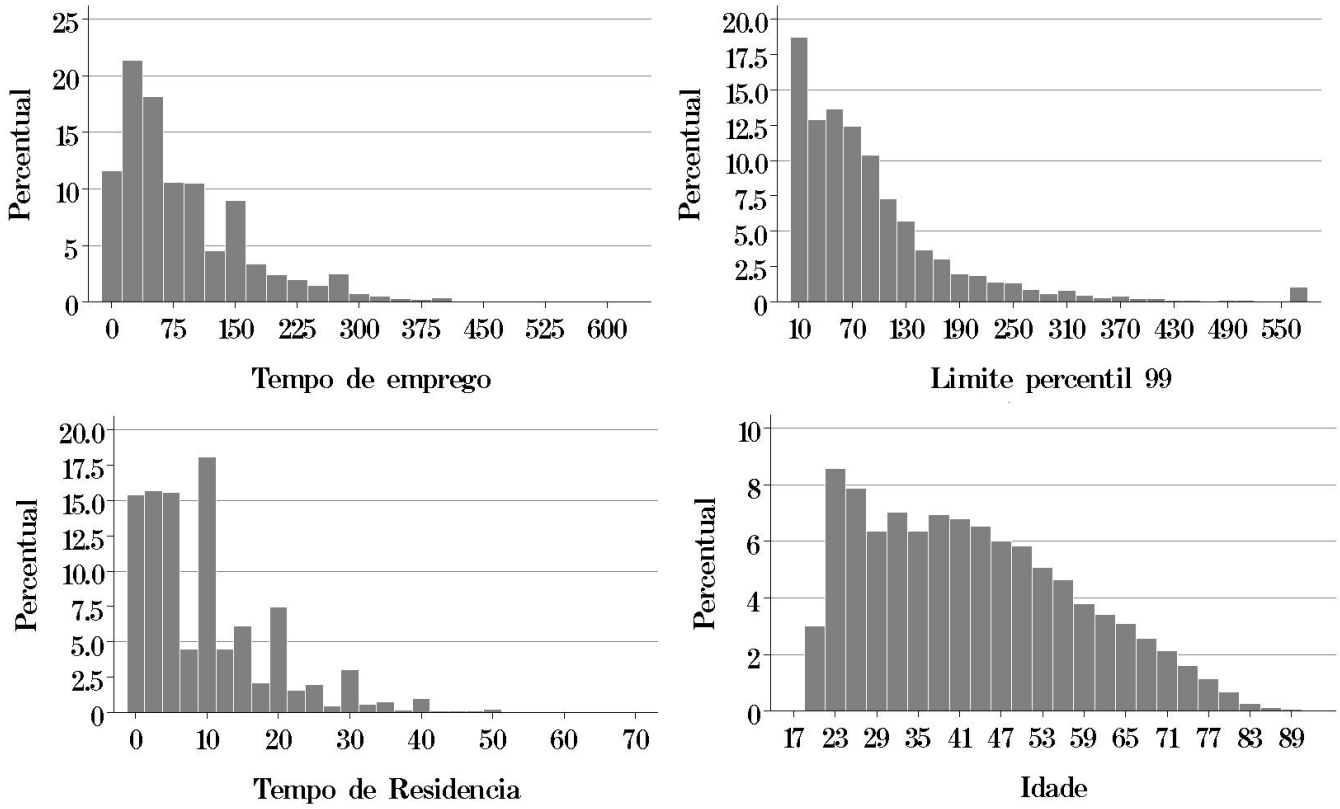

Figura 6.2: Histograma variáveis contínuas

A Tabela 6.3 apresenta as frequências cruzadas entre as variáveis categóricas e a variável de interesse "Mau". Devido ao número elevado de categorias as variáveis de região, CEP e profissão apresentadas na tabela são resultado de um processo de recategorização definido conforme a proposta da seção 6.2.1.

Antes do ajuste dos modelos, a base foi dividida em duas partes, uma utilizada para o ajuste do modelo e outra para a validação, contendo respectivamente 70% e 30%

Tabela 6.3: Tabela de frequências variáveis categóricas

		Mau Pagador		
		Não	Sim	Total
Mau		5091	2227	7318
Tpclient	1	1658	1006	2664
	2	3433	1221	4654
Sexo	F	3054	1172	4226
	M	2037	1055	3092
Estcivil	C	2807	1055	3862
	D	180	77	257
	S	1608	936	2544
	V	494	158	652
Sitresid	A	382	224	606
	P	4678	1973	6651
Regiaor	c1	2859	1044	3903
	c2	652	306	958
	c3	1307	697	2004
	c4	273	180	453
CEPr	c1	2910	1017	3927
	c2	2039	1080	3119
	c3	134	119	253
	c4	8	11	19
Profissaor	c1	2523	1005	3528
	c2	2568	1222	3790
Localiza2	c1	766	117	883
	c2	1076	291	1367
	c3	1228	466	1694
	c4	0	10	10
	c5	1601	881	2482
	c6	172	141	313
	c7	155	164	319
	c8	61	85	146
	c9	32	72	104

da base original. É interessante, se for possível, trabalhar com bases de validação que não pertençam ao mesmo período ou universo da base de desenvolvimento, avaliando e buscando garantir a robustez e a perenidade do modelo.

Tabela 6.4: Resumo *backward*

Logístico		Aditivo	
Passo	Efeito Removido	Passo	Efeito Removido
1	dm_regiao_2	1	dm_estcivil
2	dm_regiao_1	2	dm_regiao_2
3	dm_estcivil	3	dm_regiao_1
4	dm_regiao_3	4	dm_cep_3
5	dm_prof	5	dm_prof
6	dm_cep_3	6	dm_regiao_3
7	limite99		

Ajustamos então os dois modelos, MLG e MAG, utilizando as covariáveis contínuas na sua forma natural, conforme foram descritas. O ajuste parte, primeiramente, da seleção de variáveis, que pode ser feita a partir de três algoritmos automatizados conhecidos como *stepwise*, *backward* ou *forward* (Hosmer& Lemeshow, 1989). O algoritmo utilizado para a seleção dos modelos foi o *backward*, que parte do modelo completo (com todas as covariáveis) e elimina variáveis a partir da comparação dos desvios dos modelos completo (com a variável) e reduzido (sem a variável).

Tabela 6.5: Estimativa dos parâmetros e teste de significância - MLG

Parâmetro	Estimativa	Wald	Valor - p
Intercepto	0,9479	27,4293	0,0001
Idade	-0,0238	92,0195	0,0001
Tpempreg	-0,00182	14,0263	0,0002
Tempores	-0,0083	5,0551	0,0246
dm_cep_1	-0,808	26,0232	0,0001
dm_cep_2	-0,5085	10,3412	0,0013
dm_tpclient	0,4304	42,8276	0,0001
dm_sexo	-0,1594	6,2641	0,0123
dm_sitresid	0,2512	5,264	0,0218

A Tabela 6.4 mostra o resumo do algoritmo de seleção de variáveis para o MLG (logístico) e para o MAG. Podemos perceber que as variáveis eliminadas foram as mesmas para os dois modelos, exceto pela variável limite99 que não entrou no modelo logístico. Na Tabela 6.5 temos as estimativas dos parâmetros do modelo logístico e, na Tabela 6.6, as estimativas dos parâmetros da parte linear do modelo aditivo. Comparando as duas tabelas percebemos que os parâmetros estimados para a parte linear do modelo aditivo são muito próximos dos estimados pelo modelo logístico. Na Tabela 6.7 temos a estimativa dos parâmetros de alisamento e na Tabela 6.8 a análise de *deviances* contendo a significância das variáveis contínuas.

Os gráficos de alisamento por *splines* para a parte não linear das variáveis contínuas podem ser vistos na Figura 6.3, e nesses gráficos podemos observar qual a relação estimada de cada covariável contínua com a resposta. É interessante notar, que o impacto e o significado de cada variável na chance de inadimplir é diferente para diferentes faixas de valores da mesma. Temos, por exemplo, que para clientes com idade abaixo de 30 anos, maior idade implica em menor chance de inadimplir enquanto que para clientes com mais de 30 anos esta relação se inverte, notamos ainda que para as faixas entre 45 e 65 anos o impacto da variável idade é menor na inadimplência do que para outras faixas etárias. Este tipo de entendimento só é possível, pois os MAG não assumem forma rígida e buscam uma função suave para explicar a estrutura de relacionamento entre cada uma das

covariáveis e a variável de interesse, considerando as demais variáveis constantes. Temos também, por exemplo, que limites acima de 150 reais implicam em impacto positivo na inadimplência enquanto limites abaixo deste valor implicam em impacto negativo.

Tabela 6.6: Estimativa dos parâmetros e teste de significância - MAG

Parâmetro	Estimativa	T	Valor - p
Intercepto	1,00463	0,36	0,7204
dm_cep_1	-0,81457	5,12	0,0001
dm_cep_2	-0,51545	3,25	0,0012
dm_tpclient	0,33648	-4,93	0,0001
dm_sexo	-0,15719	2,45	0,0144
dm_sitresid	0,22634	-2,05	0,0401
Linear(Idade)	-0,02336	-9,56	0,0001
Linear(Limite99)	-0,0005	-1,4	0,1607
Linear(Tempores)	-0,00775	-2,12	0,0338
Linear(Tpempreg)	-0,00162	-3,44	0,0006

Temos que o maior índice de correlação entre as convariáveis selecionadas, em ambos os modelos, foi de 0,26 e, sendo assim, todas as variáveis foram mantidas no modelo (Greene, 2003). A estatística KS para o MLG foi de 0,24 tanto na amostra de desenvolvimento quanto na de validação, para o MAG o valor da estatística foi de 0,26 na amostra de desenvolvimento e 0,24 na de validação.

Tabela 6.7: Estimativa dos Parâmetros - Parte Não Linear

Componente	Parâmetro de Alisamento	GL	Número de Observações unicas
Spline(Idade)	0,999975	4	73
Spline(Limite99)	1	4	393
Spline(Tempores)	0,999953	4	60
Spline(Tpempreg)	1	4	351

Observe na Figura 6.4 que, para ambos os modelos, tanto no desenvolvimento como na validação, temos um modelo aderente, já que a probabilidade estimada é muito

Tabela 6.8: Análise de Deviance

	DF	Soma de Quadrados	Qui-Quadrado	Valor - p
Spline(idade)	4	12,20962	10,5459	0,0322
Spline(limite99)	4	10,86464	9,3842	0,0522
Spline(TEMPORES)	4	5,732088	4,951	0,2924
Spline(TPEMPREG)	4	7,473108	6,4548	0,1677

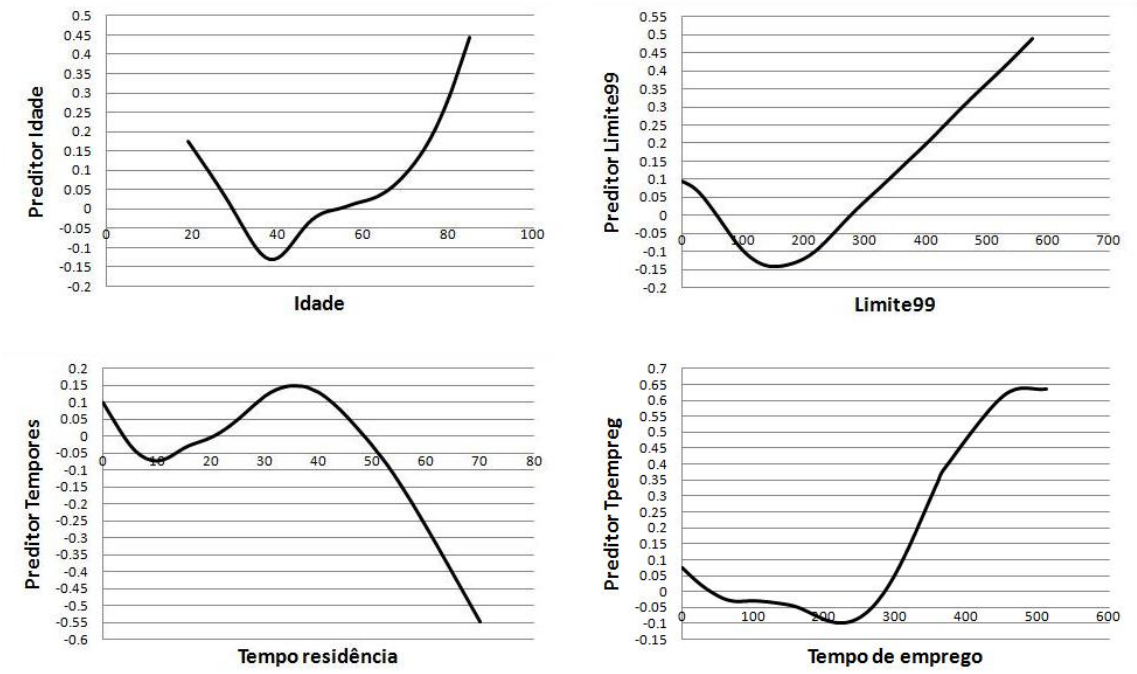

Figura 6.3: Parte não linear das variáveis contínuas

próxima da observada e ainda os gráficos de aderência apresentam monotonicidade ligeiramente maior no caso do MAG (painéis inferiores). Temos também, na Figura 6.5, que o *Lift Chart* é decrescente em ambos os modelos, tanto no desenvolvimento como na validação, sendo que a primeira faixa de *score* é ligeiramente maior no caso das duas amostras (desenvolvimento e validação) do MAG.

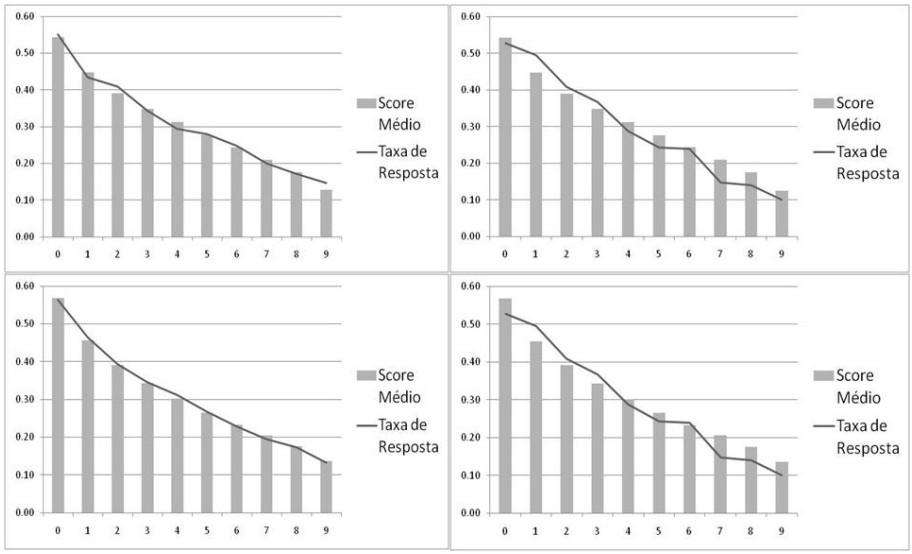

Figura 6.4: Gráficos de Aderência. Os gráficos de cima são do MLG (desenvolvimento e validação, da esquerda para a direita) e os de baixo são do MAG (desenvolvimento e validação, da esquerda para a direita)

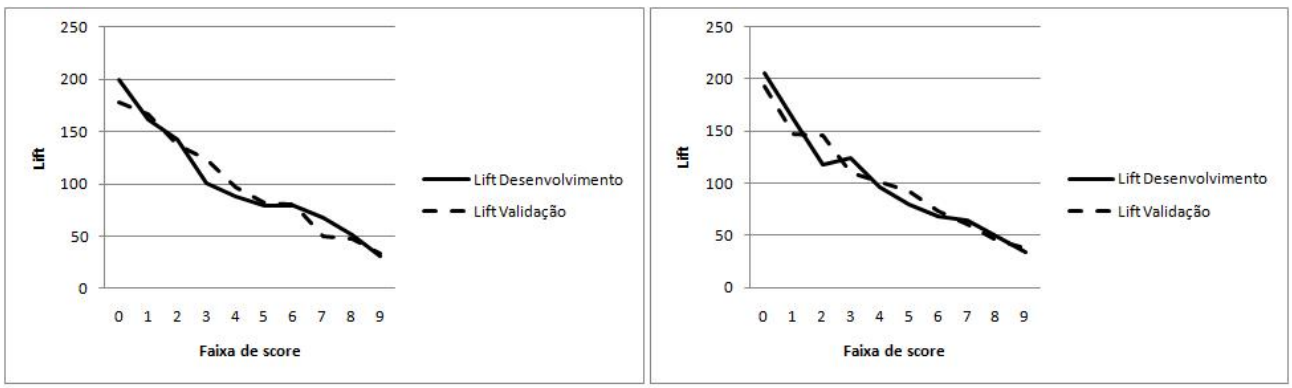

Figura 6.5: *Lift Charts*. Da esquerda para a direita, os gráficos do MLG e do MAG.

Na Figura 6.6 temos os gráficos de ganhos e observamos que, em ambos os modelos, tanto no desenvolvimento como na validação, captamos cerca de 70% da inadimplência com 50% da base.

Já no painel da Figura 6.7, estimamos intervalos de confiança para o *lift* na base de

desenvolvimento e verificamos se os *lift's* estimados na validação estavam contidos nesses intervalos. Vemos que no modelo Logístico (MLG) o *lift* da faixa de score 4 é ligeiramente mais alto que o limite superior do intervalo de confiança estimado no desenvolvimento. Já no MAG, os lifts estão sempre contidos nos intervalos estimados.

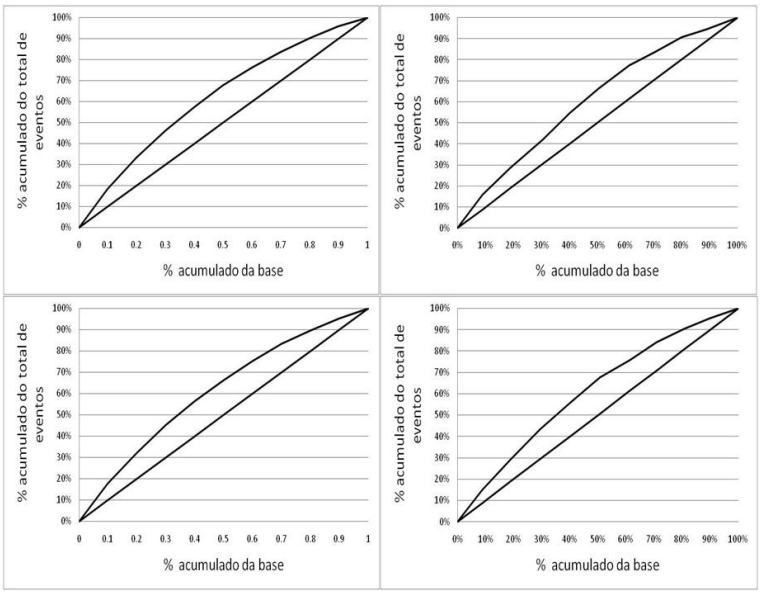

Figura 6.6: Gráficos de Ganhos. Os gráficos de cima são do MLG (desenvolvimento e validação, da esquerda para a direita) e os de baixo são do MAG (desenvolvimento e validação, da esquerda para a direita)

Na Figura 6.8 temos a curva ROC para ambos os modelos, sobrepostas. Podemos observar que o MAG apresenta pontos ligeiramente melhores que o MLG. Consequentemente vemos na Tabela 6.9 que o MAG apresenta medidas de desempenho também ligeiramente melhores que o MLG, pelo menos quando tratamos da base de desenvolvimento. Já na base de validação, vemos que o MLG se mostrou mais sensível e com um Valor Preditivo Negativo ligeiramente maior.

Por fim, pudemos verificar como se comportam os modelos MLG e MAG nas mesmas condições, isto é, mesmo tratamento de covariáveis (contínuas e categóricas) e mesma base de dados. Veremos então, nas próximas seções, como serão os resultados desses modelos na mesma base de dados, porém com tratamento de covariáveis, isto é, as covariáveis foram tratadas da forma que se sugere na literatura, de acordo com as referências que veremos ao longo do texto.

No entanto, como já mencionado as variáveis contínuas só devem ser utilizadas no

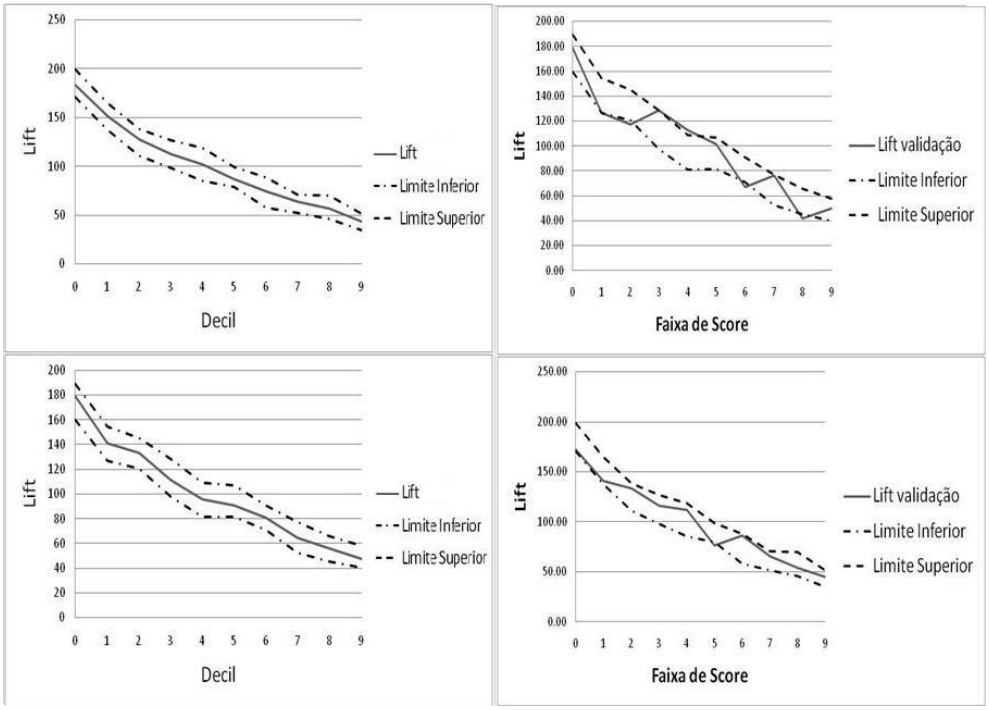

Figura 6.7: *Lift Charts* com intervalos de confiança bootstrap. Os gráficos de cima são do MLG (desenvolvimento e validação, da esquerda para a direita) e os de baixo são do MAG (desenvolvimento e validação, da esquerda para a direita)

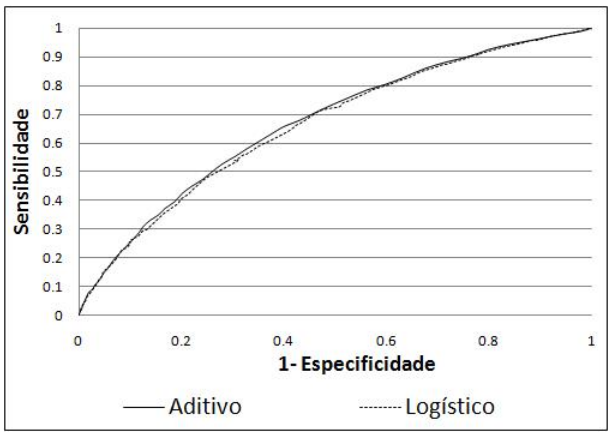

Figura 6.8: Curva ROC. MLG e MAG.

Tabela 6.9: Tabela Medidas de Desempenho

	Logístico		Aditivo	
Medida	Desempenho	Validação	Desempenho	Validação
SENSIBILIDADE	0.62	0.70	0.64	0.62
ESPECIFICIDADE	0.62	0.52	0.62	0.60
VPP	0.42	0.38	0.43	0.40
VPN	0.79	0.80	0.79	0.79
CAT	0.62	0.58	0.63	0.61
MCC	0.22	0.20	0.24	0.21

modelo de regressão logística quando o preditor da covariável apresenta relação linear com a variável resposta. Usualmente os modelos de regressão logística passam por um processo pré modelagem de análise descritiva, que busca verificar descritivamente se a suposição de linearidade do preditor realmente é observada, caso não seja tais covariáveis passam por um processo de categorização. Considerando a forma usual de se ajustar modelos, apresentaremos nas próximas seções o ajuste dos modelos MAG e MLG utilizando o procedimento mais utilizado para se obter um modelo de maior capacidade preditiva.

6.2 Aplicação do Modelo Linear Generalizado para Dados Binários

Conforme o processo de modelagem citado na seção anterior, a base foi dividida em duas partes, uma utilizada para o ajuste do modelo e outra para a validação, contendo respectivamente 70% e 30% da base original. Com já discutimos, é interessante, se for possível, trabalhar com bases de validação que não pertençam ao mesmo período ou universo da base de desenvolvimento, avaliando e buscando garantir a robustez e a perenidade do modelo.

Novamente o ajuste partiu da seleção de variáveis *backward* (Hosmer & Lemeshow, 2000). É necessário verificar se há correlação entre as covariáveis que entraram no modelo, de forma a evitar multicolinearidade (Greene, 2003). Algumas covariáveis categóricas foram recategorizadas a partir de uma análise bivariada entre a covariável e

a resposta, de acordo com o *lift* de suas categorias originais. Já as covariáveis contínuas, quando necessário, foram categorizadas de acordo com o *lift* de seus vintis. Dessa forma todas as covariáveis foram transformadas em *dummys*.

6.2.1 Análise Bivariada

Como dito anteriormente, uma etapa necessária e fundamental para o processo de modelagem é a análise bivariada, no caso dos MLG ela nos auxilia na verificação da suposição de linearidade da estrutura de relacionamento entre uma covariável contínua e a variável resposta, caso esta relação linear não exista a covariável em questão deverá ser categorizada assim como citam Gruenstein (1998) e Thomas (2002). Trata-se de uma análise descritiva das variáveis por tabelas de contingência, de forma bivariada, entre as covariáveis e a variável resposta. Esta análise nos permite identificar e conhecer as relações iniciais existentes e nos será útil posteriormente, ao analisarmos os parâmetros do modelo e sua interpretação.

A partir dessa análise que definimos possíveis categorizações de variáveis contínuas e recategorizações de variáveis categóricas. O modelo logístico (MLG) supõe que o efeito das variáveis contínuas na variável resposta tenha uma estrutura linear, dessa forma antes da inclusão de cada variável contínua ao modelo, devemos verificar a presença ou não desta estrutura linear que, caso não exista, implicará na categorização da mesma. Nesses casos, a análise bivariada parte de uma quebra das variáveis contínuas em vintis, isso é, vinte categorias de tamanhos iguais, cruzada com a variável resposta. A partir disso, podemos calcular o *lift*, que é um incremento em relação a média, em outras palavras, o quanto os clientes daquela categoria inadimplem mais do que a média geral.

A partir essa tabela bivariada, podemos observar se existe a suposta relação linear entre a covariável e a resposta, para isso basta observar se os valores do *lift* crescem ou decrescem ao aumentarem os vintis. Se essa relação linear for inexistente, devemos criar categorias para essa variável contínua, agrupando vintis que apresentem valores próximos para o *lift*, criando grupos de tamanhos razoáveis entre si. Na Tabela 6.10 temos, no painel direito, um caso de relação linear, e no painel esquerdo um caso de relação não linear.

Tabela 6.10: Exemplo fictício de estrutura de relacionamento entre uma variável continua e a variável resposta.

Grupo	Resposta		Total	Lift	Grupo	Resposta		Total	Lift
	0	1				0	1		
1	207	230	437	1,73	1	234	193	427	1,49
2	227	186	413	1,48	2	233	189	422	1,47
3	216	169	385	1,44	3	159	131	290	1,48
4	203	144	347	1,36	4	237	156	393	1,30
5	181	124	309	1,32	5	230	140	370	1,24
6	223	115	338	1,12	6	320	163	483	2,11
7	308	157	469	1,10	7	136	65	201	1,96
8	250	101	351	0,95	8	250	136	386	2,16
9	244	85	329	0,85	9	284	126	410	1,89
10	234	78	315	0,81	10	246	84	330	0,84
11	370	129	519	0,82	11	225	81	306	0,87
12	212	68	285	0,78	12	275	108	383	0,93
13	202	70	296	0,78	13	364	130	494	0,86
14	319	94	427	0,72	14	183	71	254	0,92
15	294	78	372	0,69	15	274	86	360	1,38
16	263	69	340	0,67	16	291	71	362	1,64
17	291	73	364	0,66	17	291	90	381	1,48
18	267	53	320	0,54	18	276	75	351	1,70
19	342	70	416	0,55	19	296	65	361	0,59
20	238	48	286	0,55	20	287	67	354	0,62
Total	5091	2227	7318	1,00	Total	5091	2227	7318	1,00

De acordo com a Tabela 6.10 temos que a variável à que o painel esquerdo se refere pode ser utilizada em sua forma original, no entanto a variável à que o painel direito se refere não deve ser utilizada diretamente num modelo em que há suposição de linearidade, podemos então categorizá-la. No caso da variável do painel direito, uma forma de categorização é apresentada na Tabela 6.11 Note que a categorização varia de analista para analista, no entanto, o mesmo deve fazê-la baseado nos critérios de parcimônia e homogeneidade.

As variáveis categóricas não apresentam o problema referente à linearidade, no entanto, devemos garantir que nenhuma das categorias de uma covariável apresente freqüência 0 ou muito próxima de zero em cada uma das categorias da variável resposta. No caso da ocorrência desta de baixas freqüências nas categorias o algoritmo iterativo para a estimação dos parâmetros do modelo não convergirá (Hosmer & Lemeshow, 1989).

Algumas vezes ocorre que diferentes categorias de uma covariável apresentam um efeito muito similar na variável resposta, dessa forma torna-se interessante agrupar tais variáveis. As recategorizações das variáveis discretas são feitas a partir de uma tabela de freqüência entre a covariável e a variável resposta, incrementando o *lift*, exatamente como é feito para os vintis das variáveis contínuas. Dessa forma agrupamos as categorias com *lifts* próximos, sempre buscando uma homogeneidade no volume dos grupos. Observe na Tabela 6.12 a recategorização da variável localização, que apresentava 18 categorias.

Tabela 6.11: Exemplo fictício de categorização de variável contínua

Categoria	Grupo	Resposta		Total	Lift
		0	1		
C1	1	234	193	427	1,49
C2	2	233	189	422	1,47
	3	159	131	290	1,48
C3	4	237	156	393	1,30
	5	230	140	370	1,24
C4	6	320	163	483	2,11
	7	136	65	201	1,96
	8	250	136	386	2,16
	9	284	126	410	1,89
C5	10	246	84	330	0,84
	11	225	81	306	0,87
	12	275	108	383	0,93
	13	364	130	494	0,86
	14	183	71	254	0,92
C6	15	274	86	360	1,38
	16	291	71	362	1,64
	17	291	90	381	1,48
	18	276	75	351	1,70
C7	19	296	65	361	0,59
	20	287	67	354	0,62
	Total	5091	2227	7318	1,00

Tabela 6.12: Exemplo fictício da recategorização de variáveis discretas

Novas Categorias	Categoria Original	Resposta		Total	Lift
		0	1		
C1	18	64	22	86	0,84
	1	1906	679	2585	0,86
	6	223	84	307	0,90
	4	666	259	925	0,92
C2	17	582	273	855	1,05
	3	70	33	103	1,05
C3	2	264	126	390	1,06
	15	36	18	54	1,10
	13	30	16	46	1,14
	5	494	265	759	1,15
	14	483	272	755	1,18
C4	10	237	145	382	1,25
	16	33	31	64	1,59
	11	3	4	7	1,88
	Total	5091	2227	7318	1,00

É usual ordenar as categorias de acordo com o *lift*, facilitando a visualização de categorias que se relacionam de maneira similar com a resposta. Assim, reduzimos as 18 categorias da variável em 4 novas categorias (3 *dummys*), cujo poder de relação é similar e o volume entre as categorias é maior e mais homogêneo.

Realizada essa análise já temos conhecimento suficiente das relações existentes e já preparamos as variáveis, agora nosso interesse é buscar um modelo que seja capaz de explicar essas relações de forma resumida e esteja apto a prever a ocorrência da inadimplência, servindo de apoio no processo de concessão de crédito.

6.2.2 Ajuste

Realizando a análise bivariada agora para a base de 7318, como no exemplo descrito anteriormente, observamos que nenhuma das variáveis contínuas apresentou relação linear com a variável resposta, e dessa forma todas elas foram categorizadas de acordo com a metodologia descrita na seção anterior. As variáveis discretas foram re-categorizadas segundo os critérios descritos na seção 6.2.1, exceto pelas variáveis que apresentavam, originalmente, duas categorias: sexo, tpclient e sitresid. Foram criadas *dummys* para todas as categorias, criando sempre uma *dummy* a menos que o número de categorias da variável, evitando a multicolinearidade.

Após a etapa de categorização, recategorização e criação de *dummys*, utilizou-se o método de seleção de variáveis *backward* (Hosmer & Lemeshow), na base de desenvolvimento, para buscar o conjunto de variáveis que melhor explique a relação existente com a resposta. O resumo dos passos deste algoritmo para a base em exemplo é apresentado na Tabela 6.13.

Selecionado esse conjunto de covariáveis, devemos analisar a presença de fortes correlações entre as mesmas. Para a base em exemplo observamos que a maior correlação entre covariáveis foi de 0,26, portanto todas as covariáveis selecionadas pelo método *backward* foram mantidas no modelo. Observe na Tabela 6.14 os parâmetros estimados associados a cada variável *dummy*.

Antes de validarmos e analisarmos a capacidade preditiva desse modelo devemos verificar se as estimativas para os parâmetros citados conferem com os valores do *lift* de cada *dummy* desenvolvida, por exemplo, temos que o *lift* associado à *dummy* dm_limite_3 é de 1,28, e observe que o parâmetro associado a essa covariável é de 0,450, e dessa forma podemos confirmar a relação observada na análise bivariada e afirmar que o modelo é capaz de explicar essa relação, isto é, a presença da categoria limite 3 (dm_limite_3 = 1) aumenta a chance de inadimplência em 28% e tem, na presença das outras covariáveis, um peso positivo de 0,450 no modelo de *credit score*.

Tabela 6.13: Resumo Backward

Passo	Efeito Removido
1	dm_regiao_1
2	dm_prof
3	dm_regiao_2
4	dm_regiao_3
5	dm_cep_1
6	dm_cep_2
7	dm_cep_3
8	dm_sitresid

É comum o modelo estimar, por exemplo, um parâmetro negativo para uma *dummy* que observamos um *lift* maior que um. É um caso em que o modelo não está sendo capaz de captar essa relação e isso pode ser resolvido buscando uma interação significativa entre essa *dummy* e alguma outra, ou pode-se ainda buscar uma nova categorização ou recategorização em que essa relação se mostre mais forte.

Tabela 6.14: Estimativas dos Parâmetros e Teste de Significância

Variáveis	Parâmetros estimados	Wald	Valor - p
dm_limite_3	0,450	23,18	<0,0001
dm_idade_1	0,718	30,60	<0,0001
dm_idade_2	0,568	30,36	<0,0001
dm_idade_3	0,539	23,28	<0,0001
dm_idade_5	-0,649	33,03	<0,0001
dm_idade_7	-0,317	11,96	0,0005
dm_tpempreg_1	0,188	4,43	0,0353
dm_localiza_1	-1,186	84,69	<0,0001
dm_localiza_2	-0,645	47,80	<0,0001
dm_localiza_3	-0,372	20,06	<0,0001
dm_localiza_7	0,653	18,82	<0,0001
dm_localiza_8	0,980	21,92	<0,0001
dm_localiza_9	1,199	20,98	<0,0001
dm_tpclient	0,305	20,05	<0,0001

O modelo em exemplo passou por esse processo e necessitou de algumas recategorizações, de forma a buscar uma relação mais forte e evidente. Assim, depois de todo o processo descrito, o modelo final pode ser escrito por:

$$\hat{\pi} = \frac{e^{X\hat{\beta}}}{1 + e^{X\hat{\beta}}} \tag{6.1}$$

onde $\hat{\pi}$ o vetor de probabilidades de inadimplência estimado, X é a matriz de covariáveis, $\hat{\beta}$ é a estimativa dos parâmetros e

$$X\beta = 0,450\text{dm\_limite\_3} + 0,718 * \text{dm\_idade\_1} + 0,568 * \text{dm\_idade\_2} + 0,539 * \text{dm\_idade\_3} - 0,649 * \text{dm\_idade\_5} - 0,317 * \text{dm\_idade\_7} + 0,188 * \text{dm\_tpempreg\_1} - 1,186 * \text{dm\_localiza\_1} - 0,645 * \text{dm\_localiza\_2} - 0,372 * \text{dm\_localiza\_3} + 0,653 * \text{dm\_localiza\_7} + 0,980 * \text{dm\_localiza\_8} + 1,199 * \text{dm\_localiza\_9} + 0,305 * \text{dm\_tpclient}, \quad (6.2)$$

.

Agora que temos um modelo capaz de explicar de forma conjunta e resumida todas as relações entre as covariáveis e a variável resposta, temos que validá-lo e analisar seu desempenho preditivo.

6.2.3 Análise do Desempenho e validação

O modelo descrito na seção anterior, ajustado na amostra desenvolvimento, deve agora ser avaliado quanto sua capacidade preditiva e validado em outra amostra, chamada amostra de validação. Iniciamos a análise pela estatística de Kolmogorov-Smirnov, que foi de 0,30, e observamos a seguir a curva ROC (Figura 6.9).

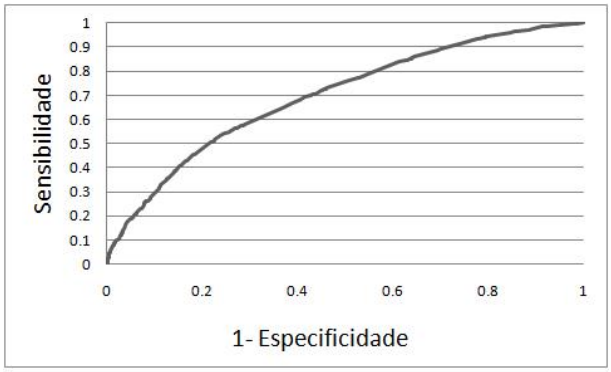

Figura 6.9: Curva ROC

A curva ROC, quanto mais próxima do canto superior esquerdo, indica um melhor poder de discriminação, pois assim indica maiores valores respectivos de sensibilidade e especificidade. Para a base em exemplo, tratando-se de um modelo de *credit score*, onde há carência em quantidade e qualidade de informações, a curva ROC apresentada indica um modelo relativamente bom, bem como o KS que foi de 30%.

Prosseguimos com a análise de decis (Tabela 6.15) e o gráfico de aderência (Figura 6.10), mostrando que a taxa real de resposta e o *score* médio tem valores muito próximos, e que a taxa real de evento é decrescente a cada decil, mostrando que o modelo esta sendo capaz de ordenar o evento. Dividindo a faixa 9 pela 0, temos que a faixa 0 de *score* é 6,4 vezes mais predisposta à resposta do que a faixa 9, mostrando que o modelo está conseguindo separar os mais propensos a inadimplência dos menos propensos.

Tabela 6.15: Análise de decis - Amostra Desenvolvimento

Decil	Resposta		Total	Taxa de Resposta	Score Médio
	0	1			
0	198	314	512	0,61	0,63
1	258	254	512	0,50	0,48
2	288	224	512	0,44	0,40
3	354	159	513	0,31	0,33
4	373	139	512	0,27	0,30
5	387	125	512	0,24	0,25
6	388	125	513	0,24	0,23
7	406	106	512	0,21	0,19
8	431	81	512	0,16	0,16
9	464	49	513	0,10	0,10
Total	3547	1576	5123	0,31	0,31

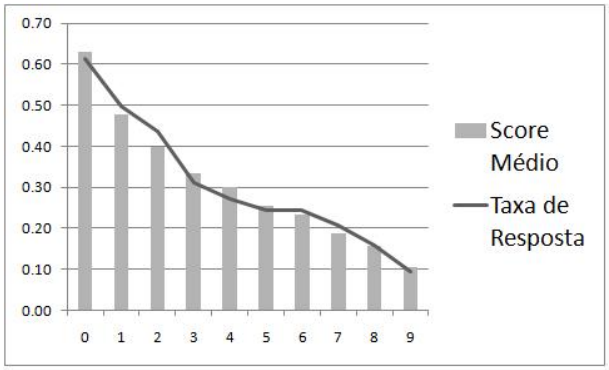

Figura 6.10: Gráfico de Aderência - Amostra Desenvolvimento

A partir da tabela de análise de decis podemos desenvolver a tabela de ganhos (Tabela 6.16), onde observamos pelo *lift* que, até a faixa dois (*lift* = 1,42), o risco de

inadimplência é maior que a média, chegando a quase o dobro quando trata-se da faixa zero (*lift* = 1,99), enquanto a faixa três mostra-se com um risco na média (*lift* = 1,01) e a partir dela até a faixa nove o risco é mais baixo que a média, sendo 69% menor na última faixa (*lift* = 0,31).

Ainda observando as três primeiras faixas de *score*, vemos através do *cum lift* que elas juntas apresentam um risco 68% maior que a média, e correspondem a apenas 30% da base. Já as duas primeiras faixas têm juntas, um risco 80% maior que a média, e correspondem a apenas 20% da base. Observamos também grandes saltos, de decil para decil, nos valores de *cum lift*, o que só reforça o bom poder de discriminação do modelo.

Tabela 6.16: Tabela de Ganhos - Amostra Desenvolvimento

Decil	Taxa de Resposta	Score Médio	Resposta acumulada	% do Total de eventos	% acumulado do total de eventos	Lift	Cum Lift
0	0,61	0,63	0,61	0,20	0,20	1,99	1,99
1	0,50	0,48	0,55	0,16	0,36	1,61	1,80
2	0,44	0,40	0,52	0,14	0,50	1,42	1,68
3	0,31	0,33	0,46	0,10	0,60	1,01	1,51
4	0,27	0,30	0,43	0,09	0,69	0,88	1,38
5	0,24	0,25	0,40	0,08	0,77	0,79	1,29
6	0,24	0,23	0,37	0,08	0,85	0,79	1,21
7	0,21	0,19	0,35	0,07	0,92	0,67	1,15
8	0,16	0,16	0,33	0,05	0,97	0,51	1,08
9	0,10	0,10	0,31	0,03	1	0,31	1
Total	0,31	0,31		1		1,00	

Observamos através do *lift chart* (Figura 6.11), construído ainda na amostra de desenvolvimento, a partir da tabela de análise de decis, que o *lift* é decrescente a cada decil, demonstrando a coerência das faixas construídas a partir dos *scores* estimados pelo modelo, e indicando que quanto maior a faixa de *score* menor o risco real a inadimplência. O gráfico de ganhos (Figura 6.12) mostra, por exemplo, que negando crédito a apenas 30% da base evitamos 50% dos inadimplentes, tendo um ganho de 67% em relação à média (abordagem aleatória, de 30% de inadimplência).

Imagine, na Figura 6.11, uma reta constante em *lift* = 1,00. Esse seria o *lift chart* se estivéssemos tratando de uma abordagem aleatória. Sendo assim, podemos considerar que quanto mais inclinada for a curva *lift*, maior o poder de discriminação do modelo. O modelo em exemplo apresenta considerável inclinação, indicando sua superioridade em relação a abordagem aleatória.

Figura 6.11: *Lift Chart* - Amostra Desenvolvimento

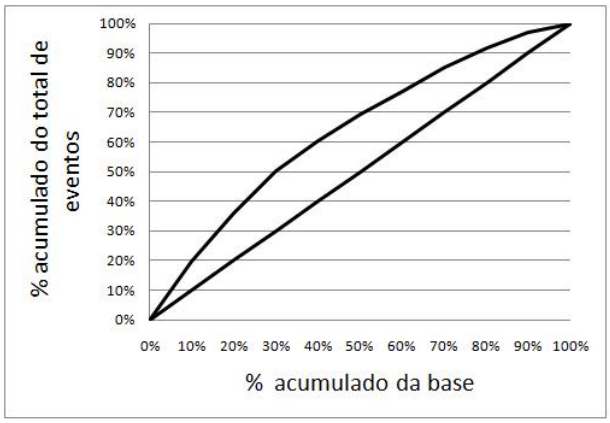

Figura 6.12: Gráfico de Ganhos - Amostra Desenvolvimento

Conforme discutido no Capítulo 4, a robustez do modelo pode ser avaliada através das amplitudes dos intervalos para o *lift*, obtidos através da re-amostragem via bootstrap. Como já dissemos, utilizar técnicas de re-amostragem nos permite fazer o que seria desejável na prática: repetir o experimento. A partir dessas re-amostras podemos construir intervalos de confiança para o *lift* resultante do modelo. Podemos observar estes intervalos na Tabela 6.17 e na Figura 6.13, em que sobrepomos ao intervalo o *lift* da amostra desenvolvimento.

De acordo com a tabela e o gráfico vemos que os intervalos contém as estimativas do *lift* para todos os decis, e podemos observar também que no geral os intervalos seguem

Tabela 6.17: Intervalo bootstrap - Amostra Desenvolvimento

Decil	Lift	Limite Inferior	Limite Superior
0	199,36	184,29	212,47
1	161,26	146,99	175,41
2	142,22	131,95	156,90
3	100,75	88,13	114,49
4	88,25	69,96	99,46
5	79,36	66,24	93,49
6	79,21	71,15	89,20
7	67,30	58,81	74,17
8	51,43	41,03	62,10
9	31,05	20,61	42,32

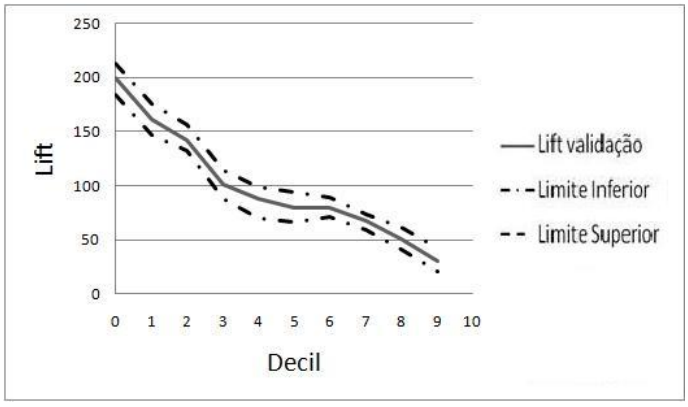

Figura 6.13: Intervalo de confiança para o *lift* - Amostra Desenvolvimento

o mesmo padrão da curva *lift*. O gráfico mostra que no decil quatro há um intervalo significativamente mais amplo que os demais.

O interesse agora é "escorar" (aplicar o modelo em outra amostra e obter os valores preditos) o modelo na amostra de validação, com o objetivo de verificar os valores da estatística KS e a análise de decil, validando o modelo e confirmando seu poder de predição em outros universos de aleatoriedade.

Partiremos da análise de decis (Tabela 6.18) e do gráfico de aderência (Figura 6.14) referentes à amostra de validação. A figura 6.14 mostra que apesar da taxa real de resposta e o *score* médio ter valores muito próximos na maioria dos decis, os decis 0 e 7 apresentam uma diferença significativa para estes valores, sendo que nestes casos, o modelo esta superestimando o risco de inadimplência. A superestimação é mais apropriada do que a subestimação nos modelos de crédito. Observe também que a taxa real de evento se mostra decrescente a cada decil e que dividindo a faixa 9 pela 0, temos que a faixa 0 de *score* é 5,3 vezes mais predisposta à resposta do que a faixa 9, mostrando que o modelo está conseguindo separar os mais propensos a inadimplência dos menos propensos num outro universo de aleatoriedade.

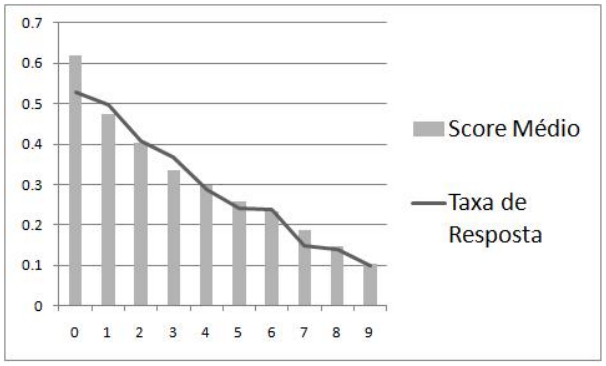

Figura 6.14: Gráfico de Aderência - Amostra Validação

A partir da tabela de análise de decis construímos a tabela de ganhos (Tabela 6.19), onde observamos pelo *lift* que, até a faixa 3 (*lift* = 1,24), o risco de inadimplência é maior que a média, chegando a 1,99 da média quando trata-se da faixa 0, enquanto a faixa quatro mostra-se com um risco ligeiramente abaixo da média (*lift* = 0,97) e a partir dela até a faixa nove o risco é mais baixo que a média, sendo 66% menor na última faixa (*lift* = 0,34).

Tabela 6.18: Análise de decis - Amostra Validação

Faixa de *Score*	Resposta		Total	Taxa de Resposta	Score Médio
	0	1			
0	92	103	195	0,53	0,62
1	122	120	242	0,50	0,47
2	71	49	120	0,41	0,40
3	270	157	427	0,37	0,34
4	74	30	104	0,29	0,30
5	181	58	239	0,24	0,26
6	150	47	197	0,24	0,23
7	226	39	265	0,15	0,19
8	160	26	186	0,14	0,15
9	198	22	220	0,10	0,10
Total	1544	651	2195	0,30	

Ainda observando as quatro primeiras faixas de *score*, vemos através do *cum lift* que elas juntas apresentam um risco 47% maior que a média, e correspondem a apenas 40% da base. Já as duas primeiras faixas têm juntas, um risco 72% maior que a média, e correspondem a apenas 20% da base. Observamos também saltos significativos, de decil para decil, nos valores de *cum lift*, o que só reforça o bom poder de discriminação do modelo.

Tabela 6.19: Tabela de Ganhos - Amostra Validação

Faixa de *Score*	Taxa de Resposta	Score Médio	Resposta acumulada	% do Total de eventos	% acumulado do total de eventos	Lift	Cum Lift
0	0,53	0,62	0,53	0,16	0,16	1,78	1,78
1	0,50	0,47	0,51	0,18	0,34	1,67	1,72
2	0,41	0,40	0,49	0,08	0,42	1,38	1,65
3	0,37	0,34	0,44	0,24	0,66	1,24	1,47
4	0,29	0,30	0,42	0,05	0,71	0,97	1,42
5	0,24	0,26	0,39	0,09	0,79	0,82	1,31
6	0,24	0,23	0,37	0,07	0,87	0,80	1,25
7	0,15	0,19	0,34	0,06	0,93	0,50	1,14
8	0,14	0,15	0,32	0,04	0,97	0,47	1,07
9	0,10	0,10	0,30	0,03	1,00	0,34	1,00
Total	0,30			1		1	

O gráfico de ganhos (Figura 6.15) mostra, por exemplo, que com 30% da base captamos e evitamos aproximadamente 45% dos inadimplentes, tendo um ganho de 50% em relação à média (abordagem aleatória, de 30% de inadimplência).

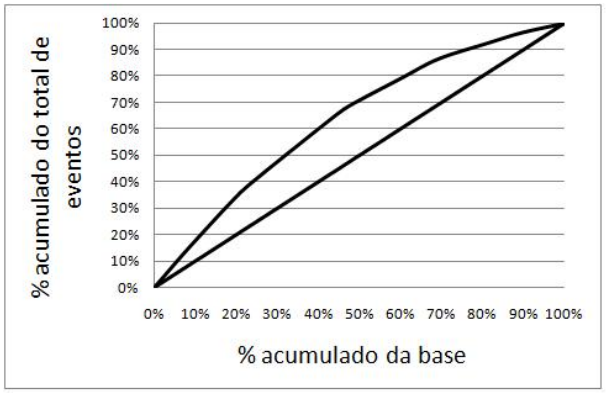

Figura 6.15: Gráfico de Ganhos - Amostra Validação

Na Figura 6.16 observamos que o *lift* resultante da amostra validação é ligeiramente maior entre as faixas três e cinco, e ligeiramente menor na faixa sete, porém o mais importante é que ele mostra um comportamento decrescente e uma inclinação considerável, assim como na amostra de desenvolvimento, validando o modelo. Observamos também que o valor do KS para a amostra de validação foi de 0,31, mostrando que o modelo não perde seu poder de predição em outros universos.

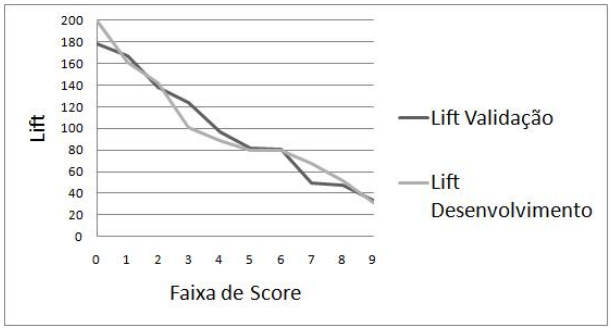

Figura 6.16: Sobreposição dos *lifts* da Amostra Desenvolvimento e Validação

Como dito na seção 4.3.1 podemos verificar, a cada faixa de *score*, se o valor do *lift* definido pela amostra validação está contido no intervalo determinado a partir da amostra desenvolvimento. Na figura 6.17 podemos perceber que, apenas nas faixas de *score* 0, 3 e 7 os valores do *lift* da amostra validação localizam-se fora do intervalo determinado pela reamostragem na amostra desenvolvimento. Notemos que ao passo que os intervalos superestimam os *lifts* nas faixas de *score* 0 e 7, que é mais apropriado para o caso de concessão de crédito do que a subestimação, isto não ocorre na faixa de *score* 3.

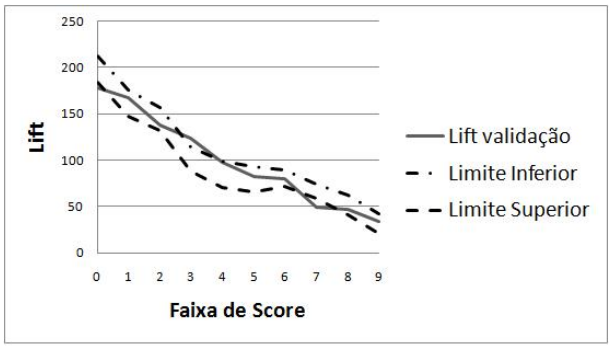

Figura 6.17: Intervalo de confiança para o *lift* - Amostra Validação

Comparando os resultados da amostra validação com os obtidos na amostra de desenvolvimento, podemos concluir que o modelo, apesar de apresentar menor desempe-

nho neste novo universo, como já era esperado, não apresentou padrões significativamente diferentes, validando assim o modelo ajustado.

Como tratamos de um modelo de *credit score*, temos o interesse em discriminar os indivíduos entre bons e maus pagadores, servindo de apoio no processo de concessão de crédito, tornando as freqüentes análises de crédito fáceis, rápidas e automatizadas. Sendo assim, temos que definir, a partir dos *scores* produzidos pelo modelo, um ponto de corte que possa discriminar os indivíduos, determinando assim uma regra de decisão.

Uma das maneiras de definir essa regra de decisão é utilizando a curva ROC. Lembremos que através da curva ROC podemos encontrar os pontos de maior sensibilidade e especificidade, garantindo uma boa classificação e conseqüentemente um bom percentual de acerto. No modelo em exemplo, esse ponto foi o *score* maior ou igual a 0,344, que mostra sensibilidade igual a 0,54 e especificidade igual a 0,76.

A partir desta classificação podemos obter as medidas de desempenho descritas na seção 4.4 e assim avaliar o modelo de classificação. Nesse ponto, também é interessante calcular essas medidas para ambas as amostras, desenvolvimento e validação, para que possamos verificar se existe proximidade entre seus valores, validando o modelo de classificação. A Tabela 6.20 contém as medidas de desempenho para ambas as amostras:

Tabela 6.20: Medidas de Desempenho - Amostra Desenvolvimento e Validação

Medida	Desenvolvimento	Validação
Sensibilidade	0,54	0,52
Especificidade	0,75	0,75
VPP	0,49	0,46
VPN	0,79	0,79
CAT	0,69	0,68
MCC	0,29	0,26

Podemos perceber que a perda da capacidade preditiva da amostra de desenvolvimento para a amostra validação não foi significativa, mostrando o poder preditivo em outros universos. Notamos ainda que os valores de Especificidade e VPN são idênticos para ambas as amostras. No entanto observamos, para o VPP e para MCC, valores rela-

tivamente baixos, porém vale ressaltar que estamos em um caso de *credit score*, onde há falta de qualidade e de quantidade nas informações.

6.3 Aplicação do Modelo Aditivo Generalizado para Dados Binários

6.3.1 Ajuste

O ajuste do MAG partiu das *dummys* criadas na seção 6.2.1 adicionadas pelas variáveis *limite99* (variável limite com correção, mais detalhes ver seção 6.1), *tpempreg*, *idade* e *tempores* em sua forma natural contínua. Desta forma as quatro variáveis cuja forma natural é contínua foram testadas de duas maneiras, na sua forma original utilizando o alisamento e na sua forma categorizada utilizando a forma paramétrica do modelo aditivo generalizado. Considerando as duas possibilidades das variaveis originalmente contínuas, foi feita a escolha do formato utilizado no modelo final através da comparação do ajuste entre as duas possibilidades (contínua e categórica).

Para o ajuste do MAG foi utilizada a *procedure* GAM, do software SAS. Neste procedimento as variáveis introduzidas de forma a buscar um padrão não linear são primeiramente segmentadas em duas partes, uma parte que pode ser explicada em seu formato linear (paramétrica) e a parte restante que não pode ser explicada linearmente (não paramétrica).

Assim como no modelo logístico, o método de seleção de variáveis utilizado foi o *backward* (Tabela 6.21), por apresentar maior número de variáveis selecionadas. Como não há implementado no PROC GAM (procedimento do software utilizado que ajusta o MAG) nenhum método de seleção de variáveis, seguimos o seguinte procedimento:

1. Ajustamos o modelo aditivo generalizado com as quatro variáveis contínuas de forma a obter funções suaves.

2. Testamos os métodos *backward*, *foward* e *stepwise* implementados no PROC LOGISTIC, utilizando como variáveis preditoras as quatro funções suaves (aplicadas

Tabela 6.21: Resumo Backward

Passo	Efeito Removido
1	dm_cep_2
2	dm_tempores_5
3	dm_localiza2_4
4	dm_estcivil
5	dm_prof
6	dm_regiao_1
7	dm_tempores_3
8	dm_cep_3
9	dm_cep_1
10	dm_regiao_2
11	dm_regiao_3
12	limite99
13	dm_tpempreg_4
14	dm_tempores_4
15	dm_tempores_2
16	dm_tpempreg_2
17	dm_localiza2_6
18	dm_tpempreg_3
19	dm_tpempreg_1

nos valores observados) obtidas no passo anterior mais todas as dummys que foram criadas a partir da análise descrita em 6.2.1.

3. As variáveis selecionadas entraram como variáveis preditoras no Modelo Aditivo Generalizado via *PROC GAM*.

Da mesma forma que fizemos com os modelos ajustados anteriormente, analisamos a correlação entre as variáveis selecionadas de forma a evitar a multicolinearidade. A partir disso, todas as variáveis selecionadas pelo *backward* foram mantidas no modelo, os parâmetros lineares estimados e sua respectiva significância seguem na Tabela 6.22, os parâmetros dos *splines* podem ser vistos na Tabela 6.23 e os respectivos testes de significância na Tabela 6.24. As variáveis *idade* e *limite99* entraram no modelo na sua forma continua. Os gráficos de alisamento por *splines* para *idade* e *limite99* podem ser vistos respectivamente nas Figuras 6.18 e 6.19. Estes gráficos nos mostram como a resposta muda ao longo da variável e que as relações seguem um padrão não linear. As variáveis permanecem com a mesmas estruturas de relacionamento com a resposta encontradas no ajuste do modelo com variáveis puras, mantendo seu significado.

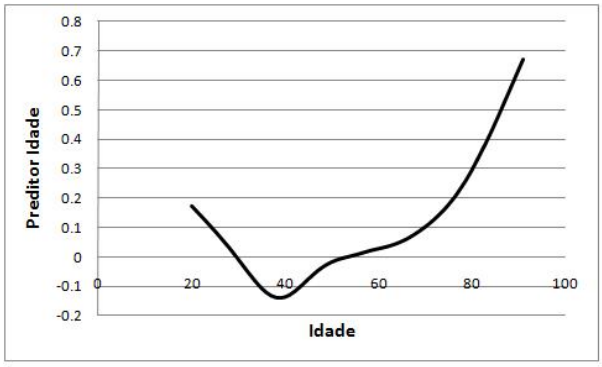

Figura 6.18: Efeito de idade na resposta

Com um modelo robusto, coerente, e capaz de explicar de maneira conjunta a relação entre as variáveis preditoras e a resposta, partimos para a validação, análise do desempenho preditivo e, por fim, a comparação com os resultados do modelo logístico.

Tabela 6.22: Estimativa dos Parâmetros - Parte Linear

Variável	Parâmetro	Valor t	Valor - p
Intercept	0,49782	4,26	<,00001
dm_tempores_1	0,19533	2,18	0,029
dm_localiza2_1	-1,21612	-9,42	<0,0001
dm_localiza2_2	-0,66387	-7,1	<0,0001
dm_localiza2_3	-0,37834	-4,56	<0,0001
dm_localiza2_7	0,66327	4,4	<0,0001
dm_localiza2_8	0,92322	4,41	<0,0001
dm_localiza2_9	1,14193	4,34	<0,0001
dm_tpclient	0,27541	3,96	<0,0001
dm_sexo	-0,17236	-2,63	0,0086
dm_sitresid	0,28573	2,53	0,0116
dm_tpempreg_5	-0,31977	-2,49	0,013
Linear(idade)	-0,02569	-11,08	<0,0001
Linear(limite99)	-0,00039	-1,08	0,2791

Tabela 6.23: Estimativa dos Parâmetros - Parte Não Linear

Componente	Parâmetro de Alisamento	GL	Número de Observações unicas
Spline(idade)	1	4	73
Spline(limite99)	1	4	393

Tabela 6.24: Análise de Deviance

	DF	Soma de Quadrados	Qui-Quadrado	Valor - p
Spline(idade)	4	13,58	12,11	0,0165
Spline(limite99)	4	14,48	12,92	0,0117

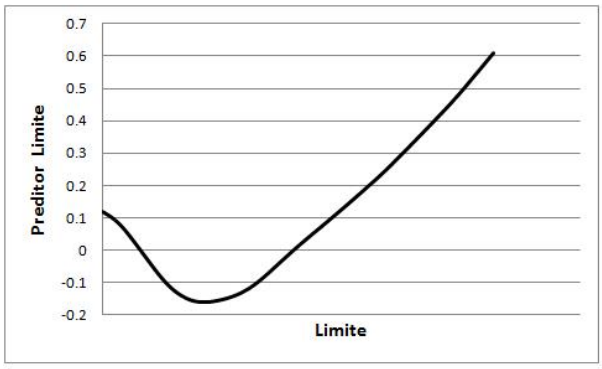

Figura 6.19: Efeito de limite na resposta

6.3.2 Análise do Desempenho e Validação

Devemos agora avaliar a capacidade preditiva e validar (na amostra validação) o modelo desenvolvido na seção 6.3.1.

Observamos uma estatística KS de 0,31 para a amostra desenvolvimento, que é ligeiramente maior que do modelo logístico (0,30). Observe no gráfico 6.20 a curva ROC do modelo aditivo e do modelo logístico, no mesmo plano. Observe que as duas curvas são muito próximas, sendo a do modelo aditivo pouco mais alta em alguns pontos, inclusive no ponto ótimo (ponto mais próximo ao canto superior esquerdo, com maior sensibilidade e especificidade), indicando um poder discriminatório maior.

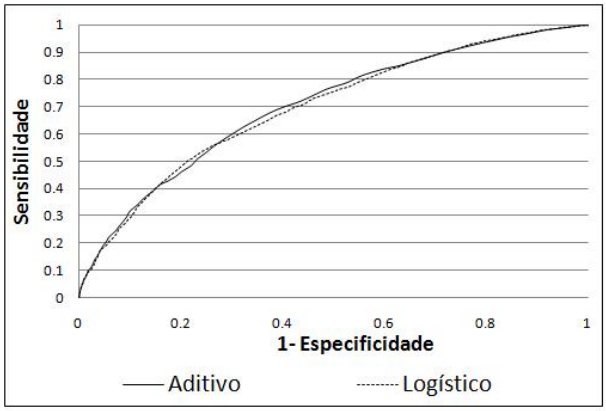

Figura 6.20: Curva ROC

Prosseguimos com a análise de decis (Tabela 6.25) e o gráfico de aderência (Figura 6.21), mostrando que a taxa real de resposta e o score médio estimado são muito próximos, e que a taxa real de resposta decresce a cada decil, exceto pequeno acréscimo de 0,02 no terceiro decil. Dividindo a faixa 9 pela 0, temos que a faixa 0 é 6,3 vezes mais predisposta a

resposta que a faixa 9, mostrando que o modelo está sendo capaz de separar os adimplentes dos inadimplentes.

Tabela 6.25: Análise de decis - Amostra Desenvolvimento

Decil	Resposta		Total	Taxa de resposta	Score Médio
	0	1			
0	189	323	512	0,63	0,63
1	259	253	512	0,49	0,48
2	326	186	512	0,36	0,40
3	316	197	513	0,38	0,34
4	360	152	512	0,30	0,30
5	387	125	512	0,24	0,26
6	405	108	513	0,21	0,23
7	411	101	512	0,20	0,19
8	434	78	512	0,15	0,15
9	460	53	513	0,10	0,10
Total	3547	1576	5123	0,31	0,31

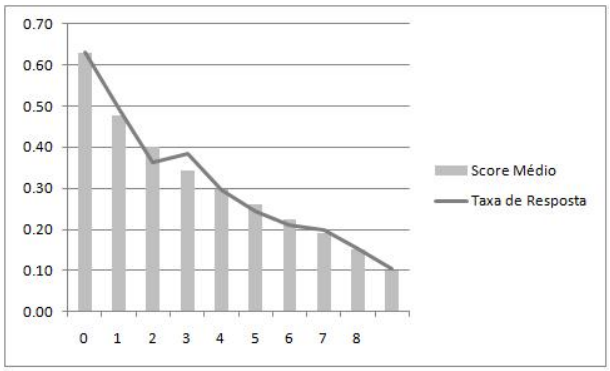

Figura 6.21: Gráfico de Aderência - Amostra Desenvolvimento

A partir da tabela de análise de decis podemos construir a tabela e o gráfico de ganhos Tabela 6.26 e Figura 6.22, respectivamente, onde observamos pelo *lift* que, até a faixa três (*lift* = 1,11), o risco a inadimplência é maior que a média, chegando a mais que o dobro quando trata-se da faixa zero (*lift* = 2,05). Observando o *cum lift* vemos, por exemplo, que as duas primeiras faixas juntas apresentam um risco 83% maior que a

média, com apenas 20% da base. Observamos também grandes saltos do *lift* de decil para decil, reforçando o bom poder discriminatório.

Tabela 6.26: Tabela de Ganhos - Amostra Desenvolvimento

Decil	Taxa de Resposta	Score Médio	Resposta acumulada	% do Total de eventos	% acumulado do total de eventos	Lift	Cum Lift
0	0,63	0,63	0,63	0,20	0,20	2,05	2,05
1	0,49	0,48	0,56	0,16	0,37	1,61	1,83
2	0,36	0,40	0,50	0,12	0,48	1,18	1,61
3	0,38	0,34	0,47	0,13	0,61	1,25	1,52
4	0,30	0,30	0,43	0,10	0,70	0,97	1,41
5	0,24	0,26	0,40	0,08	0,78	0,79	1,31
6	0,21	0,23	0,37	0,07	0,85	0,68	1,22
7	0,20	0,19	0,35	0,06	0,92	0,64	1,15
8	0,15	0,15	0,33	0,05	0,97	0,50	1,07
9	0,10	0,10	0,31	0,03	1,00	0,34	1,00
Total	0,31	0,31		1,00		1,00	

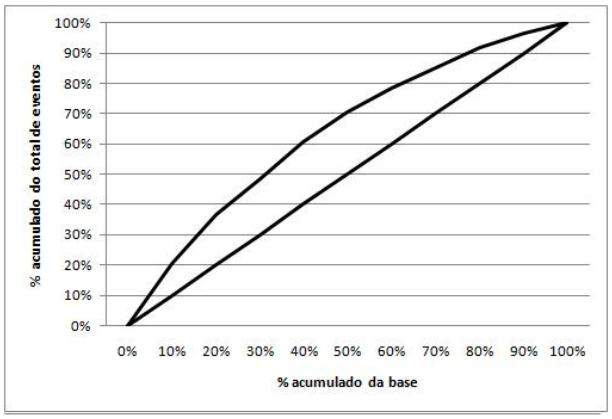

Figura 6.22: Gráfico de Ganhos - Amostra Desenvolvimento

Observamos no *lift chart* (Figura 6.23), construído ainda na amostra de desenvolvimento, a partir da tabela de análise de decis, que o *lift* é decrescente a cada decil, demonstrando a coerência das faixas construídas a partir dos scores estimados pelo modelo, e indicando que quanto maior a faixa de score menor o risco real a inadimplência.

No gráfico de ganhos (Figura 6.22) vemos, por exemplo, que com 30% da base, captamos e evitamos 48% dos inadimplentes.

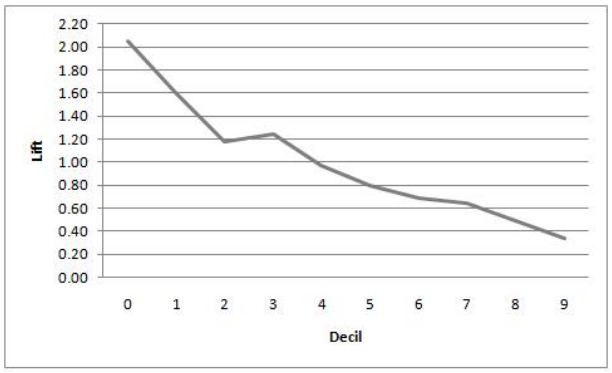

Figura 6.23: *Lift Chart* - Amostra Desenvolvimento

Utilizaremos mais uma vez a re-amostragem via bootstrap, nos permitindo realizar o que seria desejável na prática: repetir o experimento. A partir dessas re-amostras podemos construir intervalos de confiança para o *lift* resultante do modelo. Podemos observar esses intervalos na Tabela 6.27 e na Figura 6.24, em que sobrepomos ao intervalo o *lift* da amostra de desenvolvimento. A partir da tabela e do gráfico vemos que os intervalos contem as estimativas do *lift* para todos os decis, e podemos observar que os intervalos seguem o mesmo comportamento da curva do *lift*.

Tabela 6.27: Intervalo bootstrap - Amostra Desenvolvimento

Decil	Lift Desenvolvimento	Limite Inferior	Limite Superior
0	205,069	187,841	220,643
1	160,627	147,069	174,268
2	118,089	107,211	131,874
3	124,829	115,941	137,547
4	96,503	80,15	111,102
5	79,361	63,951	93,503
6	68,434	52,256	80,002
7	64,124	57,498	75,372
8	49,521	35,688	60,045
9	33,584	25,266	41,274

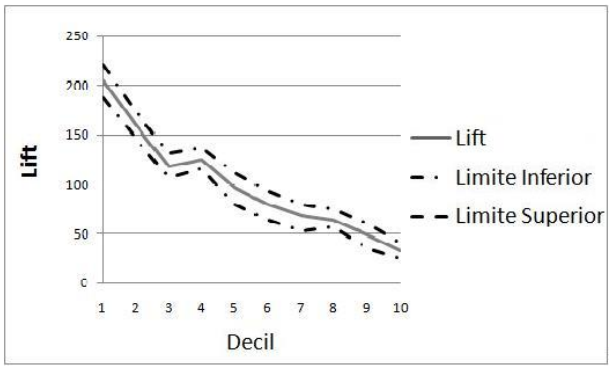

Figura 6.24: Intervalo de confiança para o *lift* - Amostra Desenvolvimento

O objetivo a partir de agora é escorar o modelo na amostra validação, verificando os valores da estatística KS e a análise de decil, validando o modelo e confirmando seu poder de discriminação em outros universos de aleatoriedade.

Podemos observar no gráfico de aderência (Figura 6.25) que a taxa real de resposta é muito próxima ao score médio, e a taxa real é estritamente decrescente conforme crescem os decis, mostrando que o modelo é capaz de ordenar o evento. Na tabela de ganhos (Tabela 6.29) vemos que até a faixa 4 (*lift* = 1,02) o risco a inadimplência é maior que a média, chegando a 1,93 quando trata-se da faixa 0.

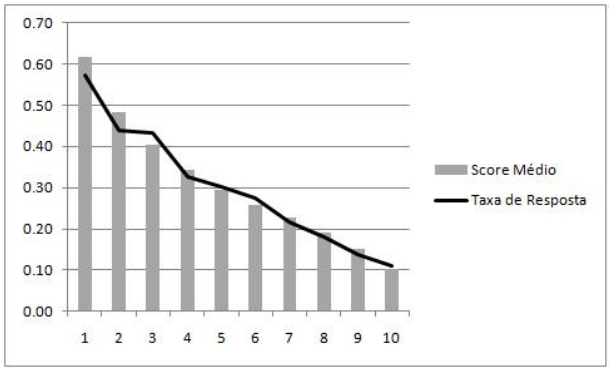

Figura 6.25: Gráfico de Aderência - Amostra Validação

Observando-se as quatro primeiras faixas, temos um risco 47% maior que a média, e apenas 40% da base. Já as duas primeiras faixas, juntas, têm um risco 70% maior que a média, com apenas 20% da base. Pelo gráfico de ganhos vemos, por exemplo, que com 30% da base captamos e evitamos 48% dos inadimplentes.

Na Figura 6.27 observamos o *lift chart* da validação sobreposto ao do desenvolvimento, e vemos que as curvas se aproximam e apresentam comportamento muito similar.

Tabela 6.28: Análise de decis - Amostra Validação

Faixa de *Score*	Resposta		Total	Taxa de Resposta	Score Médio
	0	1			
0	84	113	197	0,57	0,62
1	118	92	210	0,44	0,48
2	138	105	243	0,43	0,40
3	163	79	242	0,33	0,34
4	153	66	219	0,30	0,30
5	150	57	207	0,28	0,26
6	159	44	203	0,22	0,23
7	179	39	218	0,18	0,19
8	177	28	205	0,14	0,15
9	223	28	251	0,11	0,10
Total	1544	651	2195	0,30	0,30

Tabela 6.29: Tabela de Ganhos - Amostra Validação

Faixa de *Score*	Taxa de Resposta	Score Médio	Taxa de Resposta acumulada	% do Total de eventos	% acumulado do total de eventos	Lift	Cum Lift
0	0,57	0,62	0,57	0,17	0,17	1,93	1,93
1	0,44	0,48	0,50	0,14	0,31	1,48	1,70
2	0,43	0,40	0,48	0,16	0,48	1,46	1,61
3	0,33	0,34	0,44	0,12	0,60	1,10	1,47
4	0,30	0,30	0,41	0,10	0,70	1,02	1,38
5	0,28	0,26	0,39	0,09	0,79	0,93	1,31
6	0,22	0,23	0,37	0,07	0,85	0,73	1,23
7	0,18	0,19	0,34	0,06	0,91	0,60	1,15
8	0,14	0,15	0,32	0,04	0,96	0,46	1,08
9	0,11	0,10	0,30	0,04	1,00	0,38	1,00
Total	0,30	0,30		1,00		1,00	

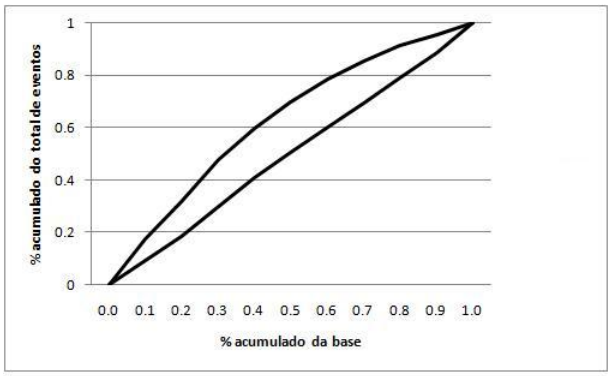

Figura 6.26: Gráfico de Ganhos - Amostra Validação

Observe também os gráficos 6.28 e 6.29 que comparam os *lifts* do modelo aditivo com o modelo logístico na amostra desenvolvimento e validação que no mesmo gráfico temos a curva de validação do modelo logístico, que também se mostra com comportamento similar as outras duas curvas, referentes ao modelo aditivo, e mostrando também valores muito próximos. Analisando a estatística KS, temos 0,30 na validação do modelo aditivo, enquanto no modelo logístico o valor foi de 0,31.

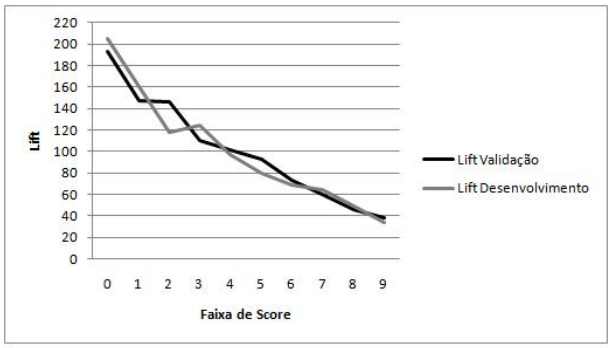

Figura 6.27: Sobreposição dos *lifts* da Amostra Desenvolvimento e Validação

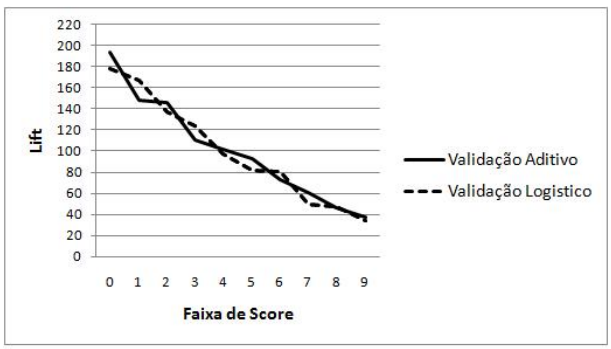

Figura 6.28: Sobreposição dos *lifts* Modelo Aditivo e Logistico para a Amostra Desenvolvimento

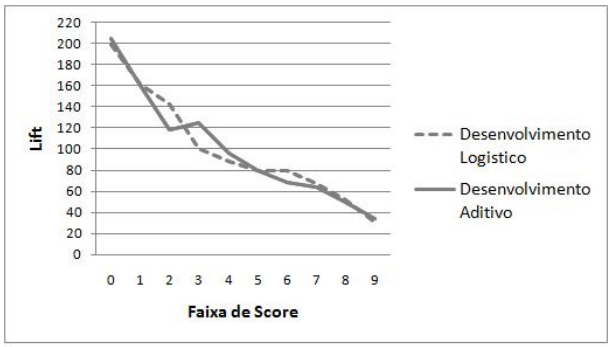

Figura 6.29: Sobreposição dos *lifts* Modelo Aditivo e Logistico para a Amostra Validação

Como vimos na seção 4.3.1, e fizemos com o modelo logístico, verificamos se a cada faixa de score o *lift* calculado na validação está contido no intervalo construído no desenvolvimento, a partir das re-amostragens via bootstrap. Observamos então na Figura 6.30 que apenas nas faixas 3 e 4 o valor do *lift* está ligeiramente fora do intervalo. Na Figura 6.24, referente ao mesmo gráfico para o modelo logístico, vimos que as faixas, 0, 3 e 7 ficaram ligeiramente fora do intervalo estimado no desenvolvimento. Comparando os resultados da validação com os resultados obtidos no desenvolvimento, concluímos que o modelo aditivo ajustado é capaz de produzir bons resultados em outro universo de aleatoriedade, não apresentando padrões significativamente diferentes. Podemos assim validar o modelo.

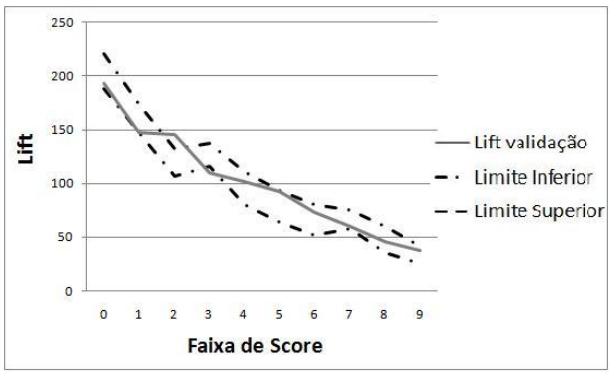

Figura 6.30: Intervalo de confiança para o *lift* - Amostra Desenvolvimento

Em um modelo de *credit score*, seja modelo logístico ou aditivo, temos o interesse em discriminar os indivíduos entre bons e maus pagadores, servindo de apoio no processo de concessão de crédito, tornando as análises de crédito rápidas e automatizadas. Dessa forma, assim como fizemos com o modelo logístico, temos que encontrar um ponto de corte, determinando uma regra de decisão. Utilizaremos mais uma vez a curva ROC (Figura

6.20) para encontrar esse ponto, buscando sempre a maior sensibilidade e especificidade, garantindo uma boa classificação e buscando uma boa taxa de acerto.

Definido esse ponto, podemos calcular as medidas de desempenho descritas na seção 4.4 e calculadas para o modelo logístico na seção 6.2.3. Observamos na Tabela 6.30 essas medidas calculadas para o desenvolvimento e para a validação do modelo aditivo.

Tabela 6.30: Medidas de Desempenho - Amostra Desenvolvimento e Validação

Medida	Desenvolvimento	Validação
Sensibilidade	0,65	0,63
Especificidade	0,65	0,64
VPP	0,45	0,43
VPN	0,81	0,81
CAT	0,65	0,64
MCC	0,28	0,25

Analisando apenas o modelo aditivo, vemos que os valores para a validação são próximos aos valores do desenvolvimento, mostrando o poder preditivo do modelo em outros universos. Comparando os valores da validação do modelo aditivo com os da validação do modelo logístico, vemos que os valores também são muito próximos, sendo ligeiramente maior no modelo aditivo para Sensibilidade e VPN, e ligeiramente menor para Especificidade, VPP, CAT e MCC.

6.4 Comentários Finais

Neste capítulo exemplificamos, utilizando uma base disponibilizada por uma instituição, todos os passos para o desenvolvimento de um modelo de *credit score*. Vimos no primeiro ajuste seção 6.1, no qual não usamos de categorizar nenhuma covariável de natureza contínua, que o MLG e o MAG apresentam capacidades preditivas muito similares, sendo o MAG ligeiramente superior em algumas medidas. No segundo ajuste, em que usamos de categorizar algumas covariáveis contínuas observamos novamente resultados muito similares, sendo o modelo aditivo ligeiramente superior em algumas medidas. Comparando o

primeiro ajuste ao segundo (seções 6.2 e 6.3), vimos que tanto o MLG quanto o MAG apresentaram melhor desempenho quando trabalhamos as variáveis contínuas categorizando-as quando necessário. Esse aumento de capacidade preditiva, do primeiro ajuste para o segundo, pode ser observada diretamente ao analisarmos as medidas de desempenho (Tabelas 6.9, 6.20 e 6.30).

Em paralelo ao segundo ajuste falamos da importância da análise bivariada inicial, do seu uso nas categorizações e recategorizações necessárias, de métodos de seleção de variáveis, eliminação de multicolinearidade, da análise e interpretação dos parâmetros. Falamos também do processo de análise de desempenho e validação desses modelos, desde a seleção da base de validação até a análise do KS, a análise de decil, o *lift chart*, a tabela de ganhos, a curva ROC, bem como sua utilização para a definição de uma regra de decisão e, por fim, mostramos como analisar as medidas de desempenho de um modelo de classificação. O desafio nesses modelos é grande, e sempre há espaço para novas técnicas e metodologias que buscam modelos com desempenhos preditivos maiores.

Mais que essa comparação, apresentamos uma metodologia alternativa para o desenvolvimento de modelos de *credit score* e quaisquer outros modelos de resposta binária. Os modelos aditivos generalizados se mostram, assim como a regressão logística, um modelo robusto, com capacidade preditiva satisfatória, fácil interpretação, ajuste e implementação. Além disso, permitem uma interpretação plausível a respeito da estrutura de relacionamento entre cada covariável e a variável resposta, considerando as demais variáveis constantes, este tipo de análise é sempre interessante para os pesquisadores sendo assim de extrema importância que não se baseie em uma suposição rígida como é o caso do ajuste em qualquer modelo linear. Certamente esta metodologia é competitiva para casos nos quais o interesse principal é a predição e mais ainda quando o interesse é a análise de impacto e significado de cada variável com relação a resposta.

Capítulo 7

Conclusão e Propostas Futuras

7.1 Conclusão

Um dos objetivos desta dissertação foi o estudo, o desenvolvimento e a comparação da capacidade preditiva entre uma técnica de modelagem ainda relativamente pouco explorada, que é o Modelo Aditivo Generalizado e uma técnica mais usual, que é o Modelo Linear Generalizado. Esta comparação foi explorada no contexto de dados cuja resposta é uma variável binária no contexto de risco de crédito. Também introduzimos no contexto de *credit score* o *lift*, medida usual no marketing (Rud, 2000) mas ainda pouco utilizada na área de finanças.

A linha de aplicação seguida nesse trabalho foi especificamente no contexto do ciclo de crédito, mais particularmente em modelos de *credit score*. Todos os passos e informações utilizadas para o ajuste desses modelos são encontrados nesta dissertação, bem como técnicas de validação e idéias de aplicabilidade no mercado financeiro.

Na aplicação, ambas metodologias (MLG e MAG) foram aplicadas no desenvolvimento de um modelo de *credit score*. Em um primeiro momento (seção 6.1) buscamos uma comparação mais bruta, na qual não usamos de categorizar nenhuma covariável de natureza contínua, neste caso observamos que o MLG e o MAG apresentam capacidades preditivas muito similares, sendo o MAG ligeiramente superior em algumas medidas. Em um segundo momento buscamos uma comparação mais detalhada entre modelos mais elaborados, seguindo as práticas de melhor incorporação das covariáveis nos mesmos. Neste

momento, discutindo todos os conceitos práticos e usuais, passando pelas etapas de análise bivariada, categorizações e recategorizações, a seleção de variáveis, a análise de correlação, o ajuste propriamente dito, a análise de decis, a sua validação e, por fim, a definição de uma regra de decisão e a análise da capacidade de classificação.

Observamos nessas análises que, para a base em estudo, os resultados do MLG e do MAG foram próximos, sendo o MAG melhor algumas vezes. Como já discutimos, os modelos de *credit score*, por tratarmos de uma etapa de concessão, onde inicia-se o ciclo de crédito, perdemos muito em quantidade e qualidade nas informações. Ainda não temos informações históricas e comportamentais dos solicitantes, apenas as informações cadastrais disponibilizadas pelos mesmos no ato da solicitação. Dessa forma, é interessante o desenvolvimento de uma análise similar para uma base, por exemplo, de *behaviour*, por possuir maior riqueza e quantidade de informações, dado que já temos o cliente na base há algum tempo Esse tema é uma proposta futura e foi abordado na Seção 7.2. Comparando o primeiro ajuste ao segundo, vimos que tanto o MLG quanto o MAG apresentaram melhor desempenho quando trabalhamos as variáveis contínuas categorizando-as quando necessário.

No estudo de simulação, tivemos como objetivo analisar o comportamento das duas metodologias na presença de variáveis preditoras de natureza contínua e uma resposta binária, em diferentes cenários: proporção de evento 1%, 10%, 25% e 50%. Para isso, foram simuladas cinqüenta bases para cada um desses cenários e ajustados MLG e MAG em todas as bases, tendo seus resultados posteriormente comparados.

Vimos que o MAG se mostrou mais adequado em todos os quatro cenários analisados, desde a significância das variáveis preditoras, até medidas como o KS e a análise de decil. Isso só reforça a proposta de utilizarmos MAG quando possuímos variáveis preditoras de natureza contínua, principalmente quando não apresentam estrutura de relação linear com a resposta, já que para este caso o MAG cairá no caso particular dos MLG.

Também vimos nessa dissertação que o MAG é um modelo de fácil entendimento, aplicabilidade, interpretação, implementação e implantação, assim como o MLG, diferente de algumas metodologias, como as Redes Neurais, que apresentam dificuldades nesses aspectos, principalmente em relação a interpretação. Dessa forma, o legado deixado por esse

trabalho é uma metodologia competitiva, e que produz resultados satisfatórios de maneira simples, diferente de outras metodologias alternativas já estudadas nesse contexto.

7.2 Propostas Futuras

Modelos de regressão logística são os mais usuais em todos os modelos do ciclo de crédito, sendo assim a metodologia mais usual nos modelos de *credit score*. Como já discutimos, nessa etapa do ciclo de crédito, a qualidade e a quantidade das informações é muito baixa, levando a modelos com desempenhos preditivos abaixo do esperado. Os modelos aditivos generalizados surgem como alternativa para esses casos (Hastie & Tibshirani, 1990). Pouco discutida nesse contexto, essa metodologia visa buscar uma função que explique a relação bivariada entre cada covariável contínua e a resposta, tendo assim um ganho de informação, já que ao invés de supormos linearidade, por exemplo, estimamos esta função, possibilitando alavancar o desempenho preditivo do modelo.

Quando o efeito de uma covariável contínua na variável resposta não apresenta uma estrutura linear, a categorização da mesma é uma boa alternativa, mas dado que a classificação pode acarretar numa perda considerável de informação ela pode não ser a melhor alternativa para o caso. Dessa forma, se fosse possível descobrir a forma estrutural de relacionamento entre a covariável e a resposta, seria interessante incorporá-la ao modelo, evitando dessa forma a perda de informação. Apesar da dificuldade em descobrir essa estrutura de relacionamento, é possível tentar incorporar novas formas ao modelo logístico. É isso que os modelos Aditivos Generalizados para dados binários propõem, baseando-se em uma forma de alisamento nas variáveis contínuas. A partir desse alisamento, é possível a utilização das variáveis contínuas no modelo, sem que a relação apresentada seja linear.

Nesse trabalho vimos a aplicação dos MAG em bases simuladas e em um problema de *credit score*, onde tínhamos apenas três variáveis contínuas, pois por tratar-se da etapa de concessão de crédito, temos poucas informações do cliente. A partir disso, uma proposta futura é repetir as análises e comparações aqui apresentadas em uma base com um maior número de variáveis contínuas, como por exemplo em uma base para o desenvolvimento de um modelo de *behaviour score*, com objetivo de analisar o comportamento dos Modelos Aditivos Generalizados nesse cenário, já que vimos aqui que

há ganho em utilizá-lo quando tratamos de variáveis preditoras contínuas. Fica aqui também um estímulo para outros estudos de simulação, envolvendo bases com maior número de covariáveis contínuas.

Capítulo 8

Apêndice

8.1 Representação por B-Splines

Seja $l = M - 1$, o grau da spline, no caso das *splines* cúbicas temos $M = 4$ e $l = 3$, uma forma eficiente de calcular os coeficientes das B-splines é dada através da relação recursiva:

$$B_{k,l+1}(x) = \left[\frac{x - \xi_k}{\xi_{k+l} - \xi_k}\right] B_{k,l}(x) + \left[\frac{\xi_{k+l+1} - x}{\xi_{k+l+1} - \xi_{k+1}}\right] B_{k+1,l}(x) \tag{8.1}$$

definindo

$$B_{k,0}(x) = \begin{cases} 1, \ \xi_k < x < \xi_{k+1} \\ 0, \ \text{caso contrário} \end{cases}$$

Dessa forma substituindo l por 3 e voltando a $s(x)$ podemos reescrever as bases B_k e estimar os coeficientes γ_k através de regressão linear usual considerando as bases como variáveis explicativas.

Considerando que a *spline* cúbica restrita (linearidade após os extremos) é subespaço da *spline* cúbica, a representação das bases B-splines será feita baseada nas *spline* cúbicas irrestritas, conforme Hastie e Tibshirani (1990). Temos que a spline cúbica pode ser representado em função das bases B na forma $s(x) = \sum_{k=1}^{n+2} \gamma_k B_k(x)$, onde γ_k são os coeficientes desconhecidos e $B_k(x)$ são as bases.

Para resolver a expressão 3.1 substituímos $f(x)$ por $s(x)$. Definindo a matriz B $n \times (n+2)$ e Ω $(n+2) \times (n+2)$ por $B_{ik} = B_k(x_i)$ e $\Omega_{ik} = \int B_i''(x) B_k''(x) dx$ podemos

reescrever 3.1 como:

$$(y - B\gamma)'(y - B\gamma) + \lambda\gamma'\Omega\gamma \tag{8.2}$$

derivando em γ e igualando a zero temos:

$$(B'B + \lambda\Omega)\hat{\gamma} = B'y \tag{8.3}$$

seja $M = B'B + \lambda\Omega$ então, de acordo com Hastie e Tibshirani (1990), podemos facilmente escrever M como $M = LL'$ através da decomposição de Scholesky. Então teremos:

$$LL'\hat{\gamma} = B'y \tag{8.4}$$

e esta equação pode ser resolvida através de *back-substitution* o que nos dá uma estimativa para γ e portanto $\hat{s}$.

Referências Bibliográficas

[1] Abreu, H.J. (2005). Aplicação da Análise de Sobrevivência em um problema de *Credit Scoring* e comparação com a Regressão Logística. Dissertação de mestrado. DEs - Ufscar.

[2] Altman E.I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance 23(4), p. 589-611.

[3] Baldi, P; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000, 16, 412-424.

[4] Beaver, W. (1967). Financial ratios predictors of failure. Journal of Accounting Research 4, p.71-111.

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

[6] Demétrio, C. G. B. (2002). Modelos Lineares Generalizados em Experimentação Agronômica. ESALQ/USP - Piracicaba, SP http://www.lce.esalq.usp.br/clarice/Apostila.pdf

[7] Griner PF, Mayewski RJ, Mushlin AI, Greenland P 1981. Selection and interpretation of diagnostic test procedures. Principles and applications. Ann Intern Med 94: 553-600.

[8] Gruenstein, J.M.L. (1998). Optimal Use of Statistical Techniques in Model Building. In: Credit Risk Modeling: Design and Application. Mays E.,81-112, New York: AMACOM.

[9] Hastie, T.J. and Tibshirani, R.J. (1990), Generalized Additive Models, New York: Chapman and Hall.

[10] Hosmer, W. & Lemeshow, S. (1989). Applied Logistic Regression. Wiley.

[11] LINNET, K., BRANDT, E. Assessing diagnostic tests once an optimal cutoff point has been selected. Clin. Chem., v. 32, n.7, p. 1341-6, 1986.

[12] Linsey, J. K. Applying Generalized Linear Models. New York: Springer-Verlag, 1997. 256p.

[13] Liu, W. Cela, J. Improving Credit Scoring by Generalized Additive Model. SAS Institute Inc, 2007.

[14] Louzada-Neto, F. Amaral G. A.; Abreu H.J.; Guirado L.; Ferreira. M.R.P.; Silva P.H.F. Medidas Estatísticas da Capacidade de Modelos de Class em *Credit Scoring*. Rev Tecnologia de Crédido - Serasa Experian, 2009, 7-28.

[15] Mazucheli, J.; Louzada-Neto, F.; Guirado,l.; Matinez E.Z. (2008). Algumas Medidas do Valor Preditivo de um Modelo de Classificação. Rev. Bras. Biom., São Paulo,v.2 6, n.2 , p.8 3-91.

[16] Martinez, E.Z.; Louzada-Neto, F. (2000). Metodologia estatística para testes diagnósticos e laboratoriais com respostas dicotomizadas. Revista de Matemática e Estatística, 18 , 83-101.

[17] Matthews, B.W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 1975, 405, 442-451.

[18] McCullagh, P.; Nelder, J. A. (1989). Generalized Linear Models, 2ed, Chapman & Hall, London.

[19] Mendonça, T.S. (2008). Aplicações e comparações de modelos de Regressão Logística Clássica, Bayesiana e de Redes Neurais em um problema de *Credit Scoring*. Dissertação de mestrado. DEs - Ufscar.

[20] Moineddin, R.; Matheson F.I., Glazier R.H. (2007). A simulation study of sample size for multilevel logistic regression models. BMC Med Res Methodol.Published online.

[21] Myers, R.H. & Montgomery, D.C. (1997). A Tutorial on Generalized Linear Models. Journal of Quality Technology, 29, 274-291.

[22] Myers R.H., Montgomery D.C. (2002), Response Surface Methodology. Wiley, New York.

[23] Nelder, J. A.; Wedderburn, R. W. M. (1972). Generalized Linear Models. J. R. Stastist. Soc. A, 135, 370-384.

[24] Piccard, R.R.,Data Splitting, The American Statistician, Vol. 44, No. 2 (May, 1990), pp. 140-147.

[25] Peduzzi P.; Concato J.; Kemper E.; Holford T.R.; Feinstein A.R.(1996). A simulation Study of the Number of Events per Variable in Logistic Regression Analysis. J Clin Epidemiol Vol 49 No 12, pp 1373-1379.

[26] Rud, O.P., Data Mining Cookbook. Wiley, 2001.

[27] Reiser, B. e Faraggi, D. (1997). Confidence intervales for the generalized ROC criterion. Biometrics 53, 192-202.

[28] Sabato G. (2009) Modelos de *Scoring* de Risco de Crédito. Rev Tecnologia de Crédido - Serasa Experian, pp 29-47.

[29] Salvatore, D.; Reagle, D. Theory and Problems of Statistics and Econometrics. Schaum‘s Outline Series. Mc Grawn-Hill 2002.

[30] SCHÄFER, H. Constructing a cut-off point for a quantitative diagnostic test. Stat. Med., v.8, p.1381-91, 1989.

[31] SCHÄFER, H. Efficient confidence bounds for ROC curves. Stat. Med., v.13, 1551-61, 1994.

[32] Thomas, L.C.; Stepanova, M.(2002) Survival analysis methods for personal loan data. Operations Research, v.50, 2, p.277-289.

[33] Webb, A. (2002), Statistical Pattern Recognition, 2a Ed., Wiley.

[34] Wiginton. C. J. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior; The Journal of Financial and Quantitative Analysis, Vol. 15, No. 3 (Sep., 1980), pp. 757-770.

[35] Wood, S.N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Amer. Statist. Ass. 99:673-686

[36] Coppock D.S. (2002) Why Lift? http://www.information-management.com/news/5329-1.html

[37] Zweig, M. H.; Campbell, G. (1993). Receiver-operating characteristic (ROC) plots. Clin. Chem.,29, 561-577.