Estratégias para tratamento de variáveis com
dados faltantes durante o desenvolvimento de
modelos preditivos

Fernando Assunção

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

Programa: Estatística

Orientadora: Profa. Dra. Lúcia Pereira Barroso

São Paulo, Junho de 2012

Estratégias para tratamento de variáveis com
dados faltantes durante o desenvolvimento de
modelos preditivos

Este exemplar corresponde à redação final da
dissertação devidamente corrigida e defendida por
Fernando Assunção e aprovada pela comissão
julgadora

Comissão Julgadora

- Profa. Dra. Lúcia Pereira Barroso (orientadora) – IME-USP
- Profa. Dra. Mônica Carneiro Sandoval – IME-USP
- Prof. Dr. Rinaldo Artes – Ibmec, SP

AGRADECIMENTOS

Primeiramente a Deus, a razão do viver, pelo amor que tem para conosco. Amor este que nos dá paz e esperança.

Aos meus pais que sempre ensinaram o caminho por onde se deve andar, através de exemplos de comportamento e caráter, sendo verdadeiros orgulhos. Agradeço também por todo esforço que fizeram para propiciar as condições necessárias para que pudéssemos nos desenvolver fisicamente, intelectualmente e religiosamente.

Também às minhas queridas irmãs, grandes amigas, com as quais sempre pude contar e que amo profundamente.

À minha esposa Ana Cecília pela companhia, apoio e paciência. Agradeço diariamente a Deus por ter me dado este presente que tanto amo.

Agradeço com toda sinceridade à Professora Lúcia, que sempre esteve disposta a ajudar, sendo muito correta e coerente, além de ter contribuído significativamente para este trabalho.

Por fim, agradeço às chefes que tive durante este período de estudo, por incentivarem meu desenvolvimento acadêmico.

“Combati o bom combate, completei a carreira, guardei a fé”.

II Timóteo 4:7

RESUMO

Modelos preditivos têm sido cada vez mais utilizados pelo mercado a fim de auxiliarem as empresas na mitigação de riscos, expansão de carteiras, retenção de clientes, prevenção a fraudes, entre outros objetivos. Entretanto, durante o desenvolvimento destes modelos é comum existirem, dentre as variáveis preditivas, algumas que possuem dados não preenchidos (*missings*), sendo necessário assim adotar algum procedimento para tratamento destas variáveis. Dado este cenário, este estudo tem o objetivo de discutir metodologias de tratamento de dados faltantes em modelos preditivos, incentivando o uso de algumas delas já conhecidas pelo meio acadêmico, só que não utilizadas pelo mercado. Para isso, este trabalho descreve sete metodologias. Todas elas foram submetidas a uma aplicação empírica utilizando uma base de dados referente ao desenvolvimento de um modelo de *Credit Score*. Sobre esta base foram desenvolvidos sete modelos (um para cada metodologia descrita) e seus resultados foram avaliados e comparados através de índices de desempenho amplamente utilizados pelo mercado (KS, Gini, ROC e Curva de Aprovação). Nesta aplicação, as técnicas que apresentaram melhor desempenho foram a que tratam os dados faltantes como uma categoria à parte (técnica já utilizada pelo mercado) e a metodologia que consiste em agrupar os dados faltantes na categoria conceitualmente mais semelhante. Já a que apresentou o pior desempenho foi a metodologia que simplesmente não utiliza a variável com dados faltantes, outro procedimento comumente visto no mercado.

Palavras-chave: *credit score*, dados faltantes, imputação múltipla, modelos preditivos, regressão logística.

ABSTRACT

Predictive models have been increasingly used by the market in order to assist companies in risk mitigation, portfolio growth, customer retention, fraud prevention, among others. During the model development, however, it is usual to have, among the predictive variables, some who have data not filled in (missing values), thus it is necessary to adopt a procedure to treat these variables. Given this scenario, the aim of this study is to discuss frameworks to deal with missing data in predictive models, encouraging the use of some already known by academia that are still not used by the market. This paper describes seven methods, which were submitted to an empirical application using a Credit Score data set. Each framework described resulted in a predictive model developed and the results were evaluated and compared through a series of widely used performance metrics (KS, Gini, ROC curve, Approval curve). In this application, the frameworks that presented better performance were the ones that treated missing data as a separate category (technique already used by the market) and the framework which consists of grouping the missing data in the category most similar conceptually. The worst performance framework otherwise was the one that simply ignored the variable containing missing values, another procedure commonly used by the market.

Key-words: credit score, logistic regression, missing values, multiple imputation, predictive models.

SUMÁRIO

1	INTRODUÇÃO	8
1.1	APRESENTAÇÃO DO PROBLEMA	8
1.2	OBJETIVOS DO ESTUDO	12
2	DESCRIÇÃO DO ESTUDO	14
2.1	DESCRIÇÃO DOS DADOS	14
3	METODOLOGIA	17
3.1	REGRESSÃO LOGÍSTICA	17
3.2	MÉTODOS DE TRATAMENTO DE DADOS FALTANTES	19
4	AVALIAÇÃO DAS TÉCNICAS	32
4.1	KS	32
4.2	CURVA ROC E COEFICIENTE DE GINI	33
4.3	CURVA DE APROVAÇÃO	37
5	APLICAÇÃO	39
5.1	ANÁLISE DESCRITIVA	39
5.2	PROCESSO DE SELEÇÃO DAS VARIÁVEIS	41
5.3	AJUSTE DOS MODELOS	42
5.4	COMPARAÇÃO DOS RESULTADOS	43
6	CONCLUSÃO	46
	APÊNDICE A – TABELAS DE ANÁLISE DESCRITIVA	48
	APÊNDICE B – MODELOS AJUSTADOS	56
	APÊNDICE C - FIGURAS	63
	REFERÊNCIAS BIBLIOGRÁFICAS	71

1 INTRODUÇÃO

1.1 APRESENTAÇÃO DO PROBLEMA

Pesquisadores de diferentes áreas têm cada vez mais consciência dos problemas que podem ser causados pelos dados faltantes (*missings*). Nos últimos anos esta questão tem sido cada vez mais estudada e metodologias foram desenvolvidas para tentar solucioná-la. Infelizmente algumas destas metodologias acabam não sendo utilizadas por uma série de motivos (falta de conhecimento, problemas computacionais, tempo, entre outros) e os pesquisadores recorrem a abordagens mais simples de tratamento de *missings*, o que pode vir a trazer mais prejuízos do que benefícios.

Paralelo a isso há um gradativo aumento de empresas utilizando modelos preditivos. Tais modelos correspondem a uma equação (ou regra) na qual através de informações históricas ou atuais prediz-se um determinado evento futuro a respeito de um indivíduo. Abaixo segue lista com alguns dos principais modelos preditivos (com a respectiva área de negócio) utilizados pelo mercado atualmente, juntamente com a função de cada um.

- *Credit Score* (Crédito): classificar solicitantes de crédito segundo seu risco de inadimplência;
- *Behaviour Score* (Crédito): classificar clientes de uma carteira segundo seu risco de inadimplência;
- *Collection Score* (Cobrança): classificar clientes inadimplentes segundo sua probabilidade de quitação do débito em aberto;
- Prevenção a Fraude na Concessão (Fraude): classificar solicitantes de crédito segundo a probabilidade de o proponente ser fraudador;
- Prevenção a Fraude Transacional (Fraude): classificar transações (ex.: compras no cartão de crédito) segundo o risco de ela ser fraudulenta;
- Modelos de *Cross-Selling* (Marketing): classificar clientes de uma carteira segundo a probabilidade de aquisição de um novo produto;

- Modelos de *Up-Selling* (Marketing): classificar clientes que possuem um determinado produto segundo a probabilidade de algum tipo de migração dentro deste mesmo produto.

Durante o desenvolvimento destes modelos é comum se deparar com variáveis que possuem dados faltantes, as quais podem ser originadas por diferentes razões como: não preenchimento cadastral, problemas no armazenamento dos dados, entre outros. Além disso, em alguns casos, uma ou mais variáveis com dados faltantes podem ser fortemente correlacionadas com o evento futuro que o modelo tenta prever (variável resposta), podendo assim incrementar consideravelmente o poder preditivo do modelo. Portanto, a escolha da forma de tratamento dos *missings* será fundamental para o sucesso do modelo.

É importante ressaltar que uma vez que a base de construção do modelo apresenta *missings* é impossível não tratá-los, pois tendo realizado qualquer procedimento sobre este conjunto de dados necessariamente adotou-se algum procedimento sobre os dados faltantes.

Alguns dos métodos mais utilizados de tratamento de *missings* envolvem a remoção ou a substituição dos dados faltantes. Eles geralmente editam os *missings* para assim produzir um conjunto de dados completo e são atraentes porque são fáceis de serem implementados. No entanto, estes métodos podem trazer sérias desvantagens. Por exemplo, um dos casos mais comuns, é tratar os dados faltantes eliminando estes casos (e construindo o modelo apenas com os dados completos), nesta situação, os resultados obtidos se tornarão viesados se os casos restantes não representarem toda a amostra. Este método é o padrão na maioria dos *softwares* estatísticos. Outro procedimento que segue a mesma linha é não incluir no modelo as variáveis que possuem dados faltantes. Neste caso, não há problemas de enviesamento da base de construção, entretanto, pode resultar em um modelo com um poder preditivo inferior ao que seria obtido caso todas as variáveis fossem testadas.

Para resolver problemas como estes, métodos mais sofisticados têm sido desenvolvidos durante as últimas décadas. A maior parte deles tem como objetivo preencher os dados faltantes, possibilitando assim realizar a análise com a base de dados completa, ou seja, com todos os indivíduos e variáveis.

O mais simples dos métodos e que também está disponível na maioria dos pacotes estatísticos, é a substituição por alguma medida resumo, no qual os *missings* são substituídos pela média (ou mediana) dos dados válidos (não *missings*) da variável em questão. Porém, pelo fato de o mesmo valor estar sendo atribuído para cada um dos dados faltantes, este método substitui artificialmente a variância da variável em questão, além de diminuir a relação com as demais variáveis.

Outras técnicas também realizam imputações, entretanto o dado faltante é substituído por valores previstos utilizando as informações (não *missings*) das demais variáveis, gerando assim estimativas mais robustas do dado faltante. São os chamados processos de imputação única. Para realizar tal estimativa são utilizadas técnicas como regressão linear, algoritmos EM (*expectation-maximization*), regressão multinomial, entre outras, variando conforme o tipo da variável (categórica ou contínua).

De qualquer forma, sempre haverá um erro gerado por este processo de estimação, e tal incerteza precisa ser considerada durante a aplicação dos resultados gerados pela base completa imputada. Foi com este objetivo que surgiram técnicas como a imputação múltipla (Rubin, 1987), que consiste, resumidamente, em repetir algum destes processos de imputação várias vezes, produzindo diversos bancos de dados imputados. A análise estatística desejada (no nosso caso o modelo preditivo) é então realizada em cada um destes “novos” bancos, produzindo diversos resultados. Por fim, estes resultados são então combinados produzindo um resultado final. Esta combinação pode ser feita de diversas formas e varia conforme o estudo em questão.

Além destes, existem outros métodos estatísticos de tratamento de dados faltantes que também têm apresentado um desempenho superior em relação aos métodos mais “tradicionais” (por exemplo, Little e Rubin, 1987; Graham et al., 1997; Schafer e Graham, 2002; Ferreira, 2009; Poletto, 2006; Nunes, 2007), pelo menos no que diz respeito à avaliação dos parâmetros do modelo. Muitos destes métodos, assim como na imputação múltipla, não se concentram somente na identificação de um substituto para o dado faltante, mas também em utilizar todas as informações disponíveis para preservar as relações existentes no conjunto de dados. Entretanto, alguns destes métodos necessitam de técnicas computacionais mais sofisticadas e

pelo fato de não estarem implementados na maioria dos *softwares* estatísticos, sua utilização acaba sendo restringida.

1.1.1 Tipos de Dados Faltantes

Outro aspecto importante da análise dos dados é que antes de aplicarmos qualquer método de tratamento de dados faltantes, é necessário entender o motivo pelo qual surgiram estes dados faltantes. Graham et al. (2003) descreve que, de maneira informal, os *missings* podem ser causados por combinações de três motivos: processos aleatórios, processos mensuráveis e processos não mensuráveis. Os métodos de tratamento de dados faltantes geralmente funcionam para as duas primeiras situações, mas não para a última. Mais formalmente, os motivos de aparecimento de *missings* são comumente classificados como estando em uma das três categorias abaixo, descritas por Little e Rubin (1987).

i) *Missing Completely at Random* (MCAR): nesta situação, as observações *missings* não são diferentes das não-*missings* em termos da análise realizada. Neste caso, os *missings* surgiram de maneira aleatória e, portanto, o único problema gerado pelos dados faltantes é a perda de poder da análise a ser realizada;

ii) *Missing at Random* (MAR): neste caso, os dados faltantes dependem das variáveis preenchidas e, portanto, podem ser totalmente explicadas pelas demais variáveis presentes no banco de dados. Logo, ao realizar o tratamento dos *missings* de forma que sejam consideradas as informações que “causam” os *missings*, é possível realizar uma análise não viesada;

iii) *Missing Not at Random* (MNAR): nesta situação os *missings* são gerados de forma não mensurável, ou seja, eles dependem de eventos que o pesquisador não consegue observar e controlar. Este é o caso mais grave, em que para tratamento dos dados faltantes, em alguns casos, são necessárias técnicas mais complicadas.

De forma resumida, é possível dividir os motivos de geração dos dados faltantes em os que podem e os que não podem ser explicados. Os que podem ser

explicados pelas informações presentes englobam os MCAR e a maioria dos MAR, já os que não podem englobam os MNAR e a parte restante dos MAR.

Na prática, o motivo dos dados faltantes raramente é composto apenas por eventos não mensuráveis. Normalmente, os *missings* são gerados tanto por fatores que podem como por fatores que não podem ser explicados. Entretanto, métodos adequados de tratamento de dados faltantes tendem a apresentar bons resultados mesmo em situações na qual o pesquisador não tem certeza de que os dados faltantes foram causados puramente por motivos mensuráveis (Graham et al., 1997; Little, 1995), possibilitando assim sua aplicação na maioria dos casos.

1.2 OBJETIVOS DO ESTUDO

Como observado, há um constante avanço na forma de tratamento dos dados faltantes, sendo possível encontrar diversos materiais, em diferentes áreas, relatando estudos em que tais técnicas auxiliaram os pesquisadores a obterem resultados mais confiáveis.

Por outro lado, durante o desenvolvimento de modelos preditivos, o mercado costuma recorrer a métodos mais simples, sendo os mais comuns:

- a) Descarte das variáveis que possuem dados faltantes;
- b) Como o mercado costuma trabalhar apenas com variáveis categóricas (inclusive categorizando as originalmente contínuas), os dados faltantes são tratados como mais uma categoria da variável, ou então os dados faltantes são agrupados à categoria com comportamento mais parecido segundo a variável resposta.

Portanto, este trabalho pretende contribuir para a difusão do uso de outras técnicas de tratamento de dados faltantes durante a construção de modelos preditivos, além de estimular a discussão se tais técnicas efetivamente trarão ganhos efetivos no que diz respeito ao poder preditivo dos modelos. Para isso, através de uma base real, a qual foi utilizada para construção de um modelo preditivo, serão desenvolvidos diversos modelos cada um deles utilizando um tipo de tratamento de dados faltantes, e, então, seus resultados serão comparados.

Este trabalho se divide da seguinte forma. No Capítulo 2 há uma descrição dos dados utilizados. Já no Capítulo 3 são apresentadas as técnicas a serem testadas, enquanto o Capítulo 4 descreve os indicadores a serem considerados na avaliação dos modelos ajustados. O Capítulo 5 mostra os resultados da aplicação das técnicas na base real e, por fim, o Capítulo 6 é utilizado para conclusões e considerações finais.

2 DESCRIÇÃO DO ESTUDO

2.1 DESCRIÇÃO DOS DADOS

Neste trabalho apresenta-se uma aplicação empírica do problema. Trata-se do desenvolvimento de um modelo preditivo, mais especificamente de *Credit Score*. Para tal desenvolvimento será utilizada a técnica de regressão logística, na qual dentre suas variáveis explicativas, uma delas possui uma quantidade representativa de dados faltantes, de tal forma que o objetivo deste estudo será testar diferentes técnicas de tratamento destes valores faltantes, avaliando o impacto de cada técnica no poder preditivo do modelo.

Para este estudo, utilizou-se uma base de dados gentilmente cedida por uma instituição financeira, e os resultados são apresentados no Capítulo 5. Tal base contém registros de pessoas físicas que contrataram um determinado tipo de empréstimo com garantia e foi utilizada para desenvolvimento de um modelo de *Credit Score* para esse produto.

Todos os 124.354 financiamentos utilizados na modelagem correspondem à produção desta instituição financeira, neste tipo de empréstimo, entre janeiro de 2009 e janeiro de 2010. Destes, os 84.987 contratos iniciados entre janeiro e setembro de 2009 foram utilizados para construção dos modelos, enquanto os 39.367 contratos iniciados entre outubro de 2009 e janeiro de 2010 para validação dos mesmos.

Para todos os contratos foi possível observar o comportamento dos indivíduos em suas doze primeiras parcelas, e assim construir a variável resposta. Foram classificados como inadimplentes (ou maus) aqueles que atrasaram o pagamento de pelo menos uma das doze primeiras parcelas por mais de 90 dias, ou que, durante os doze primeiros meses, devolveram a garantia, indicando assim que não possuíam condições financeiras para honrar o contrato. Os demais indivíduos foram classificados como adimplentes (ou bons). As distribuições da variável resposta nas bases de construção e validação podem ser vistas na Tabela 2.1.1 e Tabela 2.1.2, respectivamente.

Tabela 2.1.1: Distribuição da variável resposta na base de construção

Categorias	Total	% rep
Bom	77.511	91,2%
Mau	7.476	8,8%
Total	84.987	100,0%

Tabela 2.1.2: Distribuição da variável resposta na base de validação

Categorias	Total	% rep
Bom	35.807	91,0%
Mau	3.560	9,0%
Total	39.367	100,0%

Como variáveis explicativas, foram considerados os dados cadastrais do cliente e também informações sobre o financiamento e a garantia adquirida. Tais variáveis são apresentadas na lista abaixo.

Informações do Contrato de Financiamento

Valor de Entrada

Valor do Financiamento

Quantidade de Parcelas

Valor da Prestação

Quantidade de Dias de Carência

Percentual de Entrada

Informações da Garantia

Idade da Garantia

Ano de Fabricação

Quantidade de Cilindradas

Valor de Mercado

Marca

Modelo

Informações do Proponente

Flag de Ex-Cliente da Instituição Financeira

Valor Total de Receitas

Idade do Proponente

Sexo

Quantidade de Veículos de sua Posse
Quantidade de Meses de Conta Corrente
Quantidade de Meses no Atual Emprego
CEP
Estado Civil
Tipo de Atividade Profissional
Tipo de Residência

Dentre as variáveis explicativas, a de maior interesse é a que informa a quantidade de meses de conta corrente do indivíduo. Como a solicitação do financiamento é feita via preenchimento de um formulário, e este é um campo de preenchimento não obrigatório, muitos proponentes o deixam em branco, gerando assim os *missings*. No total, este campo possui 9.776 valores faltantes (7,9% da base). Além disso, esta informação é bastante correlacionada com a variável resposta, conforme será visto na Subseção 5.1, costumando estar presente nos modelos com bastante importância. Portanto, a forma de tratamento dos dados faltantes é fundamental para o sucesso do modelo, ou seja, interferirá diretamente no poder preditivo do modelo.

Por solicitação da instituição financeira, dentre as variáveis explicativas, apenas a que possui *missings* poderá ser identificada nas demais seções deste trabalho. As demais serão tratadas através de códigos (var1, var2, e etc.). Além disso, para manter em sigilo a identidade de cada indivíduo, o CPF (Cadastro de Pessoa Física) não foi liberado.

3 METODOLOGIA

Neste capítulo, inicialmente será apresentado um resumo sobre a técnica de regressão logística, a qual foi utilizada para o desenvolvimento dos modelos relacionados ao estudo empírico contido neste trabalho. Trata-se de uma técnica amplamente aceita e utilizada pelo mercado na construção de modelos preditivos.

Posteriormente serão descritas as técnicas de tratamento de dados faltantes que foram testadas neste estudo, sendo que cada uma das técnicas foi utilizada para desenvolvimento de um modelo sobre a base de dados descrita no Capítulo 2.

3.1 REGRESSÃO LOGÍSTICA

A técnica de regressão logística é conhecida desde a década de 1950, porém foi através de Cox (1970) que ela passou a ser mais utilizada em estudos estatísticos (ver também Cox e Snell, 1989). Ela é amplamente usada em ciências médicas e sociais. Maiores detalhes sobre esta técnica podem ser encontrados em Paula (2004) e em Agresti (1990).

Assim como para outros tipos de regressão, o principal objetivo da regressão logística é estudar a correlação entre um conjunto de variáveis explicativas com uma variável resposta, com a restrição de que tal variável seja categórica (frequentemente binária), tendo a vantagem de ser relativamente flexível em relação às suas premissas.

Outro ponto favorável à regressão logística é a possibilidade de interpretação direta dos coeficientes estimados como medidas de associação, fazendo com que esta técnica seja utilizada para entender relações de uma ou mais variáveis explicativas com a resposta.

Entretanto, é com objetivo preditivo que o mercado vem utilizando cada vez mais tal técnica. Como o resultado da aplicação deste método reflete a probabilidade de ocorrência de um determinado evento, é feita uma predição no sentido de assumir que tal evento ocorrerá para os registros que possuírem maior probabilidade

e não ocorrerá para os demais. Neste contexto, a regressão logística compete com técnicas como redes neurais, análise discriminante e árvore de decisão. Entretanto, ela acaba sendo a mais utilizada por apresentar bom desempenho preditivo e ao mesmo tempo tender a fornecer modelos mais interpretáveis.

Para explicação do método, considere uma variável binária (*bons* e *maus*) sobre a qual é desejado classificar um conjunto de indivíduos. Desta forma, definimos a variável resposta $Y \in \{0,1\}$ como:

$$Y_i = \begin{cases} 0, & \text{se o } i - \text{ésimo cliente é bom} \\ 1, & \text{se o } i - \text{ésimo cliente é mau} \end{cases}$$

Considere a matriz X contendo as $(p-1)$ variáveis explicativas para cada um dos n indivíduos da amostra. Sendo então $X_i = (1, X_{i2}, X_{i3}, \dots, X_{ip})^T$ o vetor correspondente à i -ésima linha desta matriz X , e X_{ij} o elemento correspondente a j -ésima variável ($j = 2, \dots, p$) para o i -ésimo indivíduo ($i = 1, \dots, n$). Considere também $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$ o vetor de parâmetros desconhecidos, sendo β_1 o intercepto e β_j o parâmetro associado a X_{ij} .

Em um modelo de regressão logística múltipla a probabilidade de o indivíduo ser *mau* é expressa por:

$$\pi(X_i) = P(Y_i = 1|X_i) = \frac{\exp(X_i^T \beta)}{1 + \exp(X_i^T \beta)} = \frac{\exp(\beta_1 + \beta_2 X_{i2} + \dots + \beta_p X_{ip})}{1 + \exp(\beta_1 + \beta_2 X_{i2} + \dots + \beta_p X_{ip})},$$

e a probabilidade do indivíduo ser *bom* é dada por:

$$1 - \pi(X_i) = P(Y_i = 0|X_i) = \frac{1}{1 + \exp(X_i^T \beta)} = \frac{1}{1 + \exp(\beta_1 + \beta_2 X_{i2} + \dots + \beta_p X_{ip})},$$

sendo $\pi(x_i)$ a probabilidade condicional da observação $Y_i = 1$, $i = 1, \dots, n$ dados os valores das variáveis explicativas X_i .

Portanto, a função de verossimilhança pode ser escrita da seguinte maneira:

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

Assumindo que Y_i tem distribuição Bernoulli(π_i), sendo $\pi_i = \pi(X_i)$ e usando a função de ligação *logit*, dada por:

$$\ln \left[\frac{\pi_i}{1-\pi_i} \right] = \beta_1 + \sum_{j=2}^p \beta_j X_{ij},$$

é mais simples trabalhar com o logaritmo da função de verossimilhança, dada por:

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^n y_i (X_i^T \beta) - \sum_{i=1}^n \ln [1 + \exp (X_i^T \beta)].$$

Para estimar os parâmetros relativos ao vetor β é possível utilizar o método de máxima verossimilhança, encontrando o valor de β que maximiza $l(\beta)$. Para isso o método mais utilizado é o de Newton-Raphson (Cox, 1975).

Há diversos *softwares* estatísticos que disponibilizam a técnica de regressão logística, dentre os quais estão: *SAS*, *SPSS*, *S*, *R*, *MINITAB*, entre outros. Neste estudo todos os modelos foram desenvolvidos utilizando o *SPSS*.

3.2 MÉTODOS DE TRATAMENTO DE DADOS FALTANTES

Nesta seção serão apresentados os modelos desenvolvidos sobre a base de dados real obtida, a qual foi descrita no Capítulo 2, cada um utilizando uma forma de tratamento de dados faltantes. Certamente há diversas técnicas que poderiam ser testadas, entretanto este trabalho limitou-se a testar sete metodologias, as quais foram escolhidas segundo dois critérios: aplicação pelo mercado e difusão na literatura.

3.2.1 Método 1: *Missings* como uma Categoria

Neste cenário os dados faltantes são tratados como qualquer outra categoria, ou seja, lhe atribuindo uma variável indicadora ou eliminando a categoria. Neste

estudo será aplicada a segunda situação, transformando assim a categoria dos *missings* como a de referência.

Este tipo de tratamento é comumente aplicado pelo mercado e tem a vantagem de atribuir uma pontuação específica para os dados faltantes, entretanto fica mais sensível a mudanças no perfil dos missings.

3.2.2 Método 2: Descarte da Variável com Dados Faltantes

Durante o desenvolvimento de modelos preditivos, alguns preferem descartar as variáveis que possuem dados faltantes. A vantagem deste tratamento é que desvincula o modelo de futuras mudanças no perfil dos *missings*. Por exemplo, em uma situação em que um campo de preenchimento facultativo se torna obrigatório, o modelo ficaria intacto. Entretanto, ao realizarem tal artifício podem estar deixando de considerar uma informação valiosa que incrementaria o desempenho preditivo do modelo. Em uma situação extrema, é possível que todas as principais variáveis explicativas do modelo contenham dados faltantes, assim a sua não utilização inviabilizaria a construção do modelo.

3.2.3 Método 3: Imputação via Regressão

A partir deste modelo algumas técnicas conhecidas de imputação de dados faltantes serão utilizadas durante o desenvolvimento do modelo. As etapas para seu desenvolvimento estão descritas abaixo.

3.2.3.1 Resumo das técnicas necessárias - regressão linear múltipla

Esta técnica é vastamente difundida na literatura, sendo utilizada em estudos de diversas áreas, portanto, este trabalho fará apenas uma breve explanação sobre ela, sendo possível se aprofundar no tema em (Montgomery et al., 2001).

A regressão linear múltipla possibilita avaliar o relacionamento de um conjunto de variáveis explicativas com uma variável de interesse contínua, sendo também utilizada para estimar o valor esperado desta variável de interesse (também chamada de variável resposta ou dependente). Seu resultado final é uma equação que representa a melhor predição da variável dependente a partir das variáveis independentes. Tal equação corresponde a um modelo aditivo e é utilizada para estimar o valor esperado para cada unidade amostral.

$$Y_i = \alpha + \sum_{j=1}^k \beta_j X_{ij} + \epsilon_i, \quad i = 1, \dots, n,$$

em que:

Y_i : valor da variável dependente para o i -ésimo elemento da amostra;

α e β_j : parâmetros desconhecidos;

k : número de variáveis independentes no modelo;

X_{ij} : valor (conhecido) da variável independente j para o i -ésimo elemento da amostra;

ϵ_i : erro amostral atrelado ao i -ésimo elemento da amostra.

Os valores dos parâmetros α e β_j ($j = 2, \dots, k$) são estimados utilizando o método de mínimos quadrados, que procura encontrar o melhor ajustamento no sentido de minimizar a soma dos quadrados das diferenças entre o valor predito e o valor observado.

Um modelo de regressão linear múltipla depende de algumas suposições básicas, as quais são apresentadas abaixo.

I – a variável dependente é função das variáveis independentes;

II – os valores das variáveis independentes são fixos;

III – $E(\epsilon_i) = 0$;

IV – os erros são homocedásticos, isto é, $E(\epsilon_i)^2 = \sigma^2$;

V – os erros são independentes e têm distribuição normal.

3.2.3.2 Etapas de desenvolvimento do modelo

Etapas 1: os dados faltantes para a variável quantidade de meses de conta corrente são substituídos por valores imputados via regressão linear múltipla, ajustada sobre os dados completos;

Etapas 2: a variável que sofreu imputações é novamente categorizada;

Etapas 3: o modelo de regressão logística é ajustado.

3.2.4 Método 4: Imputação via Algoritmo EM

3.2.4.1 Resumo das técnicas necessárias - *expectation maximization* (EM)

O nome *Expectation Maximization* (EM) e a formalização deste método foram propostos por Dempster et al. (1977). Antes disso, diversos autores já haviam utilizado métodos em trabalhos de diferentes áreas, os quais aparentemente não eram relacionados. Basicamente, o algoritmo EM consiste em encontrar estimativas de máxima verossimilhança de um determinado conjunto de parâmetros em um modelo estatístico, nos quais o modelo depende de variáveis não observadas, se adequando a situações nas quais a variável é preenchida em alguns casos e em outros não.

Neste trabalho será apresentada uma resumida descrição do algoritmo EM, maiores detalhes podem ser encontrados em Dempster et al. (1977).

Este método tende a ser utilizado quando o objetivo é estimar um conjunto de parâmetros θ , que obedece a uma determinada distribuição de probabilidades, a qual é obtida utilizando apenas uma parcela dos dados. De maneira mais formal,

considere que $A = \{A_1, A_2, \dots, A_m\}$ represente as m unidades amostrais para as quais os valores são conhecidos, e $B = \{B_1, B_2, \dots, B_n\}$ as não conhecidas.

Assim, seja θ o vetor de parâmetros desconhecidos, com sua função de verossimilhança $L(\theta; A, B) = p(A, B/\theta)$. Com isso, temos que a estimativa de máxima verossimilhança dos parâmetros desconhecidos é determinada pela probabilidade marginal dos dados observados.

$$L(\theta; A) = p(A|\theta) = \sum_B p(A, B|\theta)$$

Entretanto, na maioria dos casos, este cálculo é de extrema complexidade. Assim, o algoritmo EM procura encontrar a estimativa de verossimilhança iterativamente, aplicando os dois passos seguintes:

Expectation: calcular o valor esperado do logaritmo da função de verossimilhança em relação a distribuição condicional de A dado B sob a atual estimativa dos parâmetros $\theta^{(c)}$.

$$Q(\theta; \theta^{(c)}) = E_{B|A, \theta}[\log L(\theta; A, B)]$$

Com isso, os dados ausentes são preenchidos.

Maximization: encontra o valor do parâmetro θ que maximiza a verossimilhança baseada em todos os dados:

$$\theta^{(c)} = \underset{\theta}{\operatorname{argmax}} Q(\theta; \theta^{(c)})$$

Os dois passos serão repetidos até que haja convergência e os valores encontrados na última iteração são os imputados.

3.2.4.2 Etapas de desenvolvimento do modelo

Este método possui etapas de desenvolvimento semelhantes ao Método 3, com a diferença que na Etapa 1 as imputações são realizadas utilizando estimativas de máxima verossimilhança obtidas através do método iterativo EM (*expectation-maximization*), descrito na Subseção 3.2.4.1.

3.2.5 Método 5: Imputação Múltipla

Para o desenvolvimento deste modelo será utilizada a técnica de imputação múltipla, da qual uma breve descrição é apresentada a seguir e que maiores informações podem ser obtidas em Wayman (2003) ou em Nunes (2007).

3.2.5.1 Resumo das técnicas necessárias - imputação múltipla

Na imputação múltipla os dados faltantes de cada variável são previstos utilizando as informações (não *missings*) das demais variáveis. Estes valores previstos substituem os valores *missings*, resultando assim em um banco de dados completamente preenchido, chamado de banco de dados imputado. Este procedimento é repetido várias vezes, produzindo vários bancos de dados imputados. A análise estatística desejada é realizada em cada um destes “novos” bancos, produzindo diversos resultados. Por fim, estes resultados são então combinados produzindo um resultado final.

Através da imputação múltipla é possível não só restaurar a variabilidade natural dos dados, como também incorporar a incerteza causada pela estimativa feita sobre os dados faltantes. A manutenção da variabilidade original é obtida através da criação dos valores imputados, que se baseiam em variáveis correlacionadas com o motivo de geração dos dados faltantes. Já a incerteza é obtida pela criação de diferentes bancos de dados imputados e observando a variabilidade entre estes conjuntos de dados.

É importante destacar que os valores imputados não têm o objetivo de “adivinhar” o valor de um determinado *missing*, e sim de criar um conjunto de dados

que preserve a variabilidade geral da população e mantenha as relações existentes entre as variáveis. Assim, ao realizar o método de imputação múltipla o interesse é preservar as características importantes do banco de dados como um todo (por exemplo: média, variância, parâmetros de uma regressão e etc.), sendo a imputação apenas um mecanismo para possibilitar uma análise que faz uso de todas as informações possíveis.

Este método representa uma boa alternativa para tratamento de dados faltantes, porque contempla ao mesmo tempo qualidade dos dados e uma relativa facilidade de uso. Ele tem sido estudado e aplicado em diversas situações e os resultados têm sido bastante favoráveis, se mostrando capaz de produzir estimativas que reflitam as incertezas associadas às imputações sobre os dados faltantes. Além disso, tem se mostrado robusto para eventuais desvios das premissas relacionadas à normalidade dos dados e fornecido resultados adequados em estudos que possuem amostras de tamanho reduzido ou alta proporção de dados faltantes (Schafer e Graham, 2002).

Este procedimento é computacionalmente mais simples que outros métodos estatísticos e, como será mostrado mais adiante, é uma técnica intuitiva e de fácil entendimento. Embora seus princípios estatísticos não sejam triviais, existem *softwares* de fácil uso que contemplam todos os procedimentos necessários.

Abaixo será apresentado apenas um resumo da imputação múltipla. Algo mais profundo pode ser encontrado em Graham et al., 2003; Little e Rubin, 1987.

A imputação múltipla é um processo bastante intuitivo que basicamente é composto por três passos.

1) Imputação de dados: são gerados m bancos de dados completos através de técnicas adequadas de imputação;

2) Análise dos bancos de dados gerados: separadamente, os m bancos são analisados por um método estatístico qualquer, escolhido conforme o objetivo do estudo em questão;

3) Combinação dos resultados: os m resultados obtidos no passo 2 são combinados de forma simples e adequada para se obter a inferência final do parâmetro de interesse.

Passo 1 – Métodos de imputação dos dados

Esta é a parte fundamental da imputação múltipla, pois a técnica de imputação deve levar em conta ao máximo a relação entre os dados faltantes e os observados. Diversas técnicas podem ser utilizadas para realizar este passo. Neste trabalho é dado um breve resumo de uma destas técnicas: regressão linear bayesiana, a qual foi utilizada na aplicação empírica contida neste trabalho. Tal descrição pode ser vista com mais detalhes em Nunes (2007).

Neste método um modelo de regressão linear bayesiana tendo como resposta a variável que possui dados faltantes (Y) e como variáveis preditoras são utilizadas as demais variáveis presentes no banco de dados ($X_1, X_2, \dots, X_k$), onde k é o número de variáveis preditoras.

Portanto teremos um modelo com a seguinte estrutura:

$$Y_i \sim N(X_i \beta, \sigma^2).$$

a qual é a especificação para $f(Y_i/X_i, \theta)$, $\theta = (\beta, \log \sigma)$, em que $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ é um vetor de parâmetros, e σ é um escalar. Assume-se então uma distribuição *a priori* não informativa para θ e que $n > k$, sendo n o total de indivíduos na base.

Rubin (1987) mostra que na distribuição *a posteriori*:

- θ depende apenas das unidades observadas de Y ;
- β dado σ^2 é normal com média $\hat{\beta}_*$ e matriz de covariância $\sigma^2 V$, em que:

$$\hat{\beta}_* = V [\sum_{i=1}^n X_i^t Y_i^t],$$

$$V = [X_i^t X_i]^{-1};$$

- σ^2 é $\hat{\sigma}_*^2(n - k)$ dividido por uma variável aleatória χ_{n-k}^2 , em que:

$$\hat{\sigma}_*^2 = \frac{\sum_{i=1}^n (Y_i - \hat{\beta}_*)^2}{(n-k)}.$$

Dessa forma, tendo a distribuição *a posteriori* de θ descrita em termos de distribuições conhecidas, torna-se possível estimar os parâmetros a serem utilizados na imputação. E o processo de imputação se resume aos seguintes procedimentos:

- a) Simular uma variável aleatória χ_{n-k}^2 , denotada como g , e seja

$$\hat{\sigma}_{**}^2 = \frac{\hat{\sigma}_*^2(n-k)}{g};$$

b) Simular k variáveis independentes seguindo uma distribuição $N(0,1)$ para criar um vetor Z de k componentes e seja

$$\beta_* = \hat{\beta} + \sigma_{**}[V]^{1/2}Z,$$

em que $[V]^{1/2}$ é a raiz quadrada de V tal como a raiz quadrada obtida pela decomposição de Cholesky;

c) Gerar os m diferentes valores para cada dado faltante da variável Y . Por exemplo, para imputar os dados faltantes para a observação r utiliza-se:

$$Y_{rs} = X_r\beta_* + Z_{rs}\sigma_{**}, \quad s = 1, \dots, m$$

sendo que os m desvios normais Z_{rs} são simulados independentemente.

Ressaltando que, além da regressão linear bayesiana, outros métodos podem ser utilizados no passo 1 da imputação múltipla, dentre os quais se destaca o MCMC (Markov Chain Monte Carlo) que tem como objetivo simular distribuições multivariadas cujo limite é uma cadeia de Markov estacionária que tem a distribuição que se deseja encontrar (Gilks et al., 1996).

Passo 2 – Análise dos bancos de dados gerados

Uma vez que os conjuntos de dados imputados foram criados, a análise de interesse é realizada separadamente para cada um dos m bancos de dados. Esta análise é a que seria feita se não existissem os *missings* (por exemplo, modelos de regressão, ANOVA, análises descritivas e etc.). No caso da aplicação empírica contida neste trabalho, a técnica de interesse é uma regressão logística.

Passo 3 – Combinação dos resultados

Com as análises realizadas, tudo o que resta é combinar seus resultados para produzir um resultado final. Para isso utilizam-se as regras de Rubin as quais estão amplamente divulgadas na literatura, pois combinam os resultados através de regras simples e podem ser utilizadas independentemente do método utilizado para imputação dos dados faltantes (Passo 1).

A partir das análises realizadas no Passo 2 são obtidas m estimativas para o parâmetro de interesse D , ou seja, D_s , para $s=1,2,\dots,m$. Esta estimativa pode ser qualquer medida escalar como por exemplo médias, variâncias, correlações ou coeficientes de regressão. Então, a estimativa combinada será a média das estimativas individuais.

$$\bar{D} = \frac{1}{m} \sum_{s=1}^m \hat{D}_s.$$

A variância combinada é igualmente intuitiva, mas requer mais cálculos. Para os m bancos de dados imputados, a variância combinada de $\bar{D}$ é obtida através da fórmula:

$$T = \bar{E} + \left(1 + \frac{1}{m}\right) F,$$

em que, $\bar{E} = \frac{1}{m} \sum_{s=1}^m E_s$ é a média das variâncias (E) do parâmetro de interesse nos m bancos de dados e $F = \frac{1}{(m-1)} \sum_{s=1}^m (\hat{D}_s - \bar{D})^2$. Traduzindo, vemos que a variância combinada é composta por dois fatores que devemos considerar: um componente que preserva a variabilidade natural ($\bar{E}$) e um componente adicional que estima a incerteza causada pelos dados faltantes (F). Ao observar F , que é a parte da equação relacionada à variabilidade “entre” as imputações, percebe-se que se as estimativas variam muito de um conjunto de dados para outro, então a incerteza gerada pelas imputações é alta e F será grande. Se, no entanto, as estimativas dos parâmetros são todas muito semelhantes, há menos incerteza, e F será pequeno.

3.2.5.2 Etapas de desenvolvimento do modelo

Etapas 1: desenvolve-se um modelo preditivo, via regressão logística, considerando apenas os registros completamente preenchidos, ou seja, eliminando os dados faltantes na variável com dados faltantes.

Etapa 2: através de alguma análise define-se o chamado ponto de corte, a partir do qual o cliente é marcado como mau pagador, em outras palavras, é o ponto a partir do qual a instituição financeira aprova a liberação do crédito (isso para modelos de *Credit Score*, para demais tipos de modelos preditivos o raciocínio é análogo) .

Etapa 3: são gerados m bancos de dados completos através de técnicas adequadas de imputação (regressão, EM, entre outros). No exemplo empírico descrito no Capítulo 5, adotou-se $m = 3$.

Etapa 4: o modelo desenvolvido na Etapa 1 é aplicado nas m bases geradas na Etapa 3, gerando assim m diferentes *scores* para cada registro.

Etapa 5: após a Etapa 4, cada registro possuirá m diferentes *scores*, e, após a comparação deles com o ponto de corte definido na Etapa 2, m “decisões” (aprovação ou rejeição). A “decisão final” será a predominante dentre as m “decisões” (por isso recomenda-se que m seja um número ímpar).

3.2.6 Método 6: Estimação de Probabilidades

Assim como acontece no Método 5, neste método pretende-se considerar a incerteza gerada durante o tratamento dos dados faltantes. Para isso, para cada unidade amostral com dado faltante, são estimadas as probabilidade de este registro pertencer a cada uma das categorias da variável com *missings* (assume-se que a variável que possui *missings* é categorizada durante o desenvolvimento de um modelo preditivo, seguindo assim uma prática comum no mercado). Estas probabilidades são utilizadas para estimar o *score* final de cada uma das unidades amostrais da base. Este processo é apresentado abaixo:

Etapa 1: desenvolve-se um modelo preditivo considerando apenas os registros completamente preenchidos, eliminando assim as unidades amostrais que possuam dados faltantes. Tal modelo é desenvolvido da mesma forma que seria feito caso não existissem os dados faltantes.

Etapa 2: para realizar a Etapa 1 foi necessário categorizar a variável com dados faltantes. Desta forma, nesta etapa estima-se a probabilidade de cada registro

pertencer a cada uma destas categorias. Ou seja, desenvolve-se um novo modelo, no qual a variável resposta é a que possui dados faltantes e as demais variáveis do banco de dados são utilizadas como explicativas. Para desenvolvimento deste segundo modelo, novamente, são utilizados apenas os casos completamente preenchidos. Aqui, podem ser utilizadas técnicas como regressão multinomial (ou logística, no caso da variável possuir apenas duas categorias) e análise discriminante. Na aplicação apresentada no Capítulo 5 foi utilizada análise discriminante (Agresti, 1990).

Etapa 3: sendo u o número de categorias da variável com dados faltantes categorizada, aplica-se u vezes o modelo construído na Etapa 1, simulando assim a pontuação obtida pelo modelo em todos os possíveis cenários.

Etapa 4: calcula-se um “score final” através da soma dos produtos obtidos pela multiplicação das probabilidades obtidas na Etapa 2 e suas respectivas pontuações obtidas na Etapa 3, conforme ilustrado abaixo.

$$score_{final} = \sum_{l=1}^u prob_l score_l,$$

em que:

u : número de categorias da variável com dados faltantes;

$prob_l$: probabilidade estimada de o registro pertencer à categoria l da variável com dados faltantes;

$score_l$: pontuação obtida pelo modelo construído na Etapa 1 caso o registro pertença a categoria l .

3.2.7 Método 7: Agrupamento Conceitual dos Dados Faltantes

Trata-se de uma metodologia simples no qual o objetivo é tentar inferir conceitualmente em qual das categorias os dados faltantes estariam caso a informação faltante fosse conhecida. Evidentemente, para ser possível aplicar tal técnica é necessário conhecer profundamente a variável que possui dados faltantes, entendendo a sua forma de preenchimento, para assim poder supor os motivos

pelos quais os *missings* surgiram. Tendo alguma hipótese, os dados faltantes são alocados na tal categoria e o modelo é construído normalmente.

No caso do exemplo empírico deste estudo, a variável com dados faltantes é tempo de conta corrente (em meses). Por se tratar de um campo de preenchimento não obrigatório, muitos proponentes ao crédito preferem não responder, gerando assim os *missings*. O que se supõe é que tais indivíduos não preenchem esse campo, pois na realidade possuem pouco tempo de conta corrente, e assim, possivelmente instruídos por alguém que conhece o processo de avaliação de crédito, por exemplo, o dono do estabelecimento que está vendendo o bem, imaginam que ao informar, suas chances de aprovação diminuam.

Portanto, dada esta suposição, foi construído o Modelo 7, no qual os dados faltantes foram agrupados na categoria de menor tempo de conta corrente.

4 AVALIAÇÃO DAS TÉCNICAS

Este capítulo tem como objetivo apresentar as medidas que serão utilizadas na avaliação de desempenho dos modelos desenvolvidos com as diferentes técnicas de tratamento de dados faltantes. Para isso, os indicadores utilizados foram as medidas de Kolmogorov-Smirnov (KS), a curva ROC (*Receiving Operational Characteristic*), o coeficiente de Gini e a estimativa da curva de aprovação obtida em cada modelo. A fim de facilitar na descrição destas medidas de desempenho denotaremos as duas categorias da variável resposta como *bons* e *maus*, entretanto, as medidas são válidas para todos os modelos preditivos para os quais o evento a ser previsto é binário (adquiri/não adquirir, fraude/não fraude, entre outros).

4.1 KS

A estatística de Kolmogorov-Smirnov é, provavelmente, o indicador mais utilizado pelo mercado na avaliação de modelos preditivos cuja variável resposta é binária. Este indicador também é vastamente utilizado em outras áreas como teste não paramétrico a fim de avaliar se duas amostras são provenientes da mesma função de distribuição (Conover, 1999). Basicamente, esta estatística é obtida calculando a máxima diferença entre as distribuições acumuladas do *score* (pontuação resultante do modelo em questão) entre os *bons* e *maus*, como vemos abaixo. Portanto, trata-se de uma medida que varia entre 0 e 1, e valores mais altos indicam uma melhor performance.

$$KS = \max_{0 \leq x \leq \infty} |F_{mau}(score) - F_{bom}(score)|, \quad 0 \leq KS \leq 100\%$$

em que:

$F_{mau}(score)$: frequência acumulada dos *scores* dos clientes *maus*;

$F_{bom}(score)$: frequência acumulada dos *scores* dos clientes *bons*.

A Tabela 4.1.1 tem como objetivo facilitar o entendimento do processo de obtenção deste indicador, mais especificamente o KS do Modelo 1, que trata os dados faltantes como uma categoria (mais detalhes sobre o modelo serão vistos no Capítulo 5). Antes de gerar esta tabela é necessário criar faixas de *scores*; para isso a base ordenada foi dividida em vinte partes de mesma frequência, de modo que cada uma contivesse 5% dos indivíduos, sendo a primeira com os 5% de menor *score* e a última com os 5% de *score* mais alto.

Tabela 4.1.1: Tabela auxiliar para obtenção do KS

Partes	% de Bons	% de Maus	% acum. de Bons	% acum. de Maus	Diferença
1	5,4	0,6	5,4	0,6	4,8
2	5,5	1,3	10,9	1,9	9,0
3	5,2	1,6	16,1	3,5	12,6
4	5,3	1,9	21,4	5,4	16,0
5	5,2	2,1	26,7	7,6	19,1
6	5,3	2,1	32,0	9,7	22,3
7	5,2	2,7	37,2	12,4	24,8
8	5,2	2,9	42,4	15,3	27,1
9	5,1	3,5	47,5	18,8	28,7
10	5,1	3,8	52,6	22,6	30,0
11	5,1	4,2	57,7	26,8	30,9
12	5,0	4,7	62,7	31,5	31,2
13	5,0	4,8	67,8	36,4	31,4
14	4,9	5,8	72,7	42,2	30,5
15	4,9	6,6	77,6	48,8	28,7
16	4,7	7,5	82,3	56,3	26,0
17	4,7	8,0	87,0	64,3	22,7
18	4,5	9,3	91,5	73,7	17,8
19	4,5	11,4	96,0	85,1	10,8
20	4,0	14,9	100,0	100,0	0,0

Através das colunas que representam as distribuições acumuladas dos bons e maus podemos representar o KS graficamente, como vemos na Figura 4.1.1.

Figura 4.1.1: Gráfico de KS

É importante ressaltar que, apesar de ser um importante indicador, o KS não deve ser utilizado isoladamente, uma vez que ele pode refletir uma alta discriminação em apenas uma faixa de escore. Portanto, é importante analisá-lo juntamente com outras medidas, como a curva ROC e o coeficiente de Gini.

4.2 CURVA ROC E COEFICIENTE DE GINI

A Curva ROC, também conhecida por Diagrama de Lorentz, surgiu durante a segunda guerra mundial com o uso de radares para detecção de sinais. Esta curva também é utilizada em Psicologia para avaliar a capacidade de indivíduos distinguirem entre estímulo e não estímulo, e em medicina para analisar a qualidade de um determinado teste clínico. Posteriormente, passou a ser utilizada também na área financeira na avaliação de modelos preditivos de crédito.

Para a construção desta curva, após ter ajustado o modelo e ser atribuído um score para cada indivíduo da amostra, define-se o ponto de corte PC , tal que o i -ésimo indivíduo da amostra será classificado como *mau* se $score_i > PC$ e *bom* caso contrário. Após a definição do ponto de corte é possível construir a chamada matriz de confusão, representada na Tabela 4.2.1, a qual servirá como base para as demais medidas a serem apresentadas nesta seção.

Tabela 4.2.1: Matriz de Confusão

Valor Previsto	Classificação Real		
	Mau	Bom	Total
Mau	VM	FM	NM
Bom	FB	VB	NB
Total	RM	RB	n

Os valores a serem representados na tabela são:

VM: número de clientes *maus* classificados como *maus*;

FM: número de clientes *bons* classificados como *maus*;

FB: número de clientes *maus* classificados como *bons*;

VB: número de clientes *bons* classificados como *bons*;

RM: número de clientes *maus*;

RB: número de clientes *bons*;

NM: número de clientes classificados como *maus*;

NB: número de clientes classificados como *bons*;

n: número total de clientes.

Com base nessas definições, é possível definir novas medidas de desempenho, conforme visto abaixo.

Especificidade: proporção de clientes *maus* classificados corretamente.

$$Especificidade = \frac{VM}{RM}$$

Sensibilidade: proporção de clientes *bons* classificados corretamente.

$$Sensibilidade = \frac{VB}{RB}$$

Acurácia: proporção total de acertos.

$$Acurácia = \frac{(VB + VM)}{n}$$

A curva ROC é obtida a partir de todas as matrizes de confusão que podem ser geradas, considerando-se diferentes pontos de corte do modelo. Posteriormente, a especificidade e sensibilidade são representadas em um gráfico, que apresenta os valores da sensibilidade no eixo das ordenadas e o complemento da especificidade ($1 - \text{especificidade}$) no eixo das abscissas.

Desta forma, um modelo perfeito corresponde a uma linha horizontal no topo do gráfico, o que é quase impossível ser alcançado. Do outro lado, a linha diagonal ($x=y$) indica uma classificação aleatória, ou seja, um modelo com um poder preditivo nulo. Na prática, a curva estará entre a linha diagonal e a curva perfeita, sendo que quanto maior a distância da linha diagonal, melhor será o modelo. A Figura 4.2.1 mostra um exemplo da curva ROC, mais especificamente a obtida pelo Modelo 1 (o qual será detalhado no Capítulo 5).

Figura 4.2.1: Curva ROC

Para avaliação dos modelos é comum calcular a área sob a curva ROC. Entretanto, como tal área varia entre 0,5 e 1, é adequado utilizar o coeficiente de Gini, que é dado por duas vezes a área entre a curva ROC e a linha diagonal. Ou seja, sendo ROC a área sob a curva ROC, definimos o coeficiente de Gini da seguinte forma:

$$Gini = 2 \times (ROC - 0,5)$$

Portanto, o coeficiente de Gini resume a mesma informação da curva ROC, porém variando entre 0 e 1.

4.3 CURVA DE APROVAÇÃO

Além dos indicadores propostos nas subseções anteriores, os modelos serão avaliados e comparados através de um gráfico que procura refletir de forma mais simples e direta o desempenho do modelo segundo seu propósito, e que neste trabalho será chamado de curva de aprovação. Inicialmente, é preciso reforçar que, como nossa aplicação empírica é baseada em um modelo de *Credit Score*, os termos utilizados nesta explicação serão referentes a este tipo de modelo. Entretanto, análises semelhantes podem ser feitas para os demais modelos preditivos.

Para cada ponto de corte estabelecido (conceito de ponto de corte definido na Subseção 4.2) é possível calcular uma taxa de aprovação e uma taxa de inadimplência, conforme vemos abaixo.

a) taxa de aprovação (TxAp):

$$TxAp = \frac{\sum_{i=1}^n H_i}{n},$$

em que:

n : número total de indivíduos na amostra;

$H_i = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo possuir score inferior ao ponto de corte estabelecido;} \\ 0, & \text{caso contrário;} \end{cases}$

b) taxa de inadimplência (TxIn):

$$TxIn = \frac{\sum_{i=1}^n M_i}{\sum_{i=1}^n H_i},$$

em que:

n : número total de indivíduos na amostra;

$$H_i = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo possuir escore inferior ao ponto de corte estabelecido;} \\ 0, & \text{caso contrário;} \end{cases}$$

$$M_i = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo for aprovado e ter sido considerado } \textit{mau}; \\ 0, & \text{caso contrário;} \end{cases}$$

O gráfico será construído considerando no eixo das abscissas a taxa de inadimplência e no eixo das ordenadas a taxa de aprovação. Para cada um dos possíveis pontos de corte do modelo são calculados estes dois indicadores, gerando assim um ponto a ser *plotado* no gráfico. A junção destes pontos formará a denominada curva de aprovação. A Figura 4.3.1 mostra um exemplo da curva de aprovação, mais especificamente a obtida pelo Modelo 1 (o qual será detalhado no Capítulo 5).

Quanto mais rápida esta curva se aproximar da linha horizontal no topo do gráfico, melhor será o poder preditivo do modelo.

Através deste tipo de gráfico é possível simular cenários a serem obtidos no modelo a ser aplicado. No caso do modelo representado pelo gráfico acima temos que caso fossem aprovados 40% dos indivíduos, a taxa de inadimplência esperada seria de aproximadamente 3,3%. Já se a aprovação fosse de 60% a inadimplência esperada subiria para 4,6%.

5 APLICAÇÃO

Neste capítulo são apresentados os resultados obtidos da aplicação das técnicas descritas no Capítulo 3 sobre a base apresentada no Capítulo 2, obtida junto a uma instituição financeira. Tratam-se de modelos desenvolvidos utilizando a técnica de regressão logística, e replicam um problema relativo a um modelo de *Credit Score*. Como tal base possui uma variável altamente correlacionada com a variável resposta que possui dados faltantes, serão avaliadas as técnicas de tratamento de dados faltantes segundo o poder preditivo dos modelos.

Maiores detalhes sobre a forma de construção de modelos preditivos podem ser encontradas em Araújo et al. (2006), Ferreira (2008), Alves (2008), Veloso (2006) e Selau (2008).

Primeiramente são descritos os procedimentos de categorização das variáveis explicativas e análise descritiva, seguido pelo processo utilizado para seleção das covariáveis de cada modelo. Posteriormente, são apresentados os ajustes de cada modelo finalizando o capítulo com a comparação dos resultados.

Todos os resultados gerados neste estudo foram obtidos utilizando o *software* SPSS 19.

5.1 ANÁLISE DESCRITIVA

Inicialmente, foram realizadas análises descritivas univariadas das variáveis explicativas através de tabelas de frequências, no caso de variáveis categóricas, e de medidas resumo (média, desvio padrão, mediana, quartis, mínimo e máximo), para as variáveis contínuas.

Seguindo um procedimento utilizado pelo mercado, todas as variáveis explicativas contínuas foram categorizadas. Para isso, contou-se com o auxílio da técnica de árvore de decisão (CHAID – *Chi-squared Automatic Interaction Detector*, ver Kaas, 1980), que é um método exploratório para estudar as relações entre

variáveis, simulando diversos agrupamentos e sugerindo aquele mais vantajoso segundo a estatística qui-quadrado.

Posteriormente, foi analisada a relação de cada uma das variáveis explicativas (agora todas categóricas) com a variável resposta. Para isso, foram calculadas as seguintes estatísticas:

- representatividade percentual de cada categoria da variável explicativa (% rep);
- percentual de maus em cada categoria (% maus);
- risco relativo (RR): razão de riscos entre clientes *maus* e *bons*, cuja forma de cálculo é apresentada abaixo.

$$RR_l = \frac{RepBons_l}{RepMaus_l}, \quad RepBons_l = \frac{bons_l}{bons} \quad \text{e} \quad RepMaus_l = \frac{maus_l}{maus}$$

para $l = 1, 2, \dots, u$ em que:

RR_l : risco relativo de um cliente bom presente na categoria l em relação a um cliente mau;

u : número de categorias da variável em questão;

$RepBons_l$: representatividade da categoria l entre os bons;

$RepMaus_l$: representatividade da categoria l entre os maus;

$bons_l$: número de indivíduos bons na categoria l ;

$bons$: número total de indivíduos bons;

$maus_l$: número de indivíduos maus na categoria l ;

$maus$: número total de indivíduos maus.

A Tabela 5.1.1 apresenta as medidas acima para a variável quantidade de meses de conta corrente. Por ela é visto que os clientes que não informaram sua quantidade de meses possuem uma alta inadimplência (% maus), indicando que possivelmente estes proponentes, na realidade, possuem um pequeno tempo de conta corrente, e acabam propositalmente omitindo esta informação, visando assim aumentar as chances de aprovação do seu financiamento. Muitas vezes esta

omissão é aconselhada pelo lojista que está lhe vendendo a garantia e que conhece o processo de aprovação.

Tabela 5.1.1: Análise bivariada de quantidade de meses de conta corrente

Categorias	Bons	Maus	Total	% rep	% maus	RR
Até 35 meses	15.672	2.783	18.455	22%	15%	0,54
Entre 35 e 107 meses	32.311	2.725	35.036	41%	8%	1,14
Mais que 107 meses	23.557	1.066	24.623	29%	4%	2,13
Missing	5.971	902	6.873	8%	13%	0,64
Total	77.511	7.476	84.987	100%	9%	1,00

As tabelas contendo a análise bivariada das demais variáveis explicativas podem ser encontradas no Apêndice A.

Por fim, cada categoria das variáveis explicativas foi transformada em uma variável indicadora (*dummy*), assumindo 1 caso o indivíduo pertença a categoria em questão e 0 caso contrário. Este artifício é utilizado a fim de incorporar possíveis efeitos não lineares das variáveis explicativas nos modelos. Para evitar multicolinearidade é necessário excluir uma das categorias de cada variável, dado que esta categoria será combinação linear das demais. Como critério de exclusão, optou-se pela última categoria de cada variável.

5.2 PROCESSO DE SELEÇÃO DAS VARIÁVEIS

O processo de seleção de variáveis de cada modelo iniciou-se com o método *stepwise*. Trata-se de um processo sequencial em que a cada novo passo uma variável explicativa é incluída ou retirada. Neste trabalho foi utilizado o método *forward stepwise*, no qual o ponto de partida não contém nenhuma variável explicativa, e elas são adicionadas etapa a etapa, tendo como critério de parada o momento em que nenhuma nova variável acrescenta informação ao modelo, e todas as contidas nele são significantes. Previamente são definidos os níveis de significância, que neste trabalho foram de 0,05 de entrada e 0,1 de saída.

Seguindo um procedimento comum no mercado, após a seleção inicial de variáveis realizada pelo método *stepwise*, foram feitos alguns ajustes a fim de garantir modelos mais interpretáveis e consistentes. Como o principal objetivo destes modelos é comparar as formas de tratamento de dados faltantes, procurou-se realizar o menor número de ajustes possíveis. Tais ajustes correspondem à eliminação ou reagrupamento de variáveis com indícios de multicolinearidade, identificadas através da comparação dos parâmetros do modelo com o comportamento observado das variáveis explicativas em suas respectivas análises bivariadas. Portanto, permaneceram no modelo apenas variáveis para as quais a ordenação de suas categorias, segundo seus betas, é semelhante à obtida quando a ordenamos pelo Risco Relativo.

5.3 AJUSTE DOS MODELOS

A partir da base de dados descrita no Capítulo 2 foram criados sete modelos de *Credit Score*, cada um deles com uma diferente técnica de tratamento de dados faltantes. A construção dos sete modelos seguiu as etapas descritas nas Subseções 5.1 e 5.2, ou seja, análise descritiva univariada, categorizações das covariáveis, análise descritiva bivariada e seleção das covariáveis com auxílio do método *stepwise*.

O Apêndice B contém as estimativas dos parâmetros dos modelos desenvolvidos, juntamente com seus respectivos erros padrões e valores p . Os Modelos 1 a 4 são descritos nas Tabelas B1, B2, B3 e B4 respectivamente, e o Modelo 7 pela Tabela B6. Os Modelos 5 e 6 se baseiam em um modelo construído utilizando apenas os dados completos, descrito na Tabela B5, para posteriormente contemplarem os dados faltantes através das técnicas de imputação múltipla e análise discriminante, respectivamente.

Após os ajustes dos modelos é possível calcular a probabilidade de cada um dos indivíduos da base se tornar inadimplente segundo cada modelo. Para isso utiliza-se a função logística. Por exemplo, um indivíduo classificado nas categorias de referência em todas as variáveis explicativas, terá sua probabilidade de

inadimplência totalmente vinculada ao valor da constante do modelo em questão. Para o Modelo 1, por exemplo, temos que:

$$P(mau) = \frac{e^{-3,02}}{[1+e^{-3,02}]} = 0,046.$$

5.4 COMPARAÇÃO DOS RESULTADOS

As subseções anteriores se destinaram a detalhar o processo de desenvolvimento dos modelos, através do qual foi obtido o melhor modelo para cada uma das metodologias de tratamento de dados faltantes descritas no Capítulo 3.

Já esta seção tem o objetivo de comparar estes modelos segundo seu desempenho preditivo. Para isso, foram utilizados os índices apresentados no Capítulo 4. Portanto, foram calculados os índices: KS, Gini, ROC e curva de aprovação. No Apêndice C são apresentados os gráficos de KS, curva ROC e curva de aprovação dos sete modelos desenvolvidos, todos eles obtidos sobre a base de construção, sendo que as Figuras C1, C2 e C3 correspondem aos resultados do Modelo 1; C4, C5 e C6 ao Modelo 2; C7, C8 e C9 ao Modelo 3; C10, C11 e C12 ao Modelo 4; C13, C14 e C15 ao Modelo 5; C16, C17 e C18 ao Modelo 6; C19, C20 e C21 ao Modelo 7.

A Tabela 5.4.1 apresenta os KS de cada modelo tanto na base de construção quanto na base de validação. Já a Tabela 5.4.2 contém os índices de ROC e Gini também para estas duas bases. Por fim, a Figura 5.4.1 apresenta a curva de aprovação dos modelos.

Através dos indicadores analisados é possível ver que todos os modelos apresentaram um comportamento estável, apresentando índices na base de validação semelhantes aos da base de construção.

Observa-se que os Modelos 1 (*Missing* como uma categoria) e 7 (Agrupamento conceitual dos *Missings*) foram os que apresentaram melhor desempenho, tendo seus ajustes resultados em indicadores idênticos nas bases analisadas. Em seguida aparece o Modelo 6 (Estimação de probabilidades). Já o

Modelo 2 (Descarte de variável com *missings*) foi o que apresentou pior desempenho, o que já era esperado, uma vez que uma importante variável preditiva não foi considerada no modelo.

A diferença entre o desempenho dos melhores modelos e o pior, na base de construção, é de 2,1 pontos em relação ao KS e de 3,8 em relação ao índice Gini. Apesar de aparentemente não serem diferenças tão significativas, ao analisarmos indicadores mais interpretáveis, observamos diferenças mais relevantes. Por exemplo, ao fixarmos a taxa de inadimplência em 40% vemos que com os Modelos 1 e 7 conseguimos uma taxa de aprovação aproximadamente 5 pontos percentuais maior do que a obtida pelo Modelo 2.

Tabela 5.4.1: KS dos modelos

Modelo	Base Construção	Base Validação
M1 - <i>Missing</i> como uma categoria	32,0	32,7
M2 - Descarte da variável com <i>missings</i>	29,9	29,4
M3 - Imputação via regressão	31,2	32,3
M4 - Imputação via algoritmo EM	31,5	32,4
M5 - Imputação múltipla	31,1	32,2
M6 - Estimção de probabilidades	31,7	32,5
M7 - Agrupamento conceitual dos <i>missings</i>	32,0	32,7

Tabela 5.4.2: ROC e Gini dos modelos

Modelo	ROC Construção	ROC Validação	Gini Construção	Gini Validação
M1 - <i>Missing</i> como uma categoria	0,716	0,719	0,432	0,438
M2 - Descarte da variável com <i>missings</i>	0,697	0,698	0,394	0,396
M3 - Imputação via regressão	0,712	0,715	0,424	0,430
M4 - Imputação via algoritmo EM	0,713	0,716	0,426	0,432
M5 - Imputação múltipla	0,710	0,714	0,420	0,428
M6 - Estimção de probabilidades	0,714	0,717	0,428	0,434
M7 - Agrupamento conceitual dos <i>missings</i>	0,716	0,719	0,432	0,438

6 CONCLUSÃO

Muitas instituições financeiras já utilizam modelos preditivos em suas operações a fim de propiciar decisões mais objetivas e embasadas. Porém ao se depararem com variáveis que possuem dados faltantes não sabem ao certo qual a melhor forma de tratá-las. Sendo assim, o objetivo deste trabalho foi apresentar algumas técnicas de tratamento de variáveis que possuem *missings*, incentivando a discussão do tema e a aplicação de outras técnicas.

Ao todo foram descritas sete formas de tratamento de dados faltantes em modelos preditivos, dentre as quais estão o agrupamento conceitual dos *missings* e as mais utilizadas pelo mercado que consistem em descartar as variáveis que possuem dados faltantes ou o tratamento como uma categoria à parte. Além destas, técnicas já difundidas na literatura para tratamento de dados faltantes, como imputação simples (via regressão linear ou algoritmo EM) e imputação múltipla, foram adaptadas para o contexto de desenvolvimento de modelos preditivos.

Para avaliação destas técnicas utilizou-se uma base cedida por uma instituição financeira, sobre a qual foi construído um modelo de *Credit Score* e que contém uma variável correlacionada com a resposta que possui um percentual significativo de dados faltantes. Sobre esta base, as sete técnicas foram aplicadas e seus resultados foram comparados segundo indicadores de desempenho comumente utilizados pelo mercado como: KS, ROC, Gini e curva de aprovação.

Como era esperado, o modelo desenvolvido descartando a variável com dados faltantes foi o que mostrou o pior desempenho preditivo. Dado que tal modelo não contempla as informações fornecidas pela variável com dados faltantes, ele só possuiria um desempenho semelhante aos demais modelos caso houvesse alguma outra covariável (ou conjunto de covariáveis) correlacionada com a variável descartada, e que substituísse o incremento preditivo gerado pela variável descartada (o que não aconteceu na aplicação deste estudo). Portanto, a opção de desconsiderar a variável com dados faltantes, apesar de ser possivelmente a mais simples, pode acarretar em perdas significativas.

Já a técnica que trata os dados faltantes como uma categoria e a que os agrupa na categoria conceitualmente mais próxima foram as que apresentaram

melhor desempenho, resultando em modelos que chegam a aprovar aproximadamente 5 pontos percentuais a mais do que o modelo que desconsidera a variável com dados faltantes. Estas técnicas podem ser facilmente utilizadas no desenvolvimento de modelos preditivos e geram modelos de fácil aplicação.

Em relação às técnicas de tratamento de dados faltantes difundidas na literatura, e que apresentam ganhos significativos em outros contextos, todas apresentaram resultados intermediários. Dentre elas, a que gerou um modelo com maior poder preditivo foi a que estima as probabilidades de cada registro pertencer a cada uma das categorias da variável com *missings* (lembrando que tal variável, seguindo prática de mercado, é categorizada durante o desenvolvimento do modelo), e utiliza estas probabilidades para estimar um *score* final para cada uma das unidades amostrais da base. Esta metodologia apresentou resultados muito próximos aos dos modelos de melhor resultados neste estudo. Entretanto, este tipo de metodologia gera uma aplicação de modelo relativamente mais complexa, uma vez que para se obter o *score* final é necessária a aplicação de pelo menos mais uma técnica (análise discriminante, regressões, imputação múltipla, entre outras).

Tais resultados indicam que os *missings* se aproximam mais de um comportamento MNAR (*Missings Not at Random*), uma vez que o modelo que agrupa conceitualmente a variável com dados faltantes apresentou um melhor desempenho, indicando uma não aleatoriedade dos *missings*. Possivelmente, em cenários com dados faltantes MCAR ou MAR, as metodologias que utilizam técnicas como as de imputação simples ou imputação múltipla apresentassem um melhor desempenho.

Por fim, é importante reforçar que apenas algumas técnicas foram descritas e testadas neste estudo e que existem diversas outras metodologias de tratamento de dados faltantes que podem ser utilizadas na construção de modelos preditivos e que poderiam gerar resultados superiores aos obtidos através das técnicas testadas aqui.

Apêndice A
Tabelas de Análise Descritiva

Tabela A1: Análise bivariada da Var1

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR1 < 510	18.451	1.563	20.014	24%	7,8%	1,14
510 <= VAR1 < 2915	51.510	5.335	56.845	67%	9,4%	0,93
2915 <= VAR1	7.550	578	8.128	10%	7,1%	1,26
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A2: Análise bivariada da Var2

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR2 < 5290	11.732	893	12.625	15%	7,1%	1,27
5290 <= VAR2 < 6900	30.094	2.735	32.829	39%	8,3%	1,06
6900 <= VAR2 < 8450	22.897	2.363	25.260	30%	9,4%	0,93
8450 <= VAR2 < 21000	12.404	1.467	13.871	16%	10,6%	0,82
21000 <= VAR2	384	18	402	0%	4,5%	2,06
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A3: Análise bivariada da Var3

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR3 < 27	4.441	94	4.535	5%	2,1%	4,56
27 <= VAR3 < 40	13.744	696	14.440	17%	4,8%	1,90
40 <= VAR3	59.326	6.686	66.012	78%	10,1%	0,86
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A4: Análise bivariada da Var4

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR4 < 168	5.696	335	6.031	7%	5,6%	1,64
168 <= VAR4 < 204	13.864	1.044	14.908	18%	7,0%	1,28
204 <= VAR4 < 233	16.774	1.638	18.412	22%	8,9%	0,99
233 <= VAR4 < 383	38.111	4.071	42.182	50%	9,7%	0,90
383 <= VAR4 < 534	2.495	354	2.849	3%	12,4%	0,68
534 <= VAR4	571	34	605	1%	5,6%	1,62
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A5: Análise bivariada da Var5

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR5 < 45	65.127	6.148	71.275	84%	8,6%	1,02
45 <= VAR5	12.384	1.328	13.712	16%	9,7%	0,90
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A6: Análise bivariada da Var6

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR6 < 2	74.427	7.305	81.732	96%	8,9%	0,98
2 <= VAR6	3.084	171	3.255	4%	5,3%	1,74
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A7: Análise bivariada da Var7

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR7 < 2008	3.131	164	3.295	4%	5,0%	1,84
2008 <= VAR7	74.380	7.312	81.692	96%	9,0%	0,98
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A8: Análise bivariada da Var8

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR8 < 125	7.960	598	8.558	10%	7,0%	1,29
125 <= VAR8	67.526	6.715	74.241	87%	9,0%	0,97
Total	75.486	7.313	82.799	97%	8,8%	1,00

Tabela A9: Análise bivariada da Var9

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR9 < 5,122	8.347	615	8.962	11%	6,9%	1,31
5,122 <= VAR9 < 6,763	25.509	2.765	28.274	33%	9,8%	0,89
6,763 <= VAR9 < 8,173	25.405	2.256	27.661	33%	8,2%	1,09
8,173 <= VAR9	18.250	1.840	20.090	24%	9,2%	0,96
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A10: Análise bivariada da Var10

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR10 < 377	2.590	694	3.284	4%	21,1%	0,36
377 <= VAR10 < 498	18.793	3.005	21.798	26%	13,8%	0,60
498 <= VAR10 < 575	13.469	1.570	15.039	18%	10,4%	0,83
575 <= VAR10 < 635	11.832	961	12.793	15%	7,5%	1,19
635 <= VAR10 < 711	12.943	747	13.690	16%	5,5%	1,67
711 <= VAR10 < 776	8.488	327	8.815	10%	3,7%	2,50
776 <= VAR10 < 856	6.134	145	6.279	7%	2,3%	4,08
856 <= VAR10	3.262	27	3.289	4%	0,8%	11,65
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A11: Análise bivariada da Var11

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR11 < 700	6.384	429	6.813	8%	6,3%	1,44
700 <= VAR11 < 848	13.601	1.357	14.958	18%	9,1%	0,97
848 <= VAR11 < 1,482	38.952	4.644	43.596	51%	10,7%	0,81
1,482 <= VAR11 < 1,675	4.219	350	4.569	5%	7,7%	1,16
1,675 <= VAR11	14.355	696	15.051	18%	4,6%	1,99
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A12: Análise bivariada da Var12

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR12 < 19	381	129	510	1%	25,3%	0,28
19 <= VAR12 < 22	6.197	1.247	7.444	9%	16,8%	0,48
22 <= VAR12 < 26	9.742	1.417	11.159	13%	12,7%	0,66
26 <= VAR12 < 30	10.445	1.223	11.668	14%	10,5%	0,82
30 <= VAR12 < 36	12.820	1.195	14.015	16%	8,5%	1,03
36 <= VAR12 < 43	12.411	925	13.336	16%	6,9%	1,29
43 <= VAR12 < 49	9.712	586	10.298	12%	5,7%	1,60
49 <= VAR12	15.803	754	16.557	19%	4,6%	2,02
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A13: Análise bivariada da Var13

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR13 < 2	76.838	7.448	84.286	99%	8,8%	1,00
2 <= VAR13	673	28	701	1%	4,0%	2,32
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A14: Análise bivariada da Var14

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR14 < 1	11.398	913	12.311	14%	7,4%	1,20
1 <= VAR14 < 25	49.362	5.476	54.838	65%	10,0%	0,87
25 <= VAR14 < 33	8.572	765	9.337	11%	8,2%	1,08
33 <= VAR14 < 42	4.244	230	4.474	5%	5,1%	1,78
42 <= VAR14	3.935	92	4.027	5%	2,3%	4,13
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A15: Análise bivariada da Var15

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR15 < 12	2.480	270	2.750	3%	9,8%	0,89
12 <= VAR15 < 32	30.569	4.564	35.133	41%	13,0%	0,65
32 <= VAR15 < 46	8.168	793	8.961	11%	8,8%	0,99
46 <= VAR15 < 129	22.123	1.295	23.418	28%	5,5%	1,65
129 <= VAR15	14.171	554	14.725	17%	3,8%	2,47
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A16: Análise bivariada da Var16

Categorias	Bons	Maus	Total	% rep	% maus	RR
<=35	15.672	2.783	18.455	22%	15,1%	0,54
35 -- 107	32.311	2.725	35.036	41%	7,8%	1,14
> 107	23.557	1.066	24.623	29%	4,3%	2,13
Missing	5.971	902	6.873	8%	13,1%	0,64
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A17: Análise bivariada da Var17

Categorias	Bons	Maus	Total	% rep	% maus	RR
Cat1	71.575	7.218	78.793	93%	9,2%	0,96
Cat2	5.936	258	6.194	7%	4,2%	2,22
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A18: Análise bivariada da Var18

Categorias	Bons	Maus	Total	% rep	% maus	RR
Cat1	25.687	1.938	27.625	33%	7,0%	1,28
Cat2	51.824	5.538	57.362	67%	9,7%	0,90
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A19: Análise bivariada da Var19

Categorias	Bons	Maus	Total	% rep	% maus	RR
Cat1	18.585	2.306	20.891	25%	11,0%	0,78
Cat2	21.577	1.372	22.949	27%	6,0%	1,52
Cat3	37.349	3.798	41.147	48%	9,2%	0,95
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A20: Análise bivariada da Var16B

Categorias	Bons	Maus	Total	% rep	% maus	RR
0 <= VAR16B < 1	2.913	377	3.290	4%	11,5%	0,75
1 <= VAR16B < 36	14.603	2.768	17.371	20%	15,9%	0,51
36 <= VAR16B < 108	35.052	3.110	38.162	45%	8,1%	1,09
108 <= VAR16B < 180	16.205	898	17.103	20%	5,3%	1,74
180 <= VAR16B	8.738	323	9.061	11%	3,6%	2,61
Total	77.511	7.476	84.987	100%	8,8%	1,00

Tabela A21: Análise bivariada da Var16 - CatC

Categorias	Bons	Maus	Total	% rep	% maus	RR
VAR16 < 1	2.260	245	2.505	3%	9,8%	0,89
1 <= VAR16 < 36	14.457	2.790	17.247	20%	16,2%	0,50
36 <= VAR16 < 60	13.047	1.401	14.448	17%	9,7%	0,90
60 <= VAR16 < 108	23.429	1.923	25.352	30%	7,6%	1,18
108 <= VAR16 < 1788	15.857	826	16.683	20%	5,0%	1,85
178 <= VAR16	8.461	291	8.752	10%	3,3%	2,80
Total	77.511	7.476	84.987	100%	8,8%	1,00

Apêndice B

Modelos Ajustados

Tabela B1: Estimativas, erros padrões e valores p dos parâmetros do Modelo 1 – *Missings* como uma categoria

Parâmetro	B	Erro Padrão	Valor p
VAR3_cat(1)	-1,377	0,111	0,000
VAR3_cat(2)	-0,768	0,043	0,000
VAR4_cat(1)	-0,440	0,062	0,000
VAR4_cat(2)	-0,383	0,037	0,000
VAR4_cat(3)	-0,205	0,031	0,000
VAR12_cat(1)	0,672	0,055	0,000
VAR12_cat(2)	0,519	0,053	0,000
VAR12_cat(3)	0,421	0,053	0,000
VAR12_cat(4)	0,304	0,051	0,000
VAR12_cat(5)	0,203	0,052	0,000
VAR12_cat(6)	0,083	0,058	0,148
VAR19_cat(1)	0,228	0,029	0,000
VAR19_cat(2)	-0,396	0,034	0,000
VAR14_cat(1)	0,652	0,119	0,000
VAR14_cat(2)	0,816	0,114	0,000
VAR14_cat(3)	0,688	0,117	0,000
VAR14_cat(4)	0,354	0,129	0,006
VAR16_cat(1)	0,017	0,042	0,682
VAR16_cat(2)	-0,605	0,042	0,000
VAR16_cat(3)	-0,952	0,050	0,000
VAR15_cat(1)	0,557	0,054	0,000
VAR15_cat(2)	0,381	0,063	0,000
VAR15_cat(3)	0,114	0,055	0,040
Intercepto	-3,022	0,132	0,000

Tabela B2: Estimativas, erros padrões e valores p dos parâmetros do Modelo 2 – Descarte da variável com *missings*

Parâmetro	B	Erro Padrão	Valor p
VAR3_cat(1)	-1,362	0,110	0,000
VAR3_cat(2)	-0,757	0,043	0,000
VAR4_cat(1)	-0,357	0,061	0,000
VAR4_cat(2)	-0,332	0,037	0,000
VAR4_cat(3)	-0,179	0,031	0,000
VAR12_cat(1)	0,924	0,053	0,000
VAR12_cat(2)	0,657	0,052	0,000
VAR12_cat(3)	0,527	0,052	0,000
VAR12_cat(4)	0,373	0,051	0,000
VAR12_cat(5)	0,249	0,052	0,000
VAR12_cat(6)	0,120	0,057	0,037
VAR19_cat(1)	0,153	0,029	0,000
VAR19_cat(2)	-0,375	0,034	0,000
VAR14_cat(1)	0,629	0,118	0,000
VAR14_cat(2)	0,799	0,113	0,000
VAR14_cat(3)	0,691	0,117	0,000
VAR14_cat(4)	0,369	0,129	0,004
VAR15_cat(1)	0,833	0,052	0,000
VAR15_cat(2)	0,557	0,061	0,000
VAR15_cat(3)	0,230	0,054	0,000
Intercepto	-3,743	0,125	0,000

Tabela B3: Estimativas, erros padrões e valores p dos parâmetros do Modelo 3 – Imputação via regressão

Parâmetro	B	Erro Padrão	Valor p
VAR3_cat(1)	-1,373	0,111	0,000
VAR3_cat(2)	-0,770	0,043	0,000
VAR4_cat(1)	-0,408	0,062	0,000
VAR4_cat(2)	-0,367	0,037	0,000
VAR4_cat(3)	-0,195	0,031	0,000
VAR12_cat(1)	0,653	0,055	0,000
VAR12_cat(2)	0,491	0,053	0,000
VAR12_cat(3)	0,391	0,053	0,000
VAR12_cat(4)	0,274	0,052	0,000
VAR12_cat(5)	0,182	0,053	0,001
VAR12_cat(6)	0,072	0,058	0,211
VAR19_cat(1)	0,214	0,029	0,000
VAR19_cat(2)	-0,387	0,034	0,000
VAR14_cat(1)	0,657	0,119	0,000
VAR14_cat(2)	0,817	0,114	0,000
VAR14_cat(3)	0,689	0,117	0,000
VAR14_cat(4)	0,357	0,129	0,006
VAR16_catB(1)	0,899	0,083	0,000
VAR16_catB(2)	1,006	0,068	0,000
VAR16_catB(3)	0,426	0,066	0,000
VAR16_catB(4)	0,133	0,069	0,055
VAR15_cat(1)	0,557	0,055	0,000
VAR15_cat(2)	0,376	0,064	0,000
VAR15_cat(3)	0,105	0,056	0,061
Intercepto	-3,976	0,132	0,000

Tabela B4: Estimativas, erros padrões e valores p dos parâmetros do Modelo 4 – Imputação via algoritmo EM

Parâmetro	B	Erro Padrão	Valor p
VAR3_cat(1)	-1,378	0,111	0,000
VAR3_cat(2)	-0,773	0,043	0,000
VAR4_cat(1)	-0,421	0,062	0,000
VAR4_cat(2)	-0,375	0,037	0,000
VAR4_cat(3)	-0,199	0,031	0,000
VAR12_cat(1)	0,600	0,055	0,000
VAR12_cat(2)	0,450	0,053	0,000
VAR12_cat(3)	0,367	0,053	0,000
VAR12_cat(4)	0,259	0,052	0,000
VAR12_cat(5)	0,171	0,053	0,001
VAR12_cat(6)	0,066	0,058	0,249
VAR19_cat(1)	0,217	0,029	0,000
VAR19_cat(2)	-0,385	0,034	0,000
VAR14_cat(1)	0,650	0,119	0,000
VAR14_cat(2)	0,805	0,114	0,000
VAR14_cat(3)	0,679	0,117	0,000
VAR14_cat(4)	0,355	0,129	0,006
VAR16_catC(1)	0,943	0,093	0,000
VAR16_catC(2)	1,118	0,071	0,000
VAR16_catC(3)	0,631	0,073	0,000
VAR16_catC(4)	0,505	0,070	0,000
VAR16_catC(5)	0,172	0,072	0,018
VAR15_cat(1)	0,518	0,055	0,000
VAR15_cat(2)	0,328	0,064	0,000
VAR15_cat(3)	0,076	0,056	0,176
Intercepto	-4,006	0,133	0,000

Tabela B5: Estimativas, erros padrões e valores p dos parâmetros do modelo construído sem os *missings*

Parâmetro	B	Erro Padrão	Valor p
VAR3_cat(1)	-1,357	0,116	0,000
VAR3_cat(2)	-0,774	0,045	0,000
VAR4_cat(1)	-0,468	0,068	0,000
VAR4_cat(2)	-0,371	0,040	0,000
VAR4_cat(3)	-0,228	0,034	0,000
VAR12_cat(1)	0,666	0,059	0,000
VAR12_cat(2)	0,510	0,056	0,000
VAR12_cat(3)	0,413	0,056	0,000
VAR12_cat(4)	0,298	0,055	0,000
VAR12_cat(5)	0,208	0,056	0,000
VAR12_cat(6)	0,105	0,061	0,085
VAR19_cat(1)	0,244	0,030	0,000
VAR19_cat(2)	-0,395	0,036	0,000
VAR14_cat(1)	0,647	0,126	0,000
VAR14_cat(2)	0,831	0,121	0,000
VAR14_cat(3)	0,698	0,125	0,000
VAR14_cat(4)	0,339	0,138	0,014
VAR16_cat(1)	0,979	0,042	0,000
VAR16_cat(2)	0,351	0,040	0,000
VAR15_cat(1)	0,556	0,057	0,000
VAR15_cat(2)	0,405	0,066	0,000
VAR15_cat(3)	0,122	0,058	0,036
Intercepto	-3,992	0,134	0,000

Tabela B6: Estimativas, erros padrões e valores p dos parâmetros do Modelo 7 – Agrupamento conceitual dos *missings*

Parâmetro	B	Erro Padrão	Valor p
VAR3_cat			0,000
VAR3_cat(1)	-1,376	0,111	0,000
VAR3_cat(2)	-0,768	0,043	0,000
VAR4_cat			0,000
VAR4_cat(1)	-0,441	0,062	0,000
VAR4_cat(2)	-0,383	0,037	0,000
VAR4_cat(3)	-0,206	0,031	0,000
VAR12_cat			0,000
VAR12_cat(1)	0,673	0,054	0,000
VAR12_cat(2)	0,519	0,053	0,000
VAR12_cat(3)	0,421	0,053	0,000
VAR12_cat(4)	0,304	0,051	0,000
VAR12_cat(5)	0,203	0,052	0,000
VAR12_cat(6)	0,084	0,058	0,147
VAR19_cat			0,000
VAR19_cat(1)	0,228	0,029	0,000
VAR19_cat(2)	-0,397	0,034	0,000
VAR14_cat			0,000
VAR14_cat(1)	0,651	0,119	0,000
VAR14_cat(2)	0,816	0,114	0,000
VAR14_cat(3)	0,687	0,117	0,000
VAR14_cat(4)	0,354	0,129	0,006
VAR16_cat			0,000
VAR16_cat(1)	0,964	0,040	0,000
VAR16_cat(2)	0,347	0,040	0,000
VAR15_cat			0,000
VAR15_cat(1)	0,557	0,054	0,000
VAR15_cat(2)	0,380	0,063	0,000
VAR15_cat(3)	0,113	0,055	0,040
Intercepto	-3,974	0,126	0,000

Apêndice C

Figuras

Figura C1: Curva de KS do Modelo 1

Figura C2: Curva ROC do Modelo 1

Figura C3: Curva de aprovação do Modelo 1

Figura C4: Curva de KS do Modelo 2

Figura C5: Curva ROC do Modelo 2

Figura C6: Curva de aprovação do Modelo 2

Figura C7: Curva de KS do Modelo 3

Figura C8: Curva ROC do Modelo 3

Figura C9: Curva de aprovação do Modelo 3

Figura C10: Curva de KS do Modelo 4

Figura C11: Curva ROC do Modelo 4

Figura C12: Curva de aprovação do Modelo 4

Figura C13: Curva de KS do Modelo 5

Figura C14: Curva ROC do Modelo 5

Figura C15: Curva de aprovação do Modelo 5

Figura C16: Curva de KS do Modelo 6

Figura C17: Curva ROC do Modelo 6

Figura C18: Curva de aprovação do Modelo 6

Figura C19: Curva de KS do Modelo 7

Figura C20: Curva ROC do Modelo 7

Figura C21: Curva de aprovação do Modelo 7

Referências Bibliográficas

AGRESTI, A. **Categorical Data Analysis**. New York: Wiley, 1990.

ALVES, M. C. **Estratégias para o desenvolvimento de modelos de Credit Score com inferência de rejeitados**. São Paulo: USP, 2008.

ARAÚJO, E. A.; CARMONA, C. U. M.; DEBOÇÃ, L. P. Risco de crédito: construção de modelos de Credit Scoring com abordagem de regressão logística para a análise da inadimplência de uma instituição de microcrédito. **Seminários em Administração Fea-Usp**, n. 9, São Paulo, 2006.

CONOVER, W. J. **Practical Nonparametrical Statistics**. 3 ed. New York: John Wiley and Sons, 1999.

COX, D. R. **The analysis of binary data**. London: Methuen, 1970, 142 p.

COX, D. R. Partial likelihood. **Biometrika**, n. 62, p. 269–276, 1975.

COX, D. R.; SNELL, E. J. **Analysis of Binary Data**. 2 ed. London: Chapman and Hall, 1989.

DEMPSTER, A.; LAIRD, N.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society**. series B, n. 39, v. 1, p. 1–18, 1977.

FERREIRA, J. **Modelos de previsão de perdas para crédito massificado**. São Paulo, 2008.

FERREIRA, N.M. **Presença de dados missing em modelos de regressão logística**. São Carlos, 2009.

GILKS, W.R.; RICHARDSON, S.; SPIEGELHALTER, D.J.; **Markov chain Monte Carlo in practice**. London: Chapman and Hall, 1996.

GRAHAM, J. W.; CUMSILLE, P. E. ELEK-FISK, E. Methods for handling missing data. In: SCHINKA, J. A.; VELICER, W. F. (Eds.) **Research Methods in Psychology**, v. 2, New York: John Wiley and Song, p. 87–114, 2003.

GRAHAM, J. W.; HOFER, S. M.; DONALDSON, S. I.; MACKINNON, D. P.; SHAFER, J. L. **Analysis with missing data in prevention research. Modeling longitudinal and multiple-group data**: Practical issues, applied approaches and specifics examples. Washington DC: American Psychological Association, p. 201–218, 1997.

KAAS, G. V. **An exploratory technique for investigation large quantities for categorical data**, n. 29, v. 2, p. 119–127, 1980.

LITTLE, R. J. A. Modeling the drop-out mechanism in repeated-measures studies. **Journal of the American Statistical Association**, n. 90, p. 1112 – 1121, 1995.

LITTLE, R. J. A.; RUBIN, D. B. **Statistical Analysis with missing data**. New York: John Wiley and Song, 1987.

MONTGOMERY, D. C.; PECK, E.A. E.; VINNING, G. G. **Introduction To Linear Regression Analysis**. 4 ed., Wiley, 2001.

NUNES, L. N. **Métodos de imputação de dados aplicados na área da saúde**. Porto Alegre, 2007.

PAULA, G. A. **Modelos de regressão com apoio computacional**. São Paulo: Instituto de Matemática e Estatística, Universidade de São Paulo, 2004.

POLETO, F. Z. **Análise de dados categorizados com omissão**. São Paulo, 2006.

RUBIN, D. B. **Multiple Imputation for Nonresponse in Surveys**. New York: John Wiley and Sons, 1987.

SCHAFER, J. L.; GRAHAM, J. W. **Missing data**: our view of the state of the art. *Psychological Methods*, n. 7, p. 147–177, 2002.

SELAU, L. P. R. **Construção de modelos de previsão de risco de crédito**. Porto Alegre, 2008.

VELOSO, A. D. S. **Fidelidade em telefonia celular: proposição e validação de um índice para previsão da fidelidade de um cliente**. Belo Horizonte, 2006.

WAYMAN, J.C. Multiple imputation for missing data: what is it and how can I use it?
Annual Meeting Of The American Educational Research, Chicago, 2003.