Henrique de Oliveira Teixeira

Avaliação de metodologias paramétricas e não paramétricas de credit score

Campinas

06/12/2012

Henrique de Oliveira Teixeira

Avaliação de metodologias paramétricas e não paramétricas de credit score

Orientadora

Prof[a] Dra. Ivette Luna

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE ECONOMIA

Campinas

06/12/2012

Monografia de Projeto Final de Graduação sob o título *Avaliação de metodologias paramétricas e não paramétricas de credit score* , apresentada por Henrique de Oliveira Teixeira e aprovada em 06/12/2012, em Campinas, Estado do São Paulo, pela banca examinadora constituída pelos professores:

Profª Dra. Ivette Luna
Orientadora

Profª Dr. Alexandre Gori Maia
Professor convidado

Resumo

O objetivo desse trabalho é analisar e comparar três diferentes modelos de classificação de padrões com foco na concessão de crédito. Esses modelos, conhecidos como modelos de *credit score*, se tornaram ferramentas indispensáveis na gestão do risco nas instituições financeiras, principalmente após o forte crescimento do crédito e a difusão de meios eletrônicos de pagamento nas ultimas décadas, que levaram a uma vertiginosa aceleração do número de transações e do volume de recursos transacionado.

Os modelos considerados são o modelo paramétrico de Regressão Logística e os modelo de Árvore de Decisão e Redes Neurais, ambos modelos não paramétricos da área de *machine learning*.

Os modelos serão validados sobre um conjuntos de dados dentro e fora da amostra de treinamento, utilizando como métricas de desempenho os Índice de Kolmogorov-Smirnov e de Gini.

Abstract

The aim of this study is to analyze and compare three different models of classification standards with focus on credit. These models, known as models of textit credit score, have become indispensable tools in risk management of financial institutions, especially after the strong credit growth and diffusion of electronic payments in last years, leading to a vertiginous acceleration the number of transactions and the volume of funds traded.

The models considered are the parametric model of Logistic Regression Model and Decision Tree and Neural Networks, both non-parametric models of the area *machine learning*.

The models will be validated on a datasets within and outside the training sample, using performance metrics such as the Kolmogorov-Smirnov Index and Gini Index.

Agradecimentos

Em primeiro lugar, agradeço aos meus pais, Francisco e Evani pelo apoio incondicional e pela imensurável dedicação com que eu pude e sempre poderei contar; a meus irmãos, Daniel, Lais e Isis, por todos os bons momentos; Paula, a quem devo grande parte das realizações nos últimos anos, aos jovens e brilhantes economistas, Dirceu, Eduardo, Felipe, Daniel, Leonardo, Renata, Paulo e Thais, por todas as conversas e reflexões não só acadêmicas; a minha orientadora, Ivette Luna, que desde o primeiro momento demonstrou interesse pelo tema e aceitou o complexo e árduo desafio de me orientar.

“O senhor poderia me dizer, por favor, qual o caminho que devo tomar para sair daqui?”
“Isso depende muito de para onde você quer ir”, respondeu o Gato.
“Não me importo muito para onde...”, retrucou Alice.
“Então não importa o caminho que você escolha!” disse o Gato.
Lewis Carroll - Alice no País das Maravilhas

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A expansão da atividade economia dos anos pré crise *sub-prime* e do crescimento do volume de crédito concedido além da própria crise, trouxeram à tona a necessidade de uma análise mais profunda do risco e do grau de exposição que as instituições financeiras fornecedoras de crédito estão sujeitas. As ferramentas de *credit score* se tornaram instrumentos fundamentais na análise e gestão do risco de crédito. Estes modelos automatizam o processo de classificação de novos e potenciais tomadores de crédito, além de viabilizar um caráter mais científico e menos subjetivo à análise de crédito. A partir de dados de operações passadas, como informações pessoais e macroeconômicas, busca-se avaliar futuros demandantes, atribuindo-lhes um *score* relativo ao seu risco potencial.

Uma avaliação incorreta do risco de crédito em operações financeiras pode trazer dois inconvenientes principais às instituições financeiras. Em primeiro lugar, classificar um potencial cliente como bom quando na verdade ele é mau [1] traria prejuízos com a perda resultante da inadimplência. Por outro lado, avaliar um cliente bom como mau privaria a instituição de receitas e lucros decorrentes da operação de crédito que não será realizada. Como apontado por Lahsasna (2010), muitas instituições financeiras norte americanas sofreram pesadas perdas decorrente de uma equivocada avaliação dos créditos imobiliários, cujos riscos não foram avaliados de forma correta, desembocando em uma grave crise sistêmica.

[1] Ao longo do trabalho, serão considerados como bons os indivíduos adimplentes, e maus os inadimplentes.

Nenhum modelo está livre de equívocos, o erro sempre está presente. Por isso, avanços nas técnicas de classificação podem trazer ganhos consideráveis para as instituições financeiras que têm no crédito uma importante fonte de lucros. Nesse sentido, este trabalho busca o estudo comparativo de modelos de *credit score*, com o objetivo de avaliar a capacidade preditiva de diferentes técnicas. O trabalho está dividido da seguinte forma:

O capítulo 2 posiciona este trabalho definindo os principais conceitos associados ao risco de crédito. No capítulo 3 é feita a descrição dos dados utilizados no desenvolvimento do estudo, assim como o processo de *stepwise* utilizado na seleção das variáveis explicativas. O capítulo contêm uma descrição detalhada das variáveis selecionadas e que serão utilizadas na construção dos modelos.

O capítulo 4 descreve e demonstra o processo de construção dos três modelos utilizados para *credit score*, Regressão Logística, Árvore de Decisão e Rede Neural. Os indicadores empregados na avaliação dos modelos são descritos no capitulo 5, assim como a apresentação e discussão dos resultados na seção 5.

Por final, a conclusão proveniente do estudo é apresentada no capitulo 5.3, além de considerações com propostas para futuros estudos.

2 *Referencial Teórico*

2.1 Risco

Inerente aos fatos da vida, o risco esta presente em todos os momentos, em todas as situações, seja na esfera economia, profissional ou pessoal. Apesar do seu caráter onipresente, a definição de risco é abstrata. Uma das formas de defini-lo, com foco no contexto econômico é ligar o conceito de risco ao de incerteza, ambos inerentes da atividade econômica. Um dos primeiros autores a fazer essa ligação foi Knight (1921) que apesar de tratar ambos de uma forma distinta, porem intimamente ligados, definiu o risco, de forma simples, como a mensuração da incerteza:

> "*Uncertainty must be taken in a sense radically distinct from the familiar notion of Risk, from which, it never been properly separated... The essential fact is that "risk" means in some cases a quantity susceptible of measurement, while at other times it is something distinctly not of the character; and there are far-reaching and crucial diferences in the bearing of the phenomenon depending on which of the two is really present and operation. It will appear that a mensurable uncertainty, or "risk" proper, as we shall use the term, is so far different from an unmensurable one, that it is not in effect an uncertainty at all.* " (KNIGHT, 1921, p. 34)

Essa capacidade de mensurar a incerteza depende do ambiente econômico, do custo e da quantidade de informações disponíveis ou mesmo simplesmente da capacidade dos agentes de atribuir uma expectativa de quê determinados eventos venham ou não a ocorrer. Por isso uma das formas mais elementares de quantificar o risco é atribuir uma probabilidade à

sua ocorrência, a ocorrência dos eventos incertos, tomando como base fontes concretas ou abstratas, independentemente da posterior concretização ou não da expectativa previa.

Quando não é possível atribuir qualquer tipo de probabilidade (ou qualquer outra medida) ao evento, seja porque não há qualquer base, mesmo que frágil, para inferir sua trajetória ou por completo desconhecimento de suas fontes ou suas origens, tem-se um evento intrinsecamente incerto. Keynes (1937), define essa categoria de eventos da seguinte forma:

> "*The sense in which I am using the term [uncertain] is that in which the prospect of a European war is uncertain, or the price of cooper and the rate of interest twenty years hence, or the obsolescence of a new invention, or the position of private wealthowners in the social system in 1970. About these matters there is no scientific basis on which to form any calculable probability whatever. We simply do not know.* "(KEYNES, 1937, p. 214)

A diferença fundamental entre risco e incerteza, portanto, reside no fato do primeiro ser uma mensuração do segundo, seja por meio de informações sólidas, que trariam consigo uma maior confiança por reduzirem a incerteza, ou mesmo por uma base frágil de informações que reduziria a capacidade de previsão. Segundo Keynes (1937), os agentes veem o presente como uma melhor capacidade de predizer o futuro do que as situações muito pretéritas, e além disso, o senso comum é baseado em melhores informações, em detrimento da capacidade individual de compor um cenário, por isso é uma boa *proxy* para o que pode vir a ocorrer no futuro.

Ou seja, uma boa fonte de informações para a previsão de eventos futuros seria olhar para o presente ou mesmo para o passado recente e supor sua continuidade ou pequena oscilação. Alem disso, o senso comum dos agentes quando analisados em grupo pode ser considerada uma boa previsão para a redução da incerteza quanto a trajetória de eventos futuros, por ser considerada a média das opiniões individuais.

Assim, a definição de risco dada vai de encontro com a definição proposta por Hans-

son (2005), que faz a ligação da mensuração da incerteza com a teoria da probabilidade, ao afirmar que o risco é o valor esperado da ocorrência ou não de eventos incertos. Segundo o autor, na teoria da decisão a distinção fundamental é a de que na tomada de decisão sob risco assume-se que as probabilidades são conhecidas, ao contrario das decisões sob incerteza, onde não há qualquer conhecimento sobre as causas e os possíveis resultados de determinada decisão.

Apesar dessa definição de risco servir aos propósitos deste trabalho, podemos incluir a contribuição de Holton (2004), que afirma que para se configurar um evento de risco, não basta apenas atribuir uma probabilidade ao mesmo, mas também que esses eventos sejam relevantes em termos de utilidade para o individuo[1], no sentido estritamente econômico. O autor deixa claro esse conceito com o exemplo de uma pessoa que retira bolas de uma urna. Esse evento é incerto, mas apesar de ser factível a atribuição de uma probabilidade (dado determinado conjunto de informações previamente fornecidas), a retirada de bolas não vem acompanhada de qualquer beneficio ao indivíduo. Se, entretanto, atrelarmos um valor de uso as distintas bolas, essa atividade se converte em um evento de risco uma vez que há concomitantemente a probabilidade do evento e a utilidade atrelada aos distintos espaços amostrais, podendo ser realizada uma estimativa do valor esperado do evento.

Por fim ele nos dá o exemplo de um individuo que pula de um avião sem paraquedas, *"se há a certeza de que ele vai morrer, não há risco"*[2].Dado que a incerteza não existe, não no sentido da impossibilidade da atribuição de probabilidade ao evento, pelo contrario, podemos atrelar a probabilidade de 100% do evento ocorrer (a morte do individuo), este não é um evento incerto, e portanto, não se caracteriza como um evento de risco.

[1]Tomando o individuo como consumidor de bens e serviços, Krugman (2004) define utilidade como "a mensuração da satisfação que o individuo adquire ao consumir bens e serviços."

[2](HOLTON, 2004, p. 100)

Assim, o conceito que pode ser derivado dessa análise, é o do risco como a mensuração da incerteza em termos de probabilidade, devendo essa probabilidade ser maior do que 0 (não há a incerteza da ocorrência do evento, mas a certeza da sua não ocorrência) e menor do que 1 (da mesma forma, há a certeza da ocorrência do evento). Além disso, deve haver uma utilidade na ocorrência do evento.

2.1.1 Classificação de Risco

A incerteza que ao ser mensurada da origem ao risco é derivada de diversas fontes, se relacionando com diversas variáveis e correlacionando-se com distintos eventos econômicos, seja no tocante micro ou no macroeconômico. Crouhy Dan Galai (2005) propõe uma tipologia de classificação de risco financeiro apresentada na Figura 2.1.

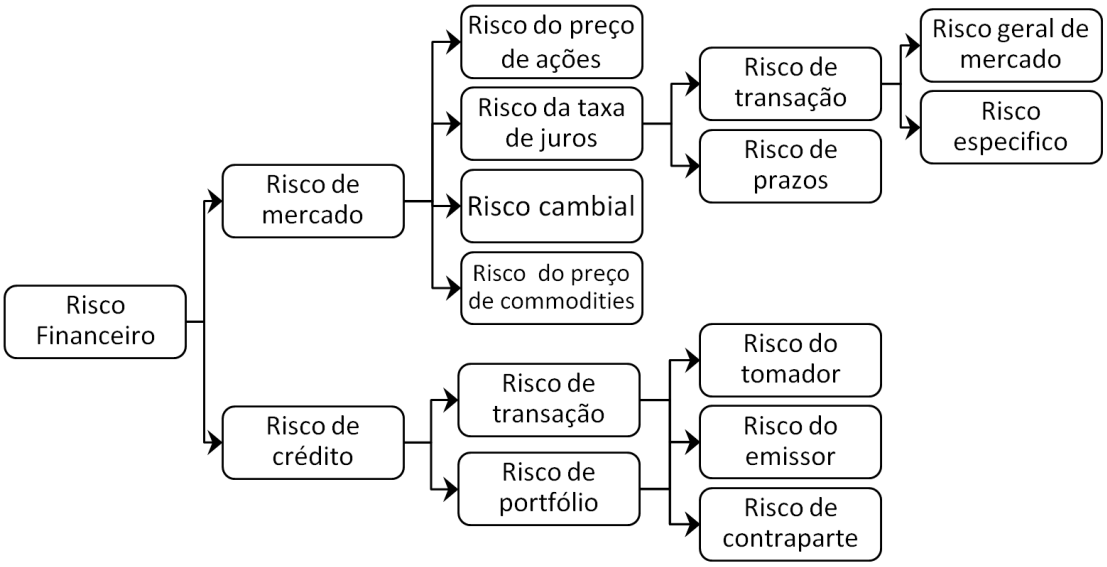


Figura 2.1: Categorias de riscos financeiros, adaptado de (CROUHY DAN GALAI, 2005)

O risco financeiro se divide, segundo a tipologia apresentada acima, em dois grupos principais. Risco de Crédito, que será discutido individualmente na seção 2.3, e Risco de Mercado, oriundo da incerteza quanto a trajetória dos preços, das taxas de juros, da demanda

e da oferta na economia, subdividindo-se, segundo Crouhy Dan Galai (2005) em:

a) Risco do preço das ações, associado a volatilidade do preço das ações, uma vez que a especulação e a flutuação são intrínsecas desse ramo, podendo ser minimizado a partir da diversificação da carteira, operações de *hedge* ou outras técnicas;

b) Risco da taxa de juros, proveniente de variações das taxas de juros, se subdividindo em risco de transação, que aborda o risco geral do mercado, componente genérica da incerteza do mercado em que se está operando e o risco especifico de cada operação. E por fim o risco de prazos que pode ser entendido como a oscilação das taxas de juros dentro de determinado período, conhecido como *gap risk*.

c) O risco cambial, oriundo das variações cambiais e da incerteza quanto suas trajetórias, dada a imprevisibilidade na ação dos agentes ou mesmo da politica cambial dos países;

d) Risco de preço de *commodities*, proveniente da variação no preço das *commodities*, seja por eventuais problemas na oferta ou na demanda desses produtos, ou mesmo pela ação especulativa.

Por ser o foco do presente trabalho, a definição de Risco de Crédito será feita em uma seção a seguir. Para sua correta compreensão, é necessária uma breve definição do conceito de crédito, a ser feito na próxima seção.

2.2 Crédito

Desde os primórdios, quando a humanidade passou a se comunicar e se organizar socialmente, mesmo que de forma precária, o ato de emprestar e tomar emprestado está presente (THOMAS, 2002). Tomando como exemplo, Lewis (1992) descreve um registro de crédito

datado de 2000 *A.C*, em que dois primitivos fazendeiros tomam empréstimos dando a colheita como garantia. O processo de garantia é essencial na prática creditícia, tendo impacto direto na precificação da operação, uma vez que minimizam o risco do credor, dizendo de outra forma, se as garantias são escassas, exige-se um retorno maior para abrir mão dos recursos no presente. O contrario também é verdadeiro, a existência de garantias reduz o risco e consequentemente o custo da operação para o credor. A própria origem da palavra crédito elucida esse fenômeno, uma vez que ela é derivada da palavra em latin *"credo"*, que significa confiar, acreditar.

Para ilustrar o crescimento do crédito nas últimas décadas, observa-se na Figura 2.2 a evolução do crédito a partir da década de 1970 [3], período que fez parte da conhecida era de ouro do capitalismo. Este período se caracteriza por um crescimento da renda, do produto e da produtividade, apoiado no estado de bem estar social dos países centrais. Desse panorama, consequentemente, há uma forte expansão do crédito.

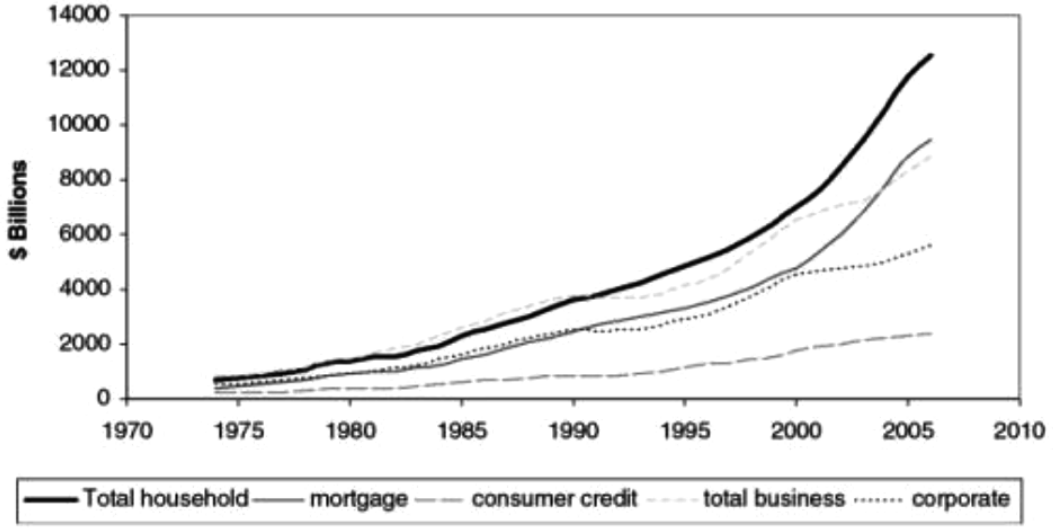


Figura 2.2: Composição da renda das famílias dos EUA, adaptado de (THOMAS, 2009)

[3]Para uma análise detalhada do crédito ao longo da história, veja (LEWIS, 1992); (THOMAS, 2002) e (THOMAS, 2009)

A Figura 2.2 exemplifica esse cenário ao demostrar o crescimento da renda das famílias nos EUA, em que podemos destacar a crescente importância e participação do crédito na composição de sua renda. Apesar dos créditos ao consumo (*consumer credit*) não ter uma expressiva participação no valor total de créditos, ele é o que possui o maior número de transações, tendo assim um baixo valor médio por transação. Isso é derivado principalmente do advento dos cartões de crédito e outros meios eletrônicos de pagamento, o que significou um maior número de transações focadas no crédito, desde pequenas compras com valores baixos, até as de maior expressão.[4]

A partir desse panorama, a definição de crédito a ser considerada, segundo a proposta de Hand (1997), é do crédito como um montante de dinheiro concedido a um agente econômico (sem fazer distinção a indivíduos, empresas ou outras instituições) por uma instituição financeira que espera receber juros, independente da forma com que será capitalizado, e o principal após determinado período de tempo acordado entre as partes.

Em outras palavras, conforme (SILVA, 2000):

> "*[...]o crédito consiste em colocar a disposição do cliente (tomador de recursos) certo valor sob a forma de empréstimos ou financiamentos, mediante uma promessa de pagamento numa data futura.* "

Da definição acima apresentada ficam evidentes duas características fundamentais do crédito. Ele está intimamente ligado à promessa de recebimento do valor concedido acrescido de um montante de juros, calculado mediante uma taxa de juros que, de forma simplificada, é uma função do custo de oportunidade e principalmente do risco que está exposto o credor. Esse risco é derivado da incerteza que se tem quanto à promessa do tomador em honrar o acordo. Por outro lado pode-se verificar um claro cenário em que há assimetria de

[4] *"[Com a introdução dos cartões de crédito] os consumidores passam a contar com um produto que lhes permite usar o crédito para quase todas as compras, de pequenos alimentos até passagens de avião"* (THOMAS, 2009, p. 2)

informação, em que o tomador de crédito está mais bem informado acerca de sua condição de honrar seus compromissos do que a instituição financeira, que tem um acesso restrito às informações do credor.

2.3 Risco de Crédito

A definição de risco de crédito será construída a partir dos conceitos de risco e crédito anteriormente definidos. Uma vez que o risco é a mensuração da incerteza via atribuição de probabilidade aos eventos relevantes em termos de utilidade, e crédito como a disponibilização de um montante de recursos, em que o ofertador espera receber do demandante após um período pré-acordado, o principal acrescido de uma taxa de juros. Podemos definir o risco de crédito como a mensuração da incerteza do ofertador de crédito não receber os recursos esperados, [5]como acordado no momento da concessão entre as partes. O risco de crédito é função de duas grandezas principais, o montante de recursos e a probabilidade de inadimplência, daqui para frente definida como *PD*, do inglês *probability of default*.

Portanto, o risco de crédito é derivado da incerteza oriunda da concessão de crédito. A gestão do risco de crédito passa obrigatoriamente pela estimação da PD, onde busca-se mensurar o grau de exposição ao risco que está sujeita uma instituição financeira, seja olhando individualmente para cada operação de crédito, seja olhando para todo o portfólio. Isso terá um grande impacto na determinação da taxa de juros a ser praticada no momento da concessão (custo da operação), ou mesmo na determinação do montante de capital que deverá ser alocado pelas instituições financeiras como provisão para a perda esperada.[6].

[5]Considera-se recursos como o principal cedido e os juros que se espera receber como a remuneração dessa concessão.

[6]A perda esperada se configura como o produto $PE = (PD \times EAD \times LGD)$, em que EAD é a exposição (montante de recursos expostos) a inadimplência, do inglês *Exposition at Default* e o LGD a perda dada a inadimplência, do inglês *Loss Given Default*. Para uma explicação detalhada sobre perda esperada, (LIMA, 2008)

Um bom exemplo da importância de uma gestão do risco de crédito, que pode ter consequências sistêmicas relevantes se realizada de forma displicente e não apoiada em sólidas e confiáveis informações é dado por Lahsasna (2010), que liga a grave crise econômica que se iniciou no mercado imobiliário norte americano de alto risco, *subprimes*, no final de 2008 a uma gestão displicente do risco de crédito, levando a um crescimento da inadimplência nesse setor. Uma vez que o mercado imobiliário é fortemente securitizado, um crescimento da inadimplência reduz a rentabilidade e a confiança nos papéis provenientes desse mercado, contaminando mesmo que de forma indireta os mais distintos mercados por todo o globo.

Ainda fazendo uso da crise dos *subprimes* no contexto do trabalho, como exemplificação da importância de uma correta avaliação do risco de crédito, para Crouhy Dan Galai (2005), os empréstimos *subprime* foram, no pré-crise, uma das mais promissoras e problemáticas atividades bancarias. Promissora porque os clientes *subprime* eram caracterizados por históricos de crédito deteriorados (*poor or impaired* segundo o autor), precisando pagar altas taxas de juros, elevando a rentabilidade dos empréstimos. Problemática no sentido que tratava-se de um ramo novo de empréstimos, não contando com um volume substancial de informações que permitisse a realização de uma predição eficiente da *PD*, ou seja, do risco que estava implícito na operação.

2.4 *Credit Score*

Desse contexto permeado pela incerteza e principalmente pela assimetria de informações, os modelos de *credit score* têm uma fundamental importância na gestão do risco e mesmo na manutenção da estabilidade econômica. Para Sicsú (2010), o *credit score* é uma medida do risco de crédito. Em outras palavras, por meio dos modelos de *credit score* é possível atribuir uma probabilidade da inadimplência, ou seja, esses modelos são instrumentos para que se

possa mensurar a *PD* das operações de crédito.

Fazendo uso da noção de risco vista na seção 2.1, as fórmulas de *credit score* têm o objetivo de inferir a probabilidade do tomador de crédito tornar-se ou não inadimplente. Para isso, faz-se uso de informações internas da instituição, externas ou de conjuntura macroeconômicas com o intuito de encontrar uma relação entre os credores do passado e do presente com os do futuro (LAI, 2006).

Portanto, o *credit score* é essencialmente uma forma de identificar diferentes grupos em uma população, buscando caracteristicas que os segregem, para que no futuro proximo novos individuos possam ser classificados corretamente de acordo com suas caracteristicas no grupo que mais lhe é semelhante (THOMAS, 2002). Em outras palavras, os modelos de *credit score* se utilizam das informações dos tomadores de crédito para estimar a *PD*, quantificando o risco desse tomador, usando como base para a inferência informações prévias da população.

Apesar do crédito, como visto na seção 2.2 ser tão antigo quanto a própria humanidade, modelos de *credit score* surgiram primeiramente com a introdução de métodos estatísticos para identificação de grupos em populações, apresentadas em Fisher (1936). Estes modelos foram pouco utilizados nas práticas comerciais, ganhando fôlego no final dos anos 1960, que como dito anteriormente, coincide com a explosão no uso de cartões de crédito, levando a uma pulverização e generalização do uso de produtos de crédito.

O crescimento no uso desses modelos se deu, segundo Yao (2009) pelo fato deles permitirem uma redução no custo da análise do crédito, acelerando a tomada de decisão, além de atribuir um caráter mais científico e impessoal.

Essas vantagens vão de encontro com o crescente volume de transações, principalmente com o uso de cartões de crédito, que não podem mais ser avaliadas individualmente como

eram feitas até então (FENSTERSTOCK, 2005). Anteriormente, a avaliação da concessão de crédito era chamada de julgamental, e tinha a desvantagem de que cada transação devia ser analisada individualmente, além de não contar com a uniformidade da resposta, uma vez que era baseada tanto em informações qualitativas, quanto na avaliação individual do analista. Com a avaliação julgamental não era possível levar em consideração características mais gerais da população, não permitindo uma visualização do todo e das tendências da população.

Em 1980, com o advento da informática e o sucesso dos modelos de *credit score* nos cartões de crédito, as instituições financeiras aplicaram esses modelos para outros produtos bancários, desde crédito imobiliário até a mensuração de retorno de campanhas publicitárias (THOMAS, 2002, p. 4). Essa ampliação das possibilidades de aplicação dos modelos comprova a eficácia das técnicas, sendo corroborado pela opinião de Gayler (2006), ao afirmar que o volume total de recursos sob controle de modelos de *credit score* é imensa, e os ganhos provenientes de melhorias que deem maior confiança e credibilidade aos modelos seriam também imensos.

Por isso, o crescente uso dessas ferramentas, o aumento no número de transações e a complexidade dos modernos produtos financeiros foram fatores que levaram a pesquisas por melhores e mais confiáveis modelos de *credit score*, em que se extrapola o campo puramente estatístico, buscando melhorias de *data mining* baseadas em diversos ramos do conhecimento, como a inteligência artificial.

3 Dados

Ao longo desse trabalho usaremos os dados da base *German Credit Data*, (FRANK; ASUNCION, 2010). A base é composta por 1000 observações, com 21 variáveis, 20 delas são variáveis explicativas (7 categóricas e 13 continuas) além de uma variável resposta binária, que classifica o individuo como adimplente ou inadimplente. A variável de saída contem 700 observações com valor 1 (adimplente) e 300 com valor 2 (inadimplente).

Não foi feita qualquer manipulação das variáveis, apenas a renomeação da categoria das variáveis categóricas, com o objetivo de simplificar o entendimento, conforme a rotina do Apêndice A.1. Um resumo dos dados se encontra na Tabela 3.1. As 13 variáveis categóricas estão resumidas na Tabela 3.2,e 7 variáveis continuas estão descritas na Tabela 3.3.

As Tabelas 3.2 e 3.3 descrevem os vinte atributos compondo inicialmente os padrões de entrada associados aos N perfis de clientes da base de dados. As variáveis categóricas encontram-se descritas na Tabela 3.2, enquanto a Tabela 3.3 descreve as variáveis que assumem valores reais.

Tabela 3.1: *German Credit Data*

Observações	Adimplentes	Inadimplentes	Var. Continuas	Var. Categorizadas	Var. Resposta
1000	700	300	7	13	1

Tabela 3.2: Variáveis Categoricas

Cód..	Descrição	Categorias
VAR1	Status de conta corrente já existente	4
VAR3	Histórico de Crédito	5
VAR4	Propósito do empréstimo	10
VAR6	Contas de poupança / Títulos	5
VAR7	Duração do atual emprego	5
VAR9	Estado civil e sexo	5
VAR10	Outros fiadores/Garantidores	3
VAR12	Patrimônio	4
VAR14	Outros planos de empréstimos	3
VAR15	Status da habitação	3
VAR17	Categoria do trabalho	4
VAR19	Tem telefone	2
VAR20	Estrangeiro ou não	2

Tabela 3.3: Variáveis Continuas

Cód..	Descrição	Mínimo	Máximo	Média	Desv.Pad
VAR2	Duração em meses do empréstimo	4,0	72,0	20,9	145,4
VAR5	Quantia do empréstimo	250	18.424	3.271	2.822,74
VAR8	% da parcela em relação a renda	1,0	8,0	3,0	1,14
VAR11	Anos na atual residência	1,0	11,0	2,9	1,09
VAR13	Idade em anos	13,0	75,0	35,5	129,4
VAR16	Número de créditos no banco	1,0	4,0	1,4	0,55
VAR18	Número de dependentes	1,0	2,0	1,2	0,31

3.1 Seleção de variáveis

Com o objetivo de simplificar nossa análise, bem como dar maior consistência aos resultados, não serão utilizadas todas as variáveis explicativas. Usaremos o algoritmo *Stepwise* proposto por Draper (1998) que afirma que esse é a melhor metodologia para seleção de variáveis, por não demandar muitos recursos computacionais além de avaliar os parâmetros escolhidos iterativamente e não de forma individual, mas em grupo[1]. O algoritmo será executado no *SAS* por meio do *Procedimento PHREG*, conforme rotina descrita no Apêndice A.2.

[1](DRAPER, 1998, p. 335)

O propósito do algoritmo é ir incluindo as variáveis ao modelo proposto e verificar se sua inclusão eleva a capacidade de predição, se sim, essa variável é mantida, e avalia-se o modelo novamente com a inclusão de uma próxima variável. Se a variável incluída não agrega qualquer poder preditivo, ela é então retirada. Esse processo é executado iterativamente, para todas as variáveis explicativas da base de dados.

As variáveis são incluídas de acordo com sua significância, mensurada pela estatística qui-quadrado [2]. A Tabela 3.4 trás as variáveis ordenadas de forma decrescente por sua significância. Ela é construída olhando-se apenas para a relação com a variável resposta.

Tabela 3.4: Ordenação das variáveis por significância

Variável	G.Liberdade	Chi-Square	Pr > ChiSq	Descrição da Variável
VAR1	3	28,993	0,000	Status de conta corrente já existente
VAR2	1	9,125	0,003	Duração em meses do empréstimo
VAR3	4	12,211	0,016	Histórico de crédito
VAR5	1	4,299	0,038	Quantia do Empréstimo
VAR6	4	8,484	0,075	Contas de poupança/Títulos
VAR15	2	3,653	0,161	Status da habitação
VAR12	3	4,907	0,179	Patrimônio
VAR13	1	1,759	0,185	Idade em anos
VAR20	1	0,175	0,187	Estrangeiro ou não
VAR14	2	2,508	0,285	Outros planos de empréstimo
VAR8	1	1,120	0,290	% da parcela em relação a renda
VAR7	4	3,789	0,435	Duração do atual emprego
VAR10	2	1,376	0,503	Outros fiadores/garantidores
VAR16	1	0,448	0,503	Número de créditos existentes no banco
VAR9	3	1,964	0,580	Estado civil e sexo
VAR4	8	6,559	0,585	Propósito do empréstimo
VAR19	1	0,281	0,596	Tem telefone
VAR17	3	0,385	0,943	Categoria do trabalho
VAR18	1	0,002	0,965	Número de dependentes
VAR11	1	0,002	0,966	Anos na atual residencia

[2] O teste estatístico que define essa distribuição e atribui um grau de significância, bem como a explicação do teste de hipóteses para as variáveis não serão abordados no presente trabalho, por fugirem do escopo. Maiores detalhes podem ser obtidos em (HOSMER; LEMESHOW, 2000, p. 116)

A cada passo do algoritmo uma tabela igual a Tabela 3.4 é construída, não mais tomando como referencia a variável resposta, mas sim as variáveis que vão compondo o modelo proposto no passo imediatamente anterior. Como critério para a entrada das variáveis no modelo foi atribuído o valor de 1 e para que se mantenha no modelo o valor de 0,5 [3]. Com isso todas as variáveis estão passiveis de entrarem no modelo e serem testadas, mas apenas as que tiverem um grau de significância $\alpha = 0,5$ serão mantidas.

Conforme a Tabela 3.4 a primeira variável a ser incluída será a $VAR1$, constituindo-se assim um modelo da seguinte forma:

$$\hat{R} = \alpha + \beta VAR \tag{3.1}$$

É então recalculada a significância de todas as variáveis contra o modelo da Equação 3.1. Esse processo é então repetido, até que nenhuma variável seja significativa, de acordo com o nível determinado previamente.

Após 13 passos, chegou-se ao resultado final, sumarizado na Tabela 3.5. Conforme dito anteriormente, são mantidas apenas as variáveis com um grau de significância maior do que 0,5. Esse é o motivo que determinou que a variável $VAR9$ fosse retirada do modelo.

Portanto, daqui em diante iremos considerar a variável resposta R como sendo uma função das variáveis eleitas, conforme a equação 3.2.

$$\hat{R} = f(VAR1, VAR2, VAR3, VAR20, VAR13, VAR10, VAR8, VAR5, VAR19, VAR14, VAR4) \tag{3.2}$$

[3] Parâmetro SLENTRY $= 1$ e SLSTAY $= 0,5$ respectivamente, na rotina SAS.

Tabela 3.5: Resultado final do *stepwise*

Passo	Var. Adicionada	Var. Removida	Gl	Chi-Square	Pr > ChiSq
1	VAR1		3	28,993	0,000
2	VAR2		1	7,070	0,008
3	VAR3		4	6,548	0,184
4	VAR20		1	1,249	0,264
5	VAR13		1	0,919	0,338
6	VAR10		2	2,159	0,340
7	VAR8		1	0,787	0,375
8	VAR5		1	0,815	0,367
9	VAR19		1	0,829	0,363
10	VAR14		2	1,767	0,413
11	VAR4		8	8,053	0,428
12	VAR9		3	1,914	0,591
13		VAR9	3	1,912	0,591

Uma vez que esse conjunto de variáveis será utilizado na construção dos modelos de *credit score*, convêm uma análise a respeito da distribuição de cada uma delas, seu intervalo de valores além de sua distribuição em relação à variável resposta.

3.2 Análise das variáveis selecionadas

A primeira variável explicativa, cujo código é *VAR1* agrupa os indivíduos pelo tempo de conta existente em 4 categorias, conforme Tabela 3.6. Como pode ser visto na Figura 3.1 seus valores para os clientes adimplentes (resposta = 0) concentram-se na categoria sem conta corrente e os inadimplentes, por sua vez, tem uma concentração maior na categoria 0, que são os clientes com menos de 0 DM. [4].

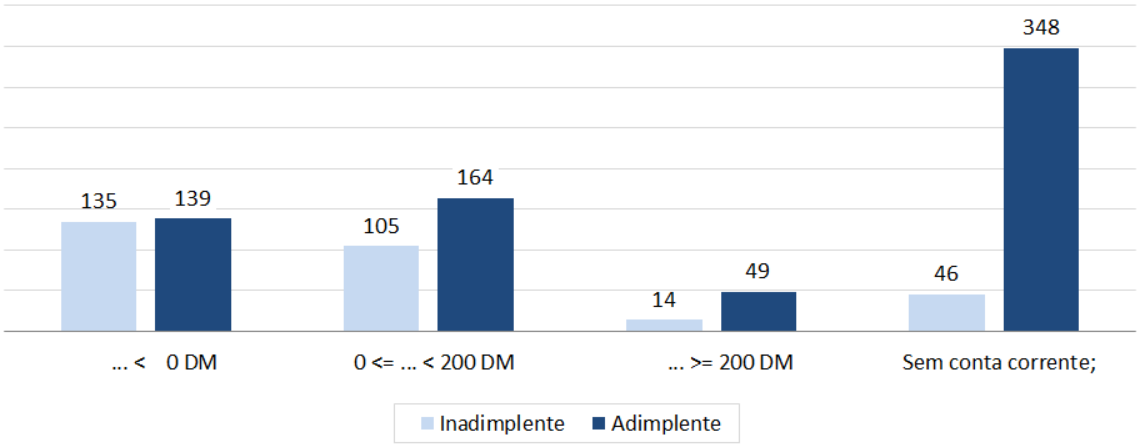
A variável *VAR2*, mensura a duração em meses do crédito concedido, com média de 20,9 meses e desvio padrão de 12,06 conforme a Tabela 3.7.

[4]DM é a sigla para *Deutsche Mark*, marcos alemães, uma vez que a base de dados é alemã e do ano de 1997.

Tabela 3.6: Variável *VAR1*

Cód.	VAR1
Descrição	Status de conta corrente já existente
0	< 0 DM
1	≥ 0 e $< 200DM$
2	$\geq 200DM$
3	Sem conta bancaria

Figura 3.1: Distribuição da Variável *VAR1* em relação a variável resposta



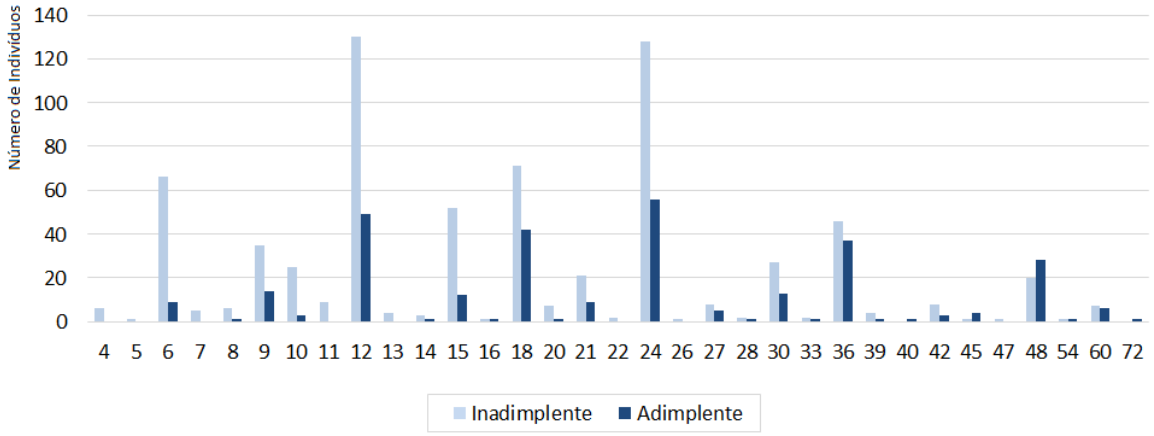
Em relação a variável resposta, os bons pagadores se concentram, principalmente nas categorias de empréstimos de 12 e 24 meses, porém com valores não desprezíveis em 6 e 18, como pode ser visto na Figura 3.2, os maus pagadores, por outro lado estão concentrados em 12, 18, 24 e 36 meses.

Tabela 3.7: Variável *VAR2*

Cód.	VAR2
Descrição	Duração em meses do empréstimo
Min.	4
Max.	72
Med.	20,9
Desv. Pad	12,06

A terceira variável selecionada *VAR3*, corresponde ao histórico de crédito do cliente. São 5 categorias, descritas na Tabela 3.8, e as observações, tanto em relação aos adimplentes

Figura 3.2: Distribuição da Variável *VAR2* em relação a variável resposta



quanto aos inadimplentes se concentram na categoria com valor 2, que são os clientes com créditos e que efetuaram o pagamento corretamente até o presente momento.

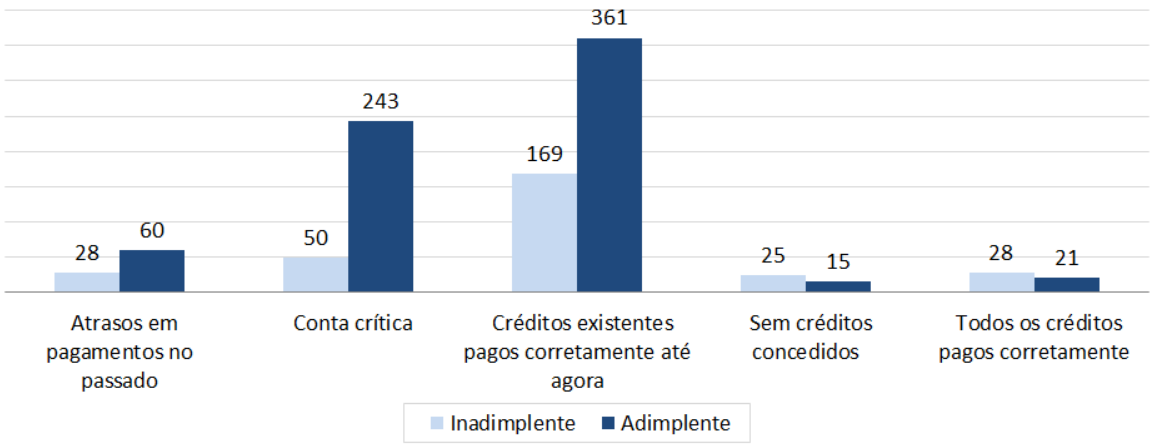
A categoria cujo valor é 4 tem um peso proporcionalmente maior quando a variável de resposta é 0 (como pode ser visto no gráfico 3.3).A descrição de que ou os clientes possuem contas criticas ou créditos em outra instituição financeira nos leva a inferir que o maior número de pessoas se encontra na segunda alternativa, uma vez que seria inconsistente supor que grande parte dos clientes adimplentes possuem contas criticas (independente do conceito de crítico nos dados).

Tabela 3.8: Variável *VAR3*

Cód.	VAR3
Descrição	Histórico de Crédito
0	Sem créditos concedidos
1	Todos os créditos pagos corretamente
2	Créditos existentes pagos corretamente ate agora
3	Atrasos em pagamentos no passado
4	Conta crítica

Seguindo a ordem da Tabela 3.5, a quarta variável a ser incluída no modelo foi a *VAR20*. Ela indica se o indivíduo e estrangeiro ou não, conforme a Tabela 3.9. Segundo pode ser constatado na Figura 3.4 grande parte dos indivíduos são estrangeiros, porem, não ha forte

Figura 3.3: Distribuição da Variável *VAR3* em relação a variável resposta



concentração entre adimplentes e inadimplentes.

Tabela 3.9: Variável *VAR20*

Cód.	VAR20
Descrição	Estrangeiro ou não
0	Sim
1	Não

A variável *VAR13* é a idade dos indivíduos em anos. De cordo com a Figura 3.5, a distribuição das idades é muito próxima entre os adimplentes e inadimplentes. Em ambas as sub populações as idades se concentram em torno da média geral de 35,5. Os valores máximo, mínimo, além do desvio padrão estão na Tabela 3.10.

Tabela 3.10: Variável *VAR13*

Cód.	VAR20
Descrição	Idade em anos
Min.	19
Max.	75
Med.	35,5
Desv. Pad	11,37

A grande maioria dos indivíduos não possuem fiadores ou garantidores. Essa informação está na variável *VAR10*. Essa informação pode ser verificada na Figura 3.6 e na Tabela 3.11.

Figura 3.4: Distribuição da Variável *VAR20* em relação a variável resposta

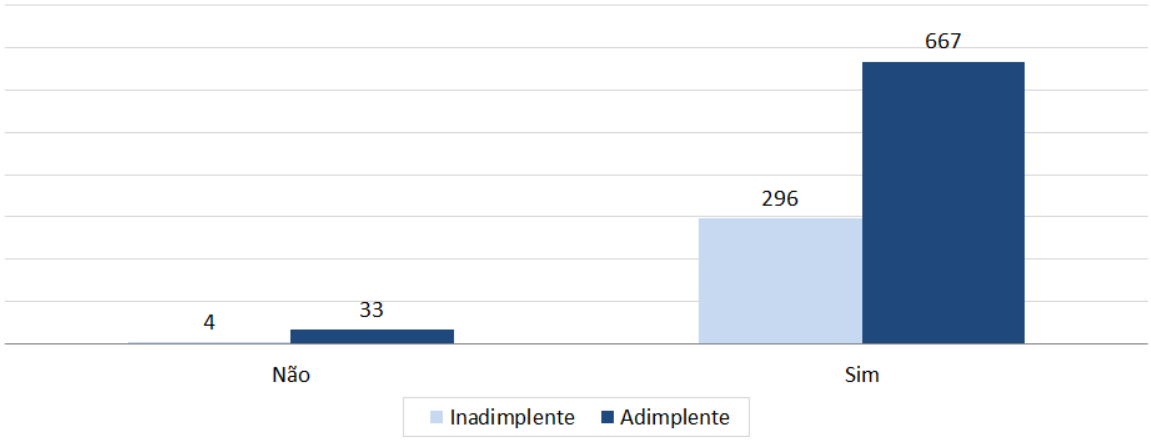
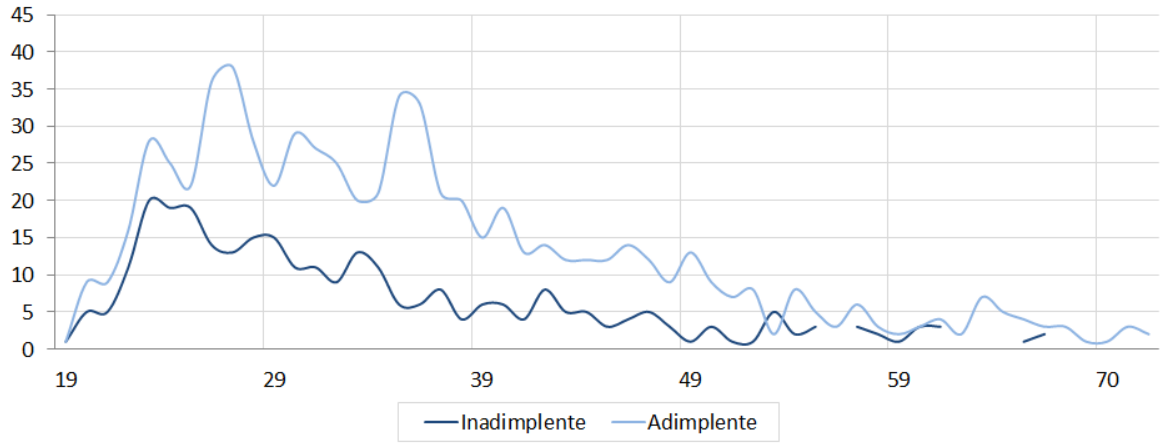


Figura 3.5: Distribuição da Variável *VAR13* em relação a variável resposta



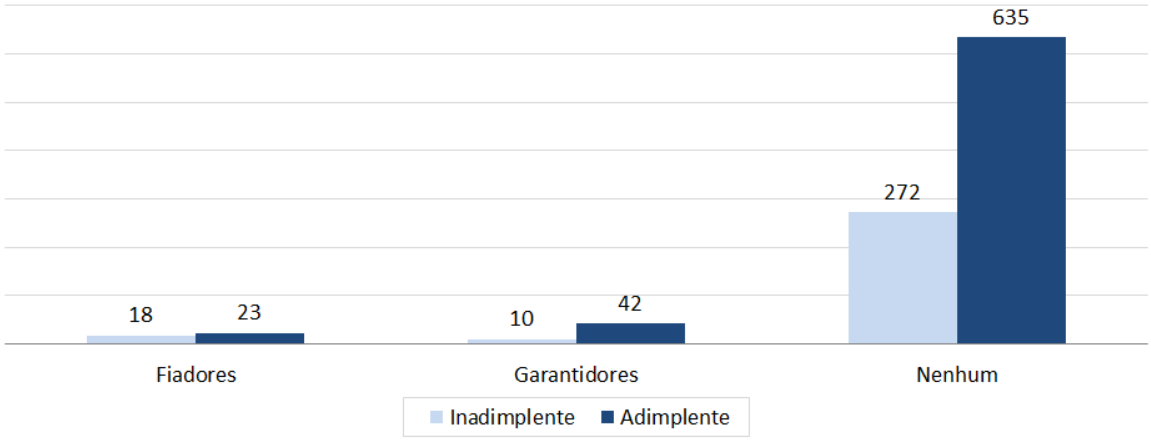
A variável *VAR8* fornece o percentual do valor da parcela em relação a renda disponível. Grande parte das observações se encontram com uma taxa de 4%, sem diferenças consideráveis entre adimplentes e inadimplentes, conforme a Figura 3.7 e a Tabela 3.12.

O valor dos empréstimos está descrito na variável *VAR5*. A Figura 3.8 representa o histograma das sub populações de indivíduos adimplentes e inadimplentes. Em ambas, os valores estão concentrados e, torno da média geral de 3.271,25, conforme a Tabela 3.13. Não há uma clara distinção entre as duas sub populações em relação ao risco de credito.

Tabela 3.11: Variável *VAR10*

Cód.	VAR10
Descrição	Outros Fiadores/Garantidores
0	Nenhum
1	Fiadores
2	Garantidores

Figura 3.6: Distribuição da Variável *VAR10* em relação a variável resposta



Conforme pode ser visto na Figura 3.9, uma parcela expressiva dos indivíduos não tem telefone, segundo os dados da variável *VAR19*. Essa variável é binaria e diz se o individuo tem ou não telefone, conforme a Tabela 3.14.

A variável *VAR14*, tem a informação dos outros empréstimos tomados pelos indivíduos (Tabela 3.15). São 3 categorias possíveis, 0, 1 e 2, representando os empréstimos a bancos, lojas ou nenhum, respectivamente. Tanto entre os adimplentes quanto os inadimplentes, grande parte das observações se concentram na ultima faixa, e o menor valor na faixa intermediaria, ou seja, poucos indivíduos tem créditos em lojas, conforme Gráfico 3.10.

Por fim, a ultima variável elegível foi a *VAR4*, que nos diz o proposito do empréstimo, o motivo da operação de credito. Segundo pode ser visto na Tabela 3.16, são 11 categorias. Duas delas não tem qualquer observação, e por isso, a variável tem efetivamente 8 categorias,

Tabela 3.12: Variável *VAR8*

Cód.	VAR8
Descrição	% da parcela em relação a renda
1	1%
2	2%
3	3%
4	4%

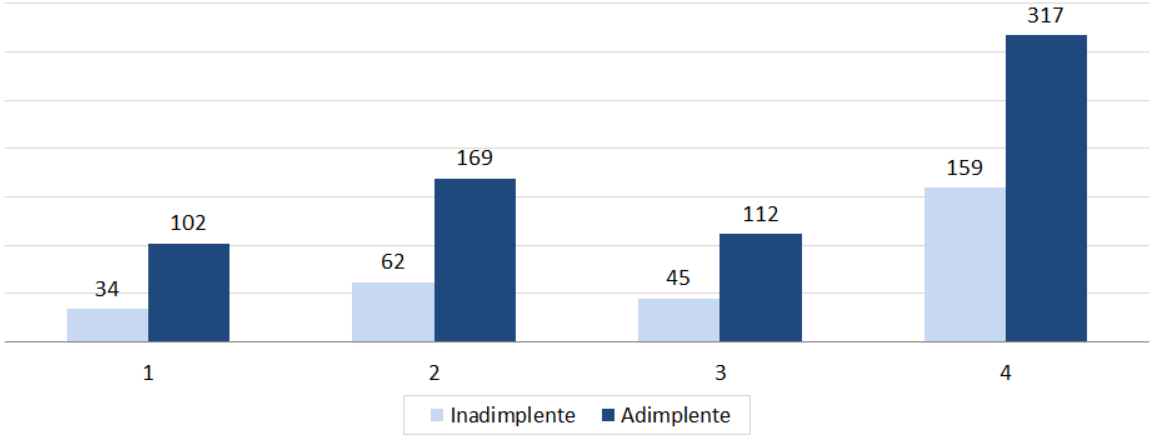


Figura 3.7: Distribuição da Variável *VAR8* em relação a variável resposta

conforme pode ser verificado na Figura 3.11.

De acordo com essa mesma figura, há uma concentração relativamente maior dos clientes

Cód.	VAR5
Descrição	Quantia do Empréstimo
Min.	250
Max.	18.424
Med.	3.271,25
Desv. Pad	2.822,74

Tabela 3.13: Variável *VAR5*

Cód.	VAR19
Descrição	Tem telefone
0	Não
1	Sim

Tabela 3.14: Variável *VAR19*

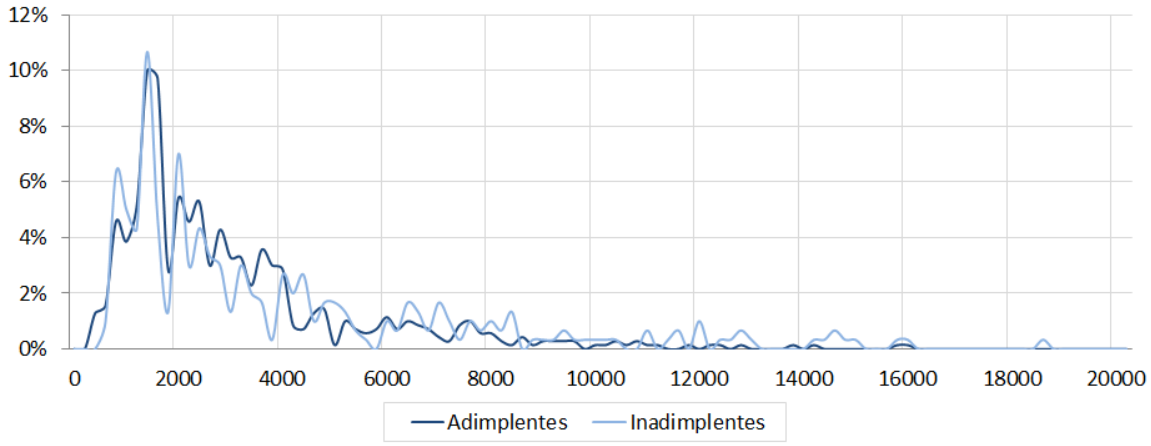


Figura 3.8: Distribuição da Variável *VAR5* em relação a variável resposta

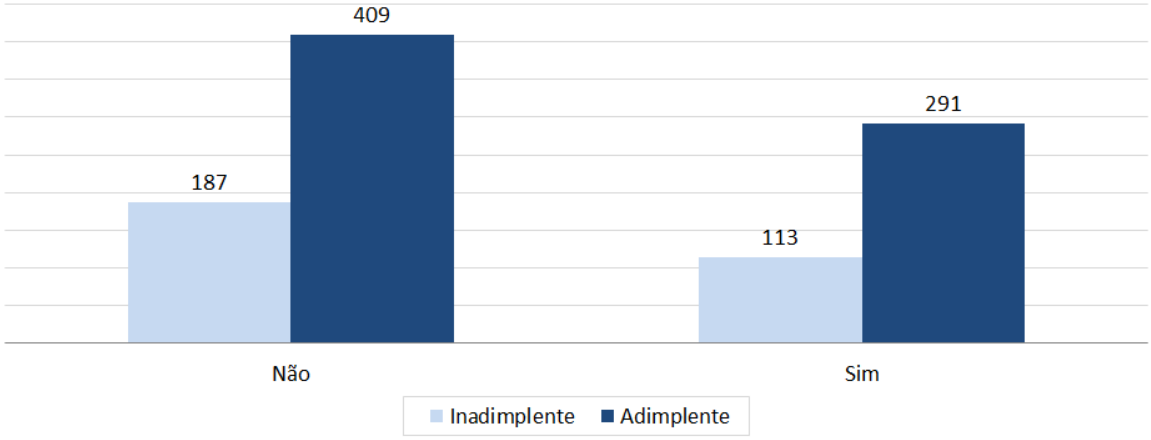


Figura 3.9: Distribuição da Variável *VAR19* em relação a variável resposta

Cód.	VAR14
Descrição	Outros planos de empréstimos
0	Bancos
1	Lojas
2	Nenhum

Tabela 3.15: Variável *VAR14*

adimplentes nas categorias de Carros(Usado) e Radio/Televisão. Por outro lado, a categoria educação tem praticamente o mesmo numero de indivíduos adimplentes e inadimplentes.

A Tabela 3.17 sintetiza os dados que serão utilizados.

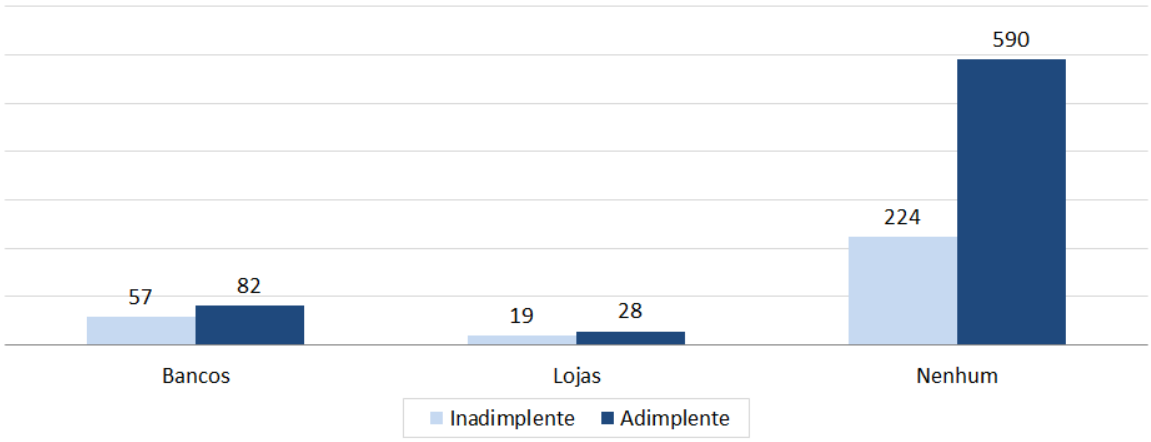


Figura 3.10: Distribuição da Variável *VAR14* contra *R*

Cód.	VAR4
Descrição	Propósito do empréstimo
0	Carro(novo)
1	Carro(usado)
2	Moveis/Equipamentos
3	Rádio/Televisão
4	Aparelhos domésticos
5	Reformas
6	Educação
7	Férias
8	Capacitação
9	Negócios
10	Outros

Tabela 3.16: Variável *VAR4*

3.3 Variáveis *Dummy*

Para a correta comparação do desempenho dos modelos a serem construídos, consideraremos a base de dados em seu formato original, sem qualquer manipulação das variáveis. Essa conduta será seguida para que qualquer tratamento não venha a favorecer uma técnica em detrimento de outra. Mesmo porque seria difícil identificar benefícios ou prejuízos decorrentes de qualquer manipulação dos dados de forma isolada.

Entretanto, o modelo de regressão logística exige que as variáveis categóricas sejam con-

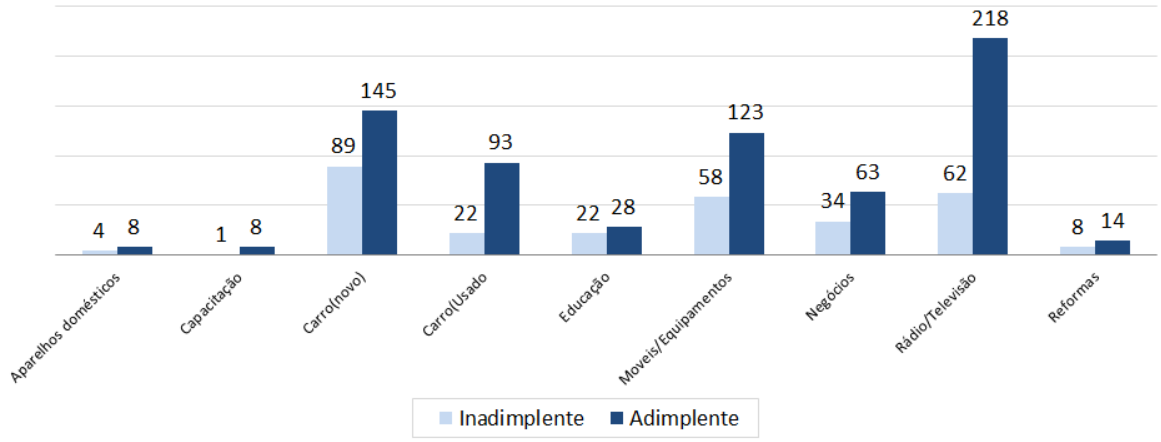


Figura 3.11: Distribuição da Variável *VAR4* em relação a variável resposta

vertidas em *dummy´s* para a correta estimação de seus coeficientes. De forma breve, esse procedimento pode ser entendido como a transformação de uma variável categórica em uma série de novas variáveis dicotômicas.

Sendo n o número de categorias em cada uma das variáveis *VAR1, VAR3 VAR4 VAR10 VAR14 VAR19* e *VAR20* foram consideradas $(n-1)$ categorias, ou seja, foram criadas $(n-1)$ variáveis *dummy* para cada uma das variáveis originalmente categóricas. Por exemplo, a variável *VAR20* contêm duas categorias, portanto, foi criada uma única variável *dummy´s*, que codifica em 0 a primeira categoria e 1 a outra. Ou seja, não é necessária a criação de uma variável *dummy* para cada categoria de uma variável.

A rotina de criação destas variáveis está descrita no Apêndice A.3.

3.4 Amostragem

Para o ajuste dos modelos, serão utilizados 70% dos dados originais. Os 30% restantes vão compor uma base de dados para a validação dos modelos. Com esta divisão garante-se uma fração dos dados que não foram utilizados na construção dos modelos, e portanto

não possuem suas características idiossincráticas captadas, para uma efetiva mensuração da capacidade preditiva dos modelos.

Será respeitada a fração de 70% dos clientes adimplentes e os outros 30% de inadimplentes seja na base de dados de treinamento, seja na de validação.

Tabela 3.18: *Log* do processo de amostragem

The SURVEYSELECT Procedure	
Selection Method	Simple Random Sampling
Input Data Set	DADOS
Random Number Seed	569144001
Sampling Rate	0.7
Sample Size	700
Selection Probability	0.7
Sampling Weight	1.428.571
Output Data Set	Treinamento

O processo de amostragem é feito pelo *SAS*, com um sorteio aleatório e está descrito no Apêndice A.3. A Tabela 3.18 fornece um resumo do sorteio, além da semente utilizada no processo de amostragem, permitindo uma replicação dos dados.

Tabela 3.17: Quadro final de variáveis

Variaveis/Categorias	Adimplentes	Inadimplentes	Total
VAR1 -Status de conta corrente já existente			
... < 0 DM	139	135	274
... >= 200 DM	49	14	63
0 <= ...< 200 DM	164	105	269
Sem conta corrente;	348	46	394
VAR2 -Duração em meses do empréstimo	700	300	1.000
VAR3 -Histórico de crédito			
Atrasos em pagamentos no passado	60	28	88
Conta crítica	243	50	293
Créditos existentes pagos corretamente até agora	361	169	530
Sem créditos concedidos	15	25	40
Todos os créditos pagos corretamente	21	28	49
VAR4 -Propósito do empréstimo			
Aparelhos domésticos	8	4	12
Capacitação	8	1	9
Carro(novo)	145	89	234
Carro(Usado	93	22	115
Educação	28	22	50
Moveis/Equipamentos	123	58	181
Negócios	63	34	97
Rádio/Televisão	218	62	280
Reformas	14	8	22
VAR5 -Quantia do Empréstimo	700	300	1.000
VAR8 -% da parcela em relação a renda			
1	102	34	136
2	169	62	231
3	112	45	157
4	317	159	476
VAR10 -Outros fiadores/garantidores			
Fiadores	23	18	41
Garantidores	42	10	52
Nenhum	635	272	907
VAR13 -Idade em anos	700	300	1.000
VAR14 -Outros planos de empréstimo			
Bancos	82	57	139
Lojas	28	19	47
Nenhum	590	224	814
VAR19 -Tem telefone			
Não	409	187	596
Sim	291	113	404
VAR20 -Estrangeiro ou não			
Não	33	4	37
Sim	667	296	963

4 Modelos

4.1 Regressão Logística

O primeiro modelo a ser considerado para a elaboração de *credit scores* será o modelo de regressão logística conforme (HOSMER; LEMESHOW, 2000). Os principais motivos para a escolha desse modelo, segundo o autor, é a sua adequação para conjuntos de dados em que a variável resposta é binária, sua flexibilidade e facilidade de uso.

Podemos interpretar essa regressão da seguinte forma. Dada uma variável resposta dicotômica y e uma única variável explicativa x, o modelo logístico pode ser representado por:

$$\hat{y}_i = E(y_i|x_i) + \varepsilon_i. \tag{4.1}$$

Em que $\hat{y}_i$ é a i-ésima estimativa da i-ésima observação y_i, $E(y_i|x_i)$ é a esperança condicional de y_i dado x_i e ε_i é o erro entre a i-ésima variável resposta observada (y_i) e estimada ($\hat{y}_i$). Ao contrario do modelo de regressão linear [1], não espera-se que ε_i seja normalmente distribuído com média $\mu = 0$ e desvio padrão σ^2 ou seja, que $\varepsilon_i \sim N(\mu, \sigma^2)$.

Substituindo $E(y_i|x_i)$ por ψi, a esperança condicional pode ser escrita como segue na equação 4.2:

[1] As características e propriedades desse modelo podem ser consultadas em (HOFFMANN, 2006)

$$\psi(x_i) = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \tag{4.2}$$

A transformação de $\psi(x)$ no *logit* $g(x)$ está na equação 4.3:

$$g(x) = ln\left[\frac{\psi(x_i)}{1-\psi(x_i)}\right] \tag{4.3}$$

A grande importância dessa transformação é que $g(x)$ é linear em seus parâmetros, continua e definida no intervalo $-\infty \leqslant g(x) \leqslant \infty$ dependendo dos valores de x. Assumindo uma função de apenas uma variável explicativa x a Figura 4.1 ilustra domínio de ψi, em que $\xi(x_i) = \alpha + \beta x_i$. As propriedades de $\xi(x_i)$ serão analisadas na seção 4.1.1, bem como a estimação dos seus coeficientes..

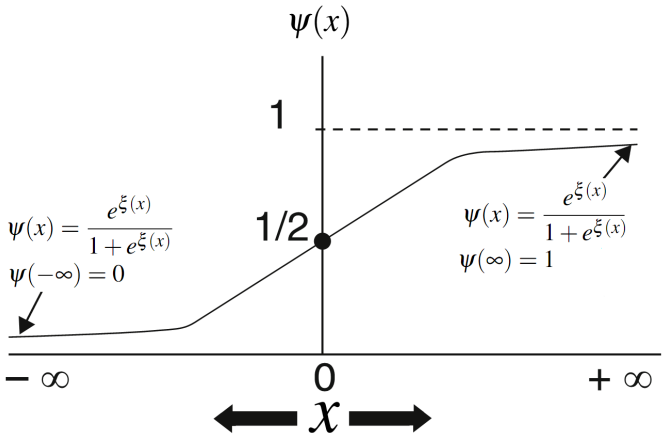


Figura 4.1: Curva padrão da regressão logística, adaptado de (KLEIN, 2010)

Como a variável resposta y_i é dicotômica, o erro ε_i pode assumir apenas dois valores. Se $y_i = 1$ então $\varepsilon_i = 1 - \psi(x_i)$ com probabilidade $\psi(x_i)$. Se $yi = 0$ então $\varepsilon_i = -\psi(x_i)$ com probabilidade $1 - \psi(x_i)$. Das propriedades da distribuição binomial[2], podemos afirmar que ε tem distribuição com média $\mu = 0$ e variância $\sigma^2 = \psi(x)[1 + \psi(x)]$. A distribuição condicional

[2]As propriedades da distribuição binomial fogem do escopo do presente trabalho, mas elas podem ser encontradas em (MEYER, 1983, cap. 4) e (ROSS, 2010, cap. 4).

da variável resposta $\hat{y}_i$ tem distribuição binomial com probabilidade dada por $\psi(x_i)$.

4.1.1 Estimação dos parâmetros

Este trabalho utilizará o modelo logístico múltiplo, que será definido a seguir. Dadas as i variáveis explicativas anteriormente selecionadas, denota-se o vetor $\mathbf{x}' = (x_1, x_2..., x_i)$. em que $x_1 \cdots x_i$ são as i variáveis da Tabela 3.17. A probabilidade condicional do i-ésimo individuo Y_i ser inadimplente ($Y_i = 1$) dado seu vetor de características $\mathbf{x}$, conforme seção 3.4 será:

$$P(Y = 1|\mathbf{x}) = \psi(\mathbf{x}) \tag{4.4}$$

Ou seja, a probabilidade do i-ésimo individuo da amostra ser inadimplente dado seu vetor de características é igual a esperança matemática de que ele seja inadimplente dado o mesmo vetor de características, conforme definido acima $\psi i = E(y_i|x_i)$. O modelo logístico será dado por:

$$\psi(\mathbf{x}) = \frac{e^{\xi(\mathbf{x})}}{1 + e^{\xi(\mathbf{x})}} \tag{4.5}$$

Sendo $\xi(\mathbf{x})$ a expressão

$$\xi(\mathbf{x}) = \beta_0 + \sum_{i=1}^{p} \beta_p \mathbf{x} \tag{4.6}$$

A estimação do modelo requer que se estime os valores do vetor $\mathbf{b} = (\beta_0, \beta_1, ..., \beta_p)$, em que β_0 será o intercepto, e β_p o coeficiente da *p-ésima* variável explicativa. Diferentemente do modelo de regressão linear, não podemos utilizar o método dos Mínimos Quadrados Or-

dinários (MQO) para estimar o vetor **b** segundo (HOSMER; LEMESHOW, 2000, p.8), não se pode garantir estimadores não viesados, homoscedasticidade e $E(\varepsilon) = 0$, o que inviabiliza a aplicação do MQO.

Com essa inviabilidade, para a estimação do vetor **b**, o Método de Máxima Verossimilhança é o indicado. Esse método, de forma geral, estima os $(p+1)$ parâmetros maximizando a probabilidade de se obter o evento a ser analisado, no contexto do trabalho, maximiza-se a probabilidade de se obter indivíduos inadimplentes.

Como a variável resposta Y é dicotômica, a equação 4.5 para um vetor **b** arbitrário tem como resultado a probabilidade condicional do individuo y_i ser igual a 1 dado seu vetor de características **x**, denotada por $P(y_i = 1|\mathbf{x})$. Posto desta forma, podemos concluir que $P(y_i = 0|\mathbf{x})$ será igual a $1 - \psi(x)$. A função de verossimilhança será:

$$\psi(x_i)^{y_i}[1 - \psi(x)]^{1-y_i} \tag{4.7}$$

Na equação 4.7, o termo $\psi(x_i)^{y_i}$ denota a contribuição para a função de verossimilhança do individuo i quando $y_i = 1$ e o termo $[1 - \psi(x)]^{1-y_i}$ quando $y_i = 0$. Tomando os indivíduos como independentes[3], a função de verossimilhança ϑ considerando os n indivíduos, para o i-ésimo indivíduo da amostra será:

$$\vartheta(\mathbf{b}) = \prod_{i=1}^{n} \left\{ \psi(x_i)^{y_i}[1 - \psi(x)]^{1-y_i} \right\} \tag{4.8}$$

Ou seja, ϑ é o produtorio da contribuição de todos os i indivíduos para a ocorrência ou não do evento, da inadimplência. Para efeito de simplificação, faz-se o ln de $\vartheta(\mathbf{b})$:

[3]Sem essa simplificação, seria necessário estimar a correlação entre todos os i indivíduos da amostra, o que desviaria o escopo do presente trabalho.

$$\Phi(\mathbf{b}) = ln(\vartheta(\mathbf{b})) = \sum_{i=1}^{n} \{y_i ln(\psi(x_i)) + (1-y_i) ln[1-\psi(x)]\} \tag{4.9}$$

Os parâmetros β_i que maximizam a equação 4.9 são encontrados diferenciando a função em relação aos $(p+1)$[4] $\beta's$ do vetor $\mathbf{b}$, que resultando nas seguintes equações:

$$\sum_{i=1}^{n} [y_i - \psi(x_i)] = 0 \tag{4.10}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} [y_i - \psi(x_i)] = 0 \tag{4.11}$$

para $i = 1,2...n$ e $j = 1,2...p$

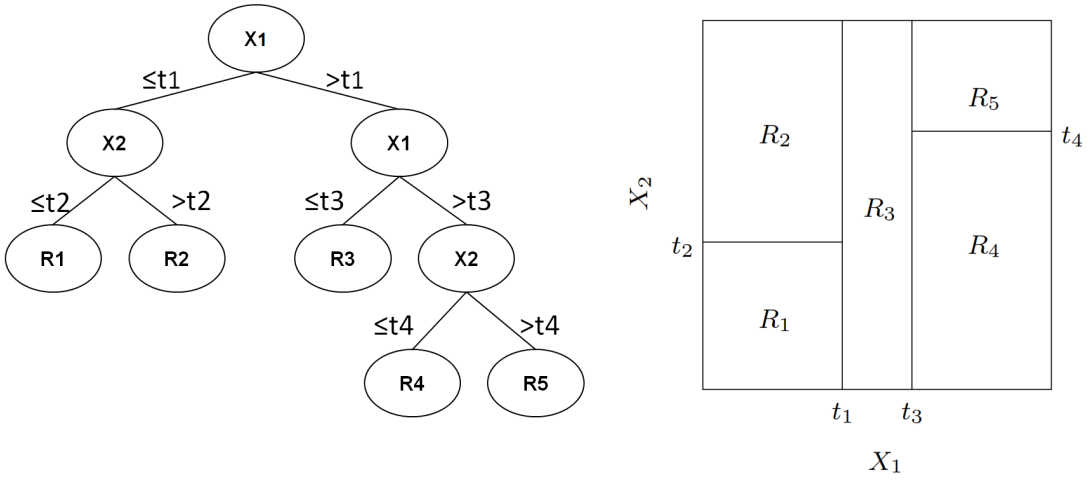
A aplicação dessa metodologia será feita no *Matlab*, através da função $glmfit$, conforme rotina descrita no Apêndice B.1. Uma vez estimados os parâmetros, é possível construir o ξ do i-ésimo individuo, conforme a equação 4.6, para que em seguida, a partir da equação 4.5 se encontre a probabilidade condicional de que o i-ésimo individuo seja inadimplente, dado seu vetor $\mathbf{x}$ de características.

4.2 Árvore de decisão

O segundo modelo a ser desenvolvido é o de Árvore de Decisão, que segundo Kantardzic (2002) é um método não paramétrico de classificação muito eficiente. O modelo é baseado em uma estrutura hierárquica de nós conectados por ramos, em que no caso das árvores de decisão binárias, como a que será desenvolvida, cada nó se subdivide em outros dois, logo, uma decisão é sempre interpretada como verdadeira ou falsa. Os últimos ramos, chamados

[4] $(p+1)$ por que estima-se além dos p coeficientes o intercepto

de folhas, são a resposta do modelo. A Figura 4.2a exemplifica essa estrutura.



(a) Árvore de decisão (b) Resposta da árvore no espaço

Figura 4.2: Exemplo de uma Árvore de Decisão e seu resultado, adaptado de (HASTIE, 2009, p. 306)

Essa estrutura pode ser melhor entendida por meio de um exemplo. Partindo de uma variável resposta continua Y e duas variáveis preditivas X_1 e X_2 pode-se construir uma árvore de decisão como a da Figura 4.2a, que segregará o plano em retângulos conforme a Figura 4.2b. A estrutura hierárquica mencionada acima consiste na apresentação dos dados inicialmente no primeiro ramo e se a variável X_1 for menor ou igual a um determinado valor t_1 segue-se para o ramo da esquerda, e se for maior do que esse valor t_1 segue para o ramo da direita, Esse processo é repetido para todos os níveis hierárquicos da árvore, até que uma folha seja alcançada.

Na Figura 4.2b estão os espaços correspondentes as folhas da árvore. Ao apresentarmos um novo conjunto de dados X_{1n} e X_{2n} a essa estrutura, ele percorrerá os ramos que satisfizerem suas condições logicas chegando assim a sua folha respectiva, que corresponderá a região a qual ele pertence. Por exemplo, se um par X_{1n} e X_{2n} for apresentado a árvore, e o valor de X_{1n} for menor do que o parâmetro t_1 e X_{1n} for menor do que t_2 esse par será classificado como

R_1, e estará na região R_1, conforme pode ser visto na Figura 4.2b.

Cada conjunto de dados apresentados a árvore será classificado em alguma das regiões e em apenas uma delas. Ou seja, A intersecção das folhas é sempre vazia e a união delas corresponde ao espaço completo.

A construção do modelo é fundada na determinação da ordem das variáveis dentro da árvore e os valores dos *t´s* que determinarão a estrutura lógica da árvore. O algoritmo a ser usado foi apresentado por Breiman et al. (1984) e é denominado CART [5], e será brevemente descrito na próxima seção.

4.2.1 Algoritmo CART

O algoritmo CART é baseado em duas características principais. A primeira delas é o fato de ser uma técnica binária, uma vez que os ramos são sempre divididos em dois ramos subsequentes. Além disso, é um algoritmo recursivo, porque cada ramo é dividido em dois ramos filhos e assim sucessivamente. Essas duas características levam a uma estrutura de regras no formato *se, então*.

O algoritmo pode ser dividido em três etapas principais:

I) **Seleção dos nós:** A ideia é partir do caso geral para o particular, ou seja, inicialmente é considerada toda a base de dados para a seleção do primeiro ramo. Em seguida, é realizado o mesmo para os dois ramos subsequentes, considerando-os de forma independente.

No primeiro momento, são considerados todas as possíveis divisões binárias de todas as variáveis preditoras com o objetivo de encontrar a divisão que eleva a pureza dos dados

[5] Do inglês, *Classification And Regression Trees.*

no nó subsequente. Pra isso e utilizado o seguinte indicador:

$$^{6}Gini(t) = \sum_{i \neq j} p(i|t)p(j|t) = 1 - \sum_{i} p(i|t)^2 \tag{4.12}$$

O termo $p(i|t)$ é a probabilidade de que a classe i esteja no nó t, e o termo $p(j|t)$ é a probabilidade de que a classe j esteja no nó t. A equação 4.12 tem valor máximo quando $p(i|t) = p(j|t)$ ou seja, quando as duas classes tem o mesmo numero de elementos. Dizendo de outra forma, o Gini máximo é alcançado quando o algoritmo é capaz de segregar o mesmo conjunto de dados em dois subconjuntos completamente distintos. O objetivo do algoritmo é maximizar esse indicador em cada um dos ramos.

II) **Crescimento da árvore:** O algoritmo acima é repetido até que não haja mais ganho com as demais repartições. O critério de parada é que o Gini do nó posterior seja menor do que do nó anterior, ou que todas as variáveis já tenham sido utilizadas.

III) **Poda:** Por ser executado repetidas vezes, até que não seja possível obter ganho com a subdivisão o algoritmo tente a construir árvores muito complexas, que se ajustam perfeitamente aos dados de treinamento, mas que não produzem resultados satisfatórios para dados que não fazem parte do conjunto de treinamento.

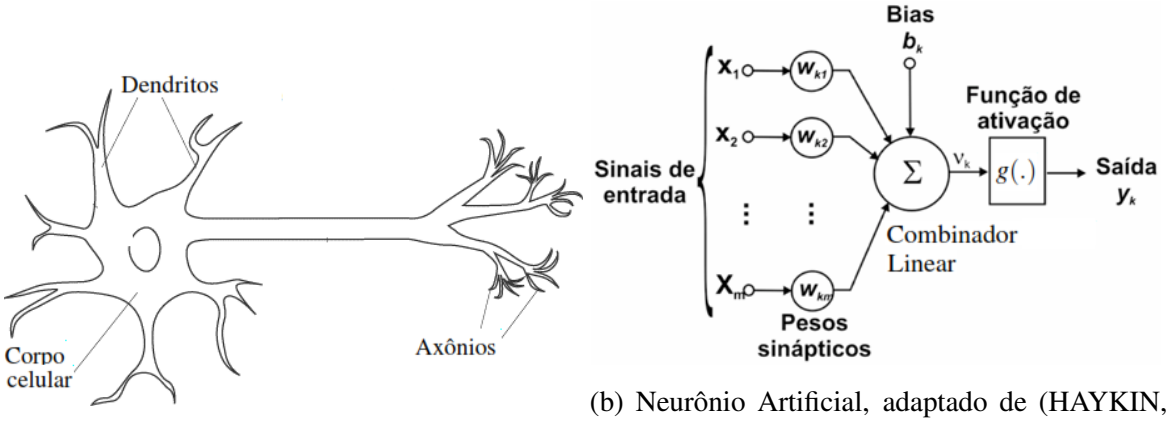
Por essa razão, quando se utiliza um menor número de nós, a árvore tende a perder um pouco a sua especificidade, o que pode vir a reduzir o desempenho nos dados de treinamento, porem, elevar o desempenho em dados fora dessa amostra. O processo de poda deve ser feito a partir do ponto em que a inclusão de novos ramos não elevar substancialmente a taxa de acertos da árvore.

[6]Esse indicador, apesar do nome, é distinto do que será construído adiante, na Seção 5.1.3. O nome foi mantido para seguir a referencia de Breiman et al. (1984, p.94)

4.3 Redes Neurais

Redes neurais são estruturas baseadas no funcionamento do cérebro dos seres vivos. Silva, Spatti e Flauzino (2010) as define como "(...) modelos computacionais inspirados no sistema nervoso dos seres vivos".

Assim como no cérebro, sua estrutura fundamental é o neurônio. De forma breve, o neurônio dos seres vivos recebe os sinais por meio dos dendritos e os encaminha ao corpo celular onde são processados e então, os resultados são encaminhados para os axônios, terminais que encaminham os sinais processados para o próximo neurônio. Essa estrutura pode ser vista na Figura 4.3a, em que está esquematizado um neurônio natural.



(a) Neurônio Natural

(b) Neurônio Artificial, adaptado de (HAYKIN, 1999)

Figura 4.3: Comparativo entre um neurônio natural e um artificial

De forma análoga, o neurônio artificial conta com essas três estruturas básicas. Os sinais são recebidos pelos terminais de entrada, na Figura 4.3b representados por $x_1, x_2 \cdots x_m$. O que no neurônio natural foi chamado de processamento, no neurônio artificial será decomposto em algumas etapas. Uma vez que os sinais são recebidos, eles são ponderado por valores $w_{k1}, w_{k2} \cdots w_{km}$ chamados de pesos sinápticos e são encaminhados para o combinador linear, que agrega todos os sinais de entrada ponderados. O Bias b_k define o valor mínimo para que

o combinador linear tenha uma saída.

O resultado proveniente do combinador linear, V_k, é então transformado por uma função de ativação $g(.)$ produzindo a saída y_k. De forma mais rigorosa, o resultado do neurônio após todas as etapas de processamento será:

$$V_k = \sum_{i=1}^{m} x_i w_i - b_k \tag{4.13}$$

$$y_k = g(V_k) \tag{4.14}$$

O neurônio até aqui analisado é a forma mais simples de uma rede com apenas um neurônio. Essa rede é conhecida como *Perceptron*[7] e é a forma mais elementar de rede neural.

A combinação de diferentes *Perceptrons* da origem a diversas formas de redes. Um exemplo pode ser visto na Figura 4.4, em que uma camada oculta formada por quatro neurônios recebe cada um deles todos os sinais de entrada provenientes da camada de entrada. O resultado desses neurônios é então encaminhado para uma próxima camada formada por dois neurônios, chamada de camada de saída. Nela cada um dos dois neurônios tem como entrada o resultado dos quatro neurônios das camadas anteriores e produzem uma saída cada um deles. Essa estrutura é chamada de *Perceptron de Múltiplas Camadas* (PMC).

A grande inovação das redes neurais é sua capacidade de aprendizado. Haykin (1999, p. 72) define o aprendizado como o processo em que os parâmetros livres (pesos sinápticos principalmente) são adaptados através de um processo de estimação através do ambiente em

[7]Um histórico detalhado do *Perceptron* e do desenvolvimento das Redes Neurais pode ser encontrado em (HAYKIN, 1999, p. 60)

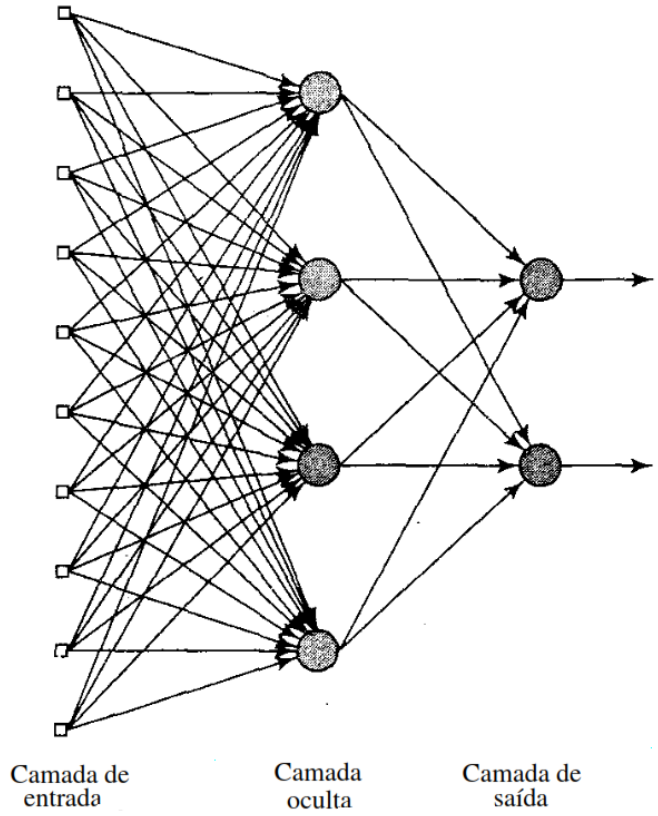


Figura 4.4: Representação de um rede neural, (HAYKIN, 1999)

que a rede está inserida. Uma vez que a rede é adaptada, ela passa a produzir resultados de acordo com essa nova configuração em detrimento das anteriores, ou seja, as redes neurais são capazes de se adaptar a mudanças nos padrões dos dados de entrada.

Dada a flexibilidade com que se podem configurar redes, exclusivas para as mais diferentes aplicações, foram desenvolvidos diversos algoritmos de aprendizado. Neste trabalho usaremos redes neurais com estrutura PMC e o algoritmo de aprendizado será o *Backpropagation*, analisado na próxima seção.

4.3.1 Algoritmo de aprendizado

Para descrevermos o funcionamento do algoritmo de aprendizado *backpropagation* usaremos de exemplo a rede da Figura 4.5, uma PMC com duas camadas intermediárias. A primeira com n_1 neurônios e a segunda com n_2 neurônios. Todas as x_n entradas estão conectadas a todos os n_1 neurônios da primeira camada, que por sua vez tem suas saídas conectadas a todos os n_2 neurônios da segunda camada. A ultima camada, de saída, é composta por n_3 neurônios.

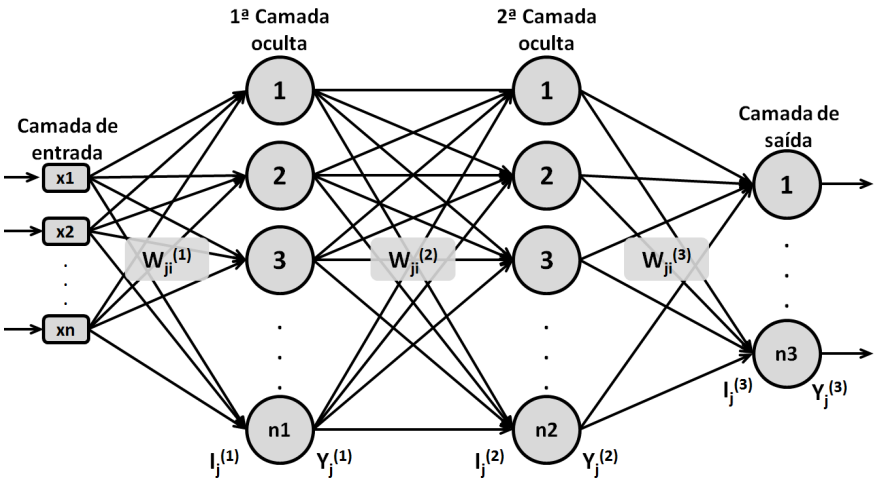


Figura 4.5: Exemplo de rede *Perceptron* multi camada. Adaptado de (SILVA; SPATTI; FLAUZINO, 2010)

Após entrarem na rede, pela camada de entrada, os i valores são ponderados pelos pesos sinápticos w de cada entrada dos j neurônios dessa primeira camada. Os pesos sinápticos no inicio do treinamento são valores aleatórios, sem qualquer relação com o formato da rede, os dados de entrada ou qualquer outra premissa. Dessa forma, temos que $W_{ij}^{(1)}$ é a matriz dos pesos sinápticos do j-ésimo neurônio da primeira camada conectada ao i-ésimo valor da camada de entrada. De forma análoga podemos definir $W_{ij}^{(2)}$ e $W_{ij}^{(3)}$, em que o primeiro é a matriz dos pesos sinápticos do j-ésimo neurônio da segunda camada ligada ao i-ésimo valor

de saída da camada anterior.

Podemos então, definir a entrada do j-ésimo neurônio da primeira camada como:

$$I_j^{(1)} = \sum_{i=1}^{n} W_{ij}^{(1)} x_i \tag{4.15}$$

Da mesma forma, a entrada do j-ésimo neurônio da segunda e da terceira camada serão, respectivamente:

$$I_j^{(2)} = \sum_{i=1}^{n1} W_{ij}^{(2)} Y_i^1 \tag{4.16}$$

$$I_j^{(3)} = \sum_{i=1}^{n2} W_{ij}^{(3)} Y_i^2 \tag{4.17}$$

Após passarem pela função de ativação característica de cada neurônio, os $I_j^{(1)}$ são convertidos em $Y_j^{(1)}$. Em outras palavras, no caso da primeira camada, $Y_j^{(1)}$ é a saída do j-ésimo neurônio, que fez uma combinação linear com as entradas $I_j^{(1)}$ e os transformou dada sua função de ativação $g(.)$. Então $Y_j^{(1)}$, $Y_j^{(2)}$ e $Y_j^{(3)}$ são, respectivamente:

$$Y_j^{(1)} = g(I_j^{(1)}) \tag{4.18}$$

$$Y_j^{(2)} = g(I_j^{(2)}) \tag{4.19}$$

$$Y_j^{(3)} = g(I_j^{(3)}) \tag{4.20}$$

Dessa forma, uma vez definidos os pesos sinápticos w, ao apresentarmos um vetor de entradas x, teremos $n3$ saídas. O algoritmo *backpropagation* é empregado para encontrar os pesos sinápticos, minimizando uma função de erro a ser definida.

Como usaremos o *Matlab* para modelar a rede, por meio do pacote de redes neurais[8], convêm analisarmos o erro quadrático médio, erro que o algoritmo tentará minimizar, conforme Beale Martin T. Hagan (2010, sec. 10-22). Seja k a k-ésima amostra de treinamento de um conjunto de treinamento p a entrar na rede neural acima apresentada, o desvio $E(k)$ pode ser definido como:

$$E(k) = \frac{1}{2}\left(\sum_{j=1}^{n_3}\left(d_j(k) - Y_j^3(k)\right)^2\right) \tag{4.21}$$

Em que $d_j(k)$ é o valor esperado dada a k-ésima amostra e Y_j^3 o resultado produzido pelo j-ésimo neurônio. Assim, o erro quadrático médio (Eqm)será:

$$Eqm = \frac{1}{p}\left(\sum_{k=1}^{p} E(k)\right) \tag{4.22}$$

O processo de aprendizagem é iterativo. Apresenta-se as k amostras de treinamento de p e calcula-se o Eqm. Os pesos sinápticos da matriz $W_{ij}^{(3)}$ são ajustados por meio do gradiente descendente[9].

Como as camadas intermediarias não visualizam diretamente o valor esperado, seus pesos (as matrizes $W_{ij}^{(1)}$ e $W_{ij}^{(2)}$) são ajustados somente após o ajuste dos pesos camada de saída. De forma objetiva, uma vez ajustados os pesos da camada de saída, o desvio do resultado produzido em relação ao valor esperado é retro propagado para os neurônios das camadas an-

[8] *Neural Network Toolbox 7*

[9] Uma completa explicação desse processo pode ser encontrada em (SILVA; SPATTI; FLAUZINO, 2010, p.101) e (HAYKIN, 1999, p.183).

teriores, ou seja, a resposta desejada dos neurônios das camadas intermediarias é diretamente determinado em função dos neurônios da camada imediatamente posterior e que já foram previamente ajustados. Esse processo é repedido quantas vezes forem necessárias, até que se atinja um fator de parada pré-determinado, que pode ser um número máximo de iterações, ou um erro mínimo desejado.

5 *Avaliação dos Modelos*

A seguir, descreveremos algumas métricas indicadas pela literatura[1] para avaliação de modelos de *credit score*. O detalhamento dos modelos implementados será feito na seção 5.2, bem como a análise dos resultados encontrados.

5.1 Medidas de desempenho

5.1.1 Índice KS

O índice de *Kolmogorov-Smirnov* mensura a máxima distância entre as curvas acumuladas da distribuição de probabilidade de adimplentes e inadimplentes. Seja P_{iM} a probabilidade do i-ésimo indivíduo da amostra ser inadimplente, dado seu vetor de características $\mathbf{x}_i$, conforme definido na Seção 4.1. Assim P_{iM} é definido por:

$$P_{iM} = P(Y_i = 1|\mathbf{x}_i) = \psi(\mathbf{x}_i) \tag{5.1}$$

Onde $\psi(\mathbf{x}_i)$ será o resultado dos modelos a serem construídos. De forma análoga, a probabilidade P_{iB} do mesmo i-ésimo individuo da amostra não ser inadimplente dado o mesmo vetor de características $\mathbf{x}_i$ é dado por:

[1] (THOMAS, 2009), (THOMAS, 2002), (SICSÚ, 2010)

$$P_{iB} = P(Y_i = 0|\mathbf{x}_i) = 1 - \psi(\mathbf{x}_i) \tag{5.2}$$

Se discretizarmos o domínio de $\psi(\mathbf{x})$ em n faixas de tamanho δ e contarmos o número de indivíduos adimplentes e inadimplentes dentro de cada faixa teremos um diagrama semelhante à Figura 5.1. Em outras palavras, se o i-ésimo individuo é inadimplente e tem probabilidade $P_{iM} = \psi(\mathbf{x}_i)$ ele estará na faixa $\psi(\mathbf{x}_i) - \delta < \psi(\mathbf{x}_i) \leqslant \psi(\mathbf{x}_i) + \delta$.

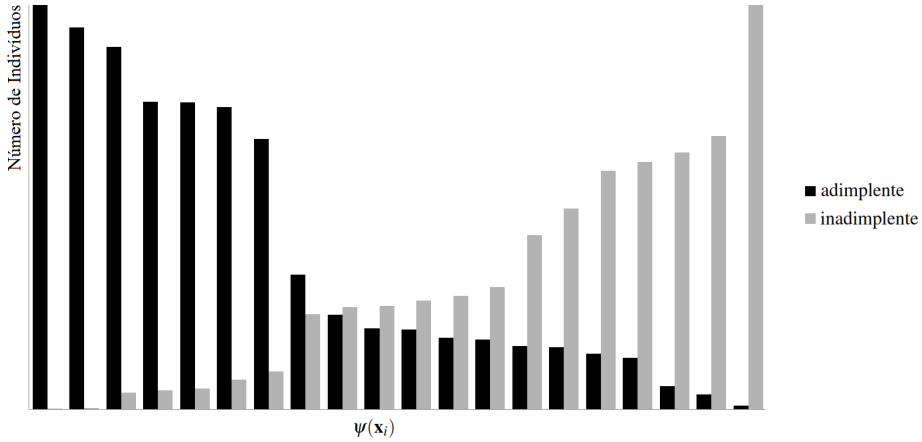


Figura 5.1: Distribuição de adimplentes e inadimplentes em faixas de $\psi(\mathbf{x}_i)$

Espera-se que a concentração nas faixas (eixo das abscissas) de menor valor de $\psi(\mathbf{x})$ contenham um número maior de adimplentes. Por outro lado, espera-se que para os valores mais altos de $\psi(\mathbf{x})$ a frequência de indivíduos inadimplentes seja maior.

Para que seja possível a comparação entre amostras de diferentes tamanhos, é necessário trabalhar com as frequências relativas acumuladas dentro da amostra. Assim, as frequências relativas acumuladas são ilustradas na Figura 5.2.

Fazendo $\delta \to \infty$ e definindo $M(\mathbf{x})$ como a função que descreve a frequência relativa acumulada de inadimplentes e $B(\mathbf{x})$ a função da frequência relativa acumulada de adimplentes, o KS é dado pela distancia máxima entre $M(\mathbf{x})$ e $B(\mathbf{x})$. Matematicamente, tem-se que:

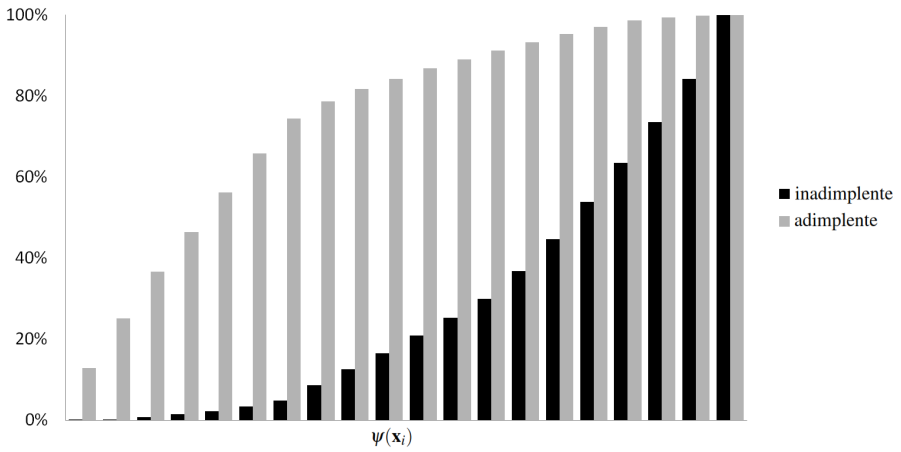


Figura 5.2: Distribuição acumulada de adimplentes e inadimplentes em faixas de $\psi(\mathbf{x}_i)$

$$KS = max|B(\psi(x_i)) - M(\psi(x_i))|() \tag{5.3}$$

A Figura 5.3 é uma representação da equação 5.3. Em um modelo ideal, a equação tem KS igual a 1 se:

$$B(\mathbf{x}) = 1 \mid \psi(x_i) = 0$$

$$M(\mathbf{x}) = 1 \mid \psi(x_i) = 1$$

Nestas condições, o modelo segrega a população inicial em duas categorias distintas, separando os inadimplentes (curva superior) dos adimplentes(curva inferior). Portanto, valores de KS quando tendem a 1 demonstram que o modelo em questão é capaz de discriminar as duas sub populações (adimplentes e inadimplentes no caso) da população inicial.

Dizendo de outra maneira, para um determinado valor de $\psi(\mathbf{x}) = \gamma$ no intervalo (0,1), considera-se como adimplentes os indivíduos com $\psi(x_i) \leq \gamma$ e inadimplentes os com $\psi(x_i) >$

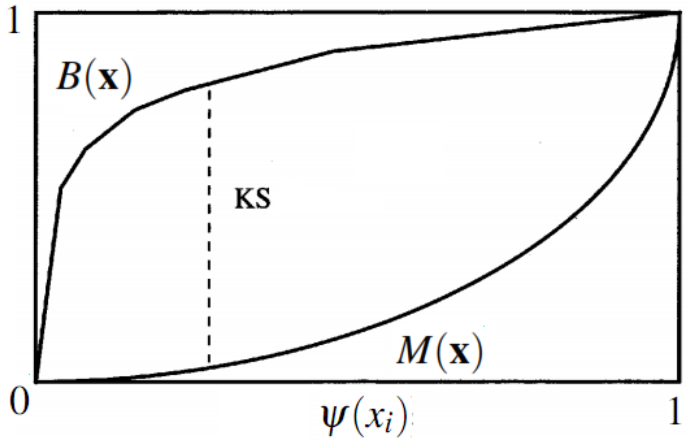


Figura 5.3: Diagrama da equação 5.3, adaptado de (THOMAS, 2002)

γ. O valor de $B(\mathbf{x})$ da Figura 5.3 neste ponto será a razão entre o número de adimplentes e o total de indivíduos e o valor de $M(\mathbf{x})$ será a razão entre o número de inadimplentes e o total de indivíduos. O KS será, para um determinado valor de $\psi(\mathbf{x}) = \gamma$, a maior distancia entre as curvas.

Uma vez que o $\psi(x_i)$ desenvolvido no presente trabalho é continuo[2] no intervalo $[0,1]$, este trabalho utilizará um $\delta = 0,01$.

5.1.2 Curva ROC

Ao contrario do *KS* que é construído com a distribuição relativa acumulada de adimplentes e inadimplentes dentro de faixas de $\psi(x_i)$ independentemente, a curva ROC, (do inglês *Receiver Operating Characteristic*) (ZWEIG, 1993), é baseada na distribuição de probabilidade acumulada condicional entre adimplentes e inadimplentes.

A curva ROC, segundo Thomas (2009, p. 115) baseia-se em dois conceitos. O primeiro deles, sensitividade é a proporção de indivíduos adimplentes classificados de forma correta,

[2] Além desse motivo, a criação de faixas suaviza as curvas $B(\mathbf{x})$ e $M(\mathbf{x})$, uma vez que o conjunto de dados é relativamente pequeno.

ou seja, a proporção de adimplentes classificados dessa forma pelo modelo considerando-se uma faixa de $\psi(x_i)$ com tamanho δ. Por outro lado, o conceito de especificidade é a proporção de clientes inadimplentes classificados corretamente, considerando-se também a mesma faixa de tamanho δ de $\psi(x_i)$.

Esse conceitos podem ser exemplificado pela Figura 5.4. Para um determinado $\psi(\mathbf{x}) = \gamma$, com γ no intervalo (0,1), os indivíduos com $\psi(x_i) \leq \gamma$ são denominados adimplentes e os com $\psi(x_i) > \gamma$ são considerados inadimplentes. Ao compararmos os indivíduos considerados inadimplentes e adimplentes pelo modelo, dado um $\psi(\mathbf{x}) = \gamma$ com os valores reais observados, é possível a construção de uma matriz como a da Figura 5.4:

		Modelo	
		Inadimplente	Adimplente
Observado	Inadimplente	VI	FA
	Adimplente	FI	VA

Figura 5.4: Matriz de Confusão

Essa matriz, chamada de Matriz de Confusão é composta por quatro valores:

a) VI: Verdadeiros Inadimplentes, são os indivíduos que foram observados como inadimplentes e o modelo corretamente o classificou como inadimplente;

b) FA: Falsos Adimplentes, indivíduos que foram classificados como adimplentes de forma equivocada pelo modelo, uma vez que são na realidade inadimplentes;

c) FI: Falsos Inadimplentes, indivíduos que são adimplentes, porem o modelo os classificou como inadimplentes;

d) VA: Verdadeiros Adimplentes, são os indivíduos classificados corretamente pelo modelo como adimplentes.

Conforme definido acima, a especificidade e a sensibilidade serão:

$$Especificidade = \frac{VI}{(VI + FA)} \tag{5.4}$$

$$Sensibilidade = \frac{VA}{(VA + FI)} \tag{5.5}$$

Na Figura 5.5, a curva ROC é a curva continua. Cada ponto da curva corresponde a um *score* $\psi(x)$, em que é calculada a Sensibilidade e o complemento da Especificidade ((1-Especificidade), que corresponde ao número de adimplentes classificados corretamente).

Novamente faremos a discretização de $\psi(x)$ em intervalos com amplitude $\delta = 0,01$, conforme definido na seção 5.1.1.

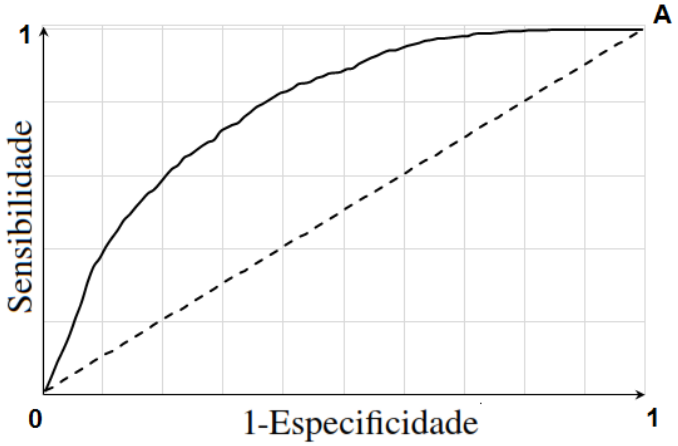


Figura 5.5: Curva *ROC*

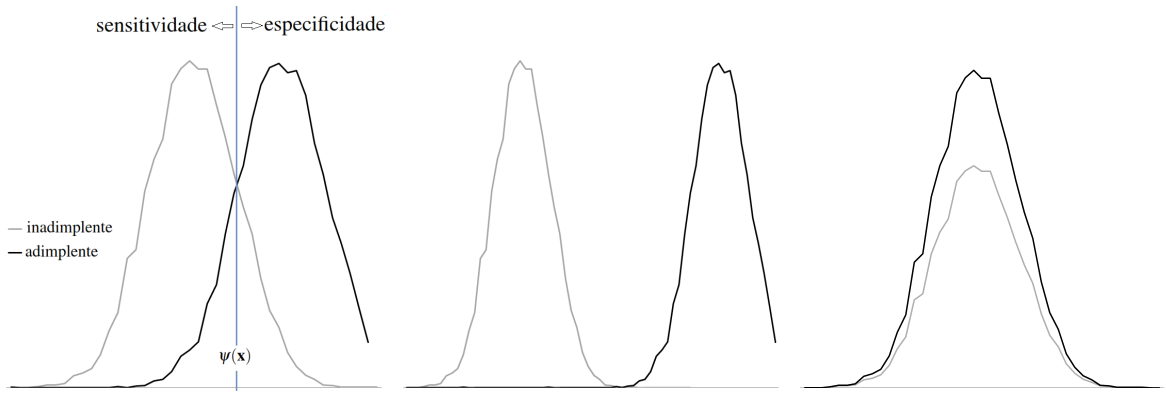
Se o modelo for capaz de discriminar as populações de adimplentes e inadimplentes com total precisão, dada a população inicial, o modelo é capaz de subdividi-la em dois grupos distintos, que apresentariam distribuições conforme a Figura 5.6b. Nela há um determinado $\psi(x)$ que divide as duas populações com exatidão, e nesse caso, a curva ROC será dois segmentos de reta, um indo do 0 ao 1 no eixo das ordenadas da Figura 5.5 e iria de 1 ao ponto

A.

No outro extremo, se o modelo for incapaz de segregar as populações, teremos uma distribuição conforme a Figura 5.6c, em que para determinado valor de $\psi(x)$ não há qualquer discriminação entre adimplentes e inadimplente. Se assim fosse, a curva ROC seria o segmento de reta pontilhado da Figura 5.5.

Uma vez traçada a curva ROC, mensura-se a área entre a curva e o segmento de reta pontilhado da Figura 5.5, conforme a equação 5.6. Chamaremos essa área de *AUROC*, do inglês *Area Under Receiver Operating Characteristic*.

$$AUROC = \int (ROC) - 0,5 \tag{5.6}$$



(a) $0 < AUROC < 0,5$ (b) $AUROC = 0,5$ (c) $AUROC = 0$

Figura 5.6: Possiveis valores para a *AUROC*

Olhando novamente para o cenário em que o modelo é capaz de separar as duas amostras com precisão, Figura 5.6b, teremos AUROC igual a 0,5. No outro extremo, em que o modelo não faz qualquer discrimina, temos AUROC igual a 0. Portanto, esperamos um valor para a AUROC entre 0 e 0,5, uma vez que os modelos discriminam as populações conforme o exemplo da Figura 5.6a.

5.1.3 Índice de Gini

O coeficiente de Gini é diretamente derivado da curva ROC. Conforme demonstrado na Figura 5.5, a área entre a curva ROC e a curva pontilhada diagonal (AUROC) vária entre 0 e 0,5. O Índice de Gini é o dobro da área entre a curva ROC e a linha diagonal pontilhada.

$$GINI = 2AUROC \tag{5.7}$$

$$GINI = 2\left(\int (ROC) - 0,5\right) \tag{5.8}$$

Assim como o KS, o Gini está no intervalo $[0,1]$, facilitando a comparação dos indicadores. Ambos serão considerados na avaliação dos modelos a serem construídos. Enquanto o KS mensura a maior distancia entre as duas subpopulações para um determinado *score* $\psi(x)$, o índice de Gini considera toda a distribuição. Por serem complementares, é importante considerar as duas métricas.

5.2 Resultados

Conforme explicitado no Capítulo 3, os dados foram manipulados utiliando-se o *softwares* SAS[3] versão 9.2. Foi feita apenas uma renomeação das observações nas variáveis categóricas com o intuito de simplificar a análise, preservando-se a integridade da base de dados, conforme a rotina do Apêndice A.1. Além disso, foi realizada a seleção de variáveis conforme descrita na sessão 3.1, cuja rotina se encontra no Apêndice A.2.

Apesar do *SAS* ter implementada uma rotina para a estimação dos coeficientes da Re-

[3]SAS é marca Registrada da SAS Institute INC, USA

gressão Logística, a versão utilizada não dispõe de rotinas para o desenvolvimento da Árvore de Decisão e da Rede Neural. Por essa razão, foi utilizado o Matlab[4] versão 7.14.0.739 (R2012a x64) para o desenvolvimento dos três modelos. Foram utilizadas as rotinas já implementadas, a fim de reduzir o esforço, mas tomando-se o devido cuidado com as opções a serem selecionadas.

As rotinas foram executadas em um PC com sistema operacional Windows7 *64bits*, processador AMD Phenom II x6 1090T e 16gb de RAM.

5.2.1 Regressão Logística

O modelo de Regressão Logística foi implementado conforme a rotina do Apêndice B.1, partindo do referencial teórico apresentado na Seção 4.1. Os coeficientes estimados encontram-se na Tabela 5.1.

A baixa significância de alguns coeficientes, *p-valor*[5] tendendo a 1, é decorrente, do reduzido número de observações na categoria em questão. Como a base de dados contem uma quantidade relativamente pequena de observações, algumas categorias acabam tendo um número muito reduzido de observações, o que acarreta na estimação de parâmetros de baixa significância.

Essas categorias poderiam ser agrupadas para que aumentasse o número de observações, o que poderia elevar a significância dos coeficientes. Esse procedimento não foi utilizado porque buscamos manter a base de dados em seu formato original, sem qualquer transformação que vise elevar o desempenho de um modelo de forma isolada. Se os tratamentos fossem realizados, conforme dito na Seção 3.3, teriam que ser replicados em todos os outros mode-

[4]Matlab é marca Registrada da The MathWorks, Inc., USA

[5]Para um dado nível de significância α, rejeita-se a hipótese nula para o coeficiente em questão se o *p-valor* for maior do que o α. Maiores detalhes podem ser encontrados em (HOSMER; LEMESHOW, 2000, p.200).

Tabela 5.1: Coeficientes do Modelo de Regressão Logística

Variável	Coeficiente	*p-valor*
Const.	-5,7622	0,0000
VAR2	0,0355	0,0008
VAR5	0,0001	0,2248
VAR8	0,2142	0,0276
VAR13	-0,0202	0,0331
VAR1_0	2,0656	0,0000
VAR1_1	1,3832	0,0000
VAR1_2	1,1692	0,0051
VAR3_0	1,6843	0,0021
VAR3_1	0,4285	0,3906
VAR3_2	0,4120	0,0873
VAR3_3	0,4782	0,1891
VAR4_0	0,7562	0,0427
VAR4_1	-0,7680	0,0955
VAR4_2	0,1721	0,6582
VAR4_3	-0,2135	0,5665
VAR4_4	-0,1536	0,8773
VAR4_5	0,3351	0,6850
VAR4_6	0,8775	0,0769
VAR4_8	-0,4033	0,7539
VAR10_0	0,5717	0,2170
VAR10_1	0,7079	0,2755
VAR14_0	0,5312	0,0416
VAR14_1	0,9019	0,0323
VAR19_0	0,1844	0,3834
VAR20_0	1,5928	0,0219

los, procedimento que poderia acarretar tanto uma melhora como a piora do desempenho nos outros modelos.

O índice KS do modelo foi de 0,454 para os dados de treinamento e de 0,465 para os dados de validação. Esses números foram obtidos das curvas da Figura 5.7. As curvas ROC para os dados de treinamento e validação estão representados na Figura 5.8, gerando um Índice de Gini de 0,625 (treinamento) e 0,577 (validação).

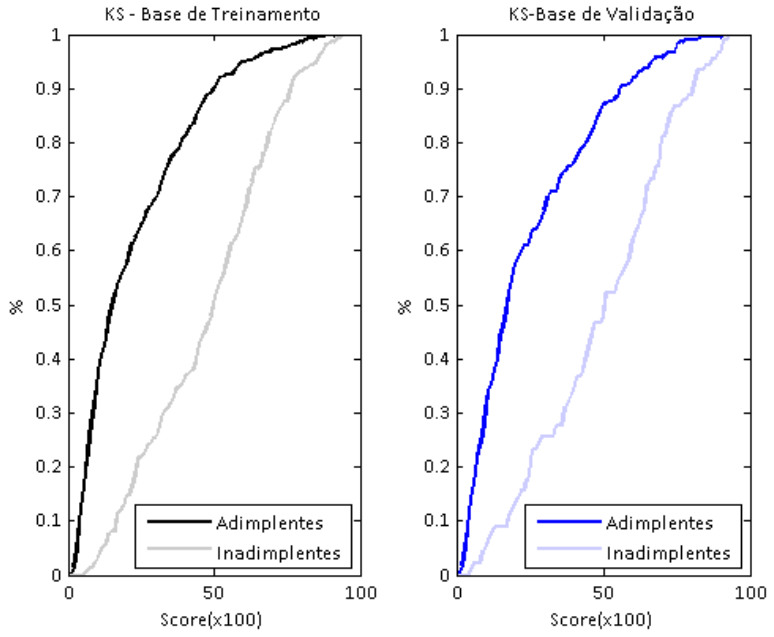


Figura 5.7: Índice KS para modelo de Regressão Logística

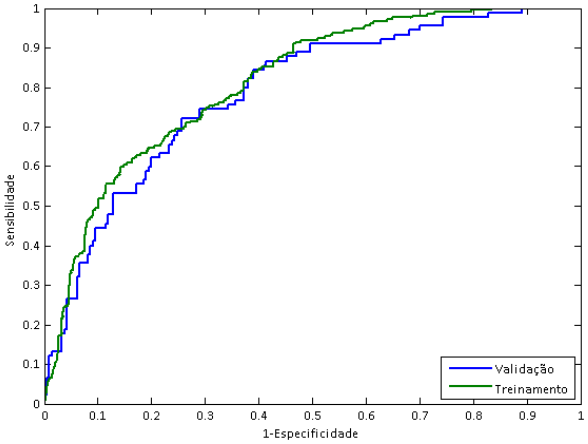


Figura 5.8: Curva ROC para modelo de regressão Logística

A rotina que gera o modelo de regressão logística está no descrita no Apêndice B.1.

5.2.2 Árvore de Decisão

O segundo modelo a ser implementado foi o modelo de Árvore de Decisão. Para este modelo as variáveis categóricas puderam ser utilizadas, uma vez que o algoritmo é capaz de considera-las, sem a necessidade de serem transformadas em *dummy*.

Conforme descrito na seção 4.2, a árvore a ser construída será um conjunto de regras no formato *se-então*, capazes de classificar qualquer novo individuo dado seu vetor de características, que o levará a uma determinada folha da árvore, atribuindo-lhe assim, um determinado *score* $\psi(x_i)$, que corresponderá a razão do número de inadimplentes em relação ao total de indivíduos que foram classificados na respectiva folha. Ou seja, se no conjunto de treinamento, uma determinada folha continha apenas inadimplentes, o *score* para os indivíduos sera 1. O inverso acontecendo para uma determinada folha em que hajam apenas adimplentes.

A árvore é desenvolvida com base na Equação 4.12. A árvore inicial contem todas as variáveis da base de dados, formando uma estrutura de 45 ramos, conforme a Figura 5.9. Apesar de ser uma estrutura complexa, e aparentemente capaz de segregar de forma correta qualquer novo indivíduo, ela apresenta o problema de *super-estimação*. Este problema é identificado quando um modelo tem um excelente desempenho em sua base de treinamento, entretanto, não tem um bom desempenho na classificação de novos dados. Isso está representado na Figura 5.10, no ponto inicial (45 ramos), onde para a base de treinamento há tanto um Índice de KS quanto um Índice de Gini elevados enquanto que para a base de dados de validação esses mesmos indicadores são muito menores.

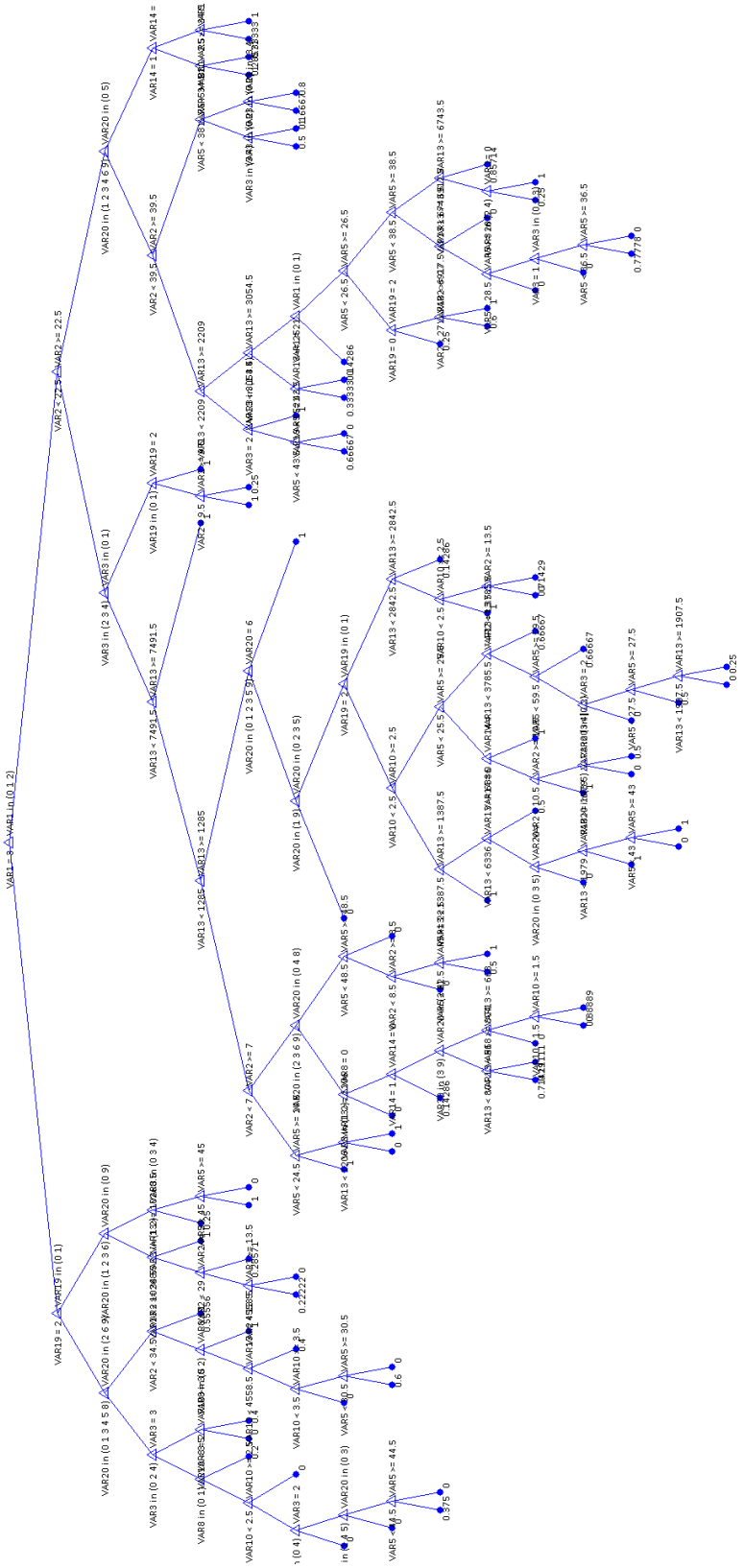
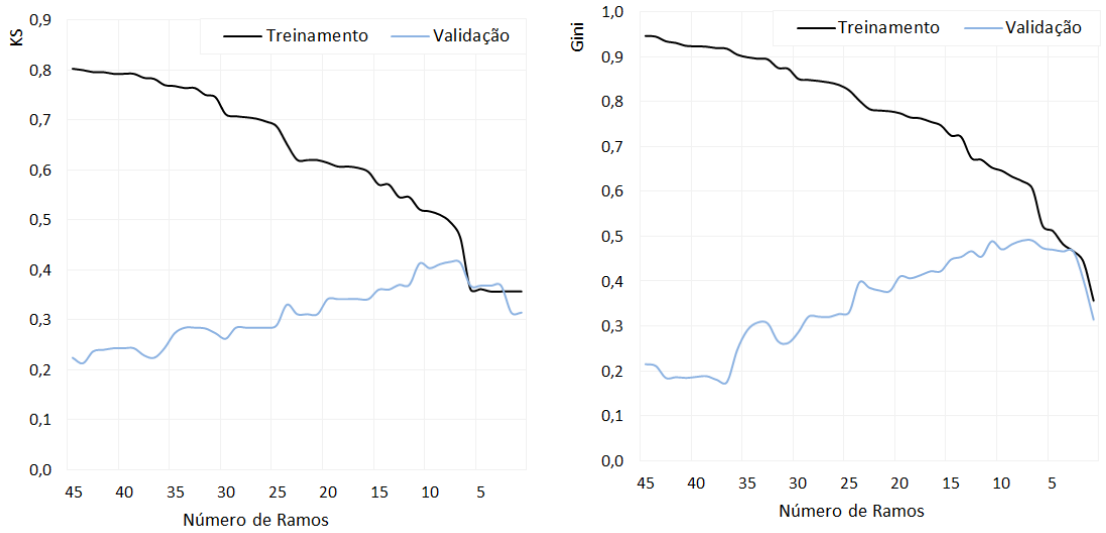


Figura 5.9: Árvore de decisão inicial

Espera-se que um modelo não apresente muitas diferenças de resultados entre os dados de treinamento e de validação, porque dessa forma seus resultados podem ser considerados robustos para conjuntos de dados não utilizados em seu desenvolvimento. Sendo assim, é necessária a execução do procedimento de poda[6] na Árvore de Decisão da Figura 5.9, com o objetivo de reduzir a discrepância entre o desempenho nos dois conjuntos de dados.

De acordo com a Figura 5.10, se o objetivo for minimizar a discrepância, deveríamos considerar a árvore com apenas 5 ramos, que é o ponto em que os Índices KS e Gini apresentam os valores mais próximos. Porém, ao considerarmos uma árvore com apenas 5 ramos, estamos considerando apenas um conjunto restrito de variáveis, que podem não ser suficientes para classificar indivíduos heterogêneos, distintos das bases de dados de treinamento e de validação. Em outras palavras, pode-se ter um resultado não robusto, uma vez que serão consideradas apenas algumas poucas variáveis desses indivíduos.



(a) KS (b) Gini

Figura 5.10: Índices KS e Gini para diferentes tamanhos de Árvore de Decisão

[6]Conforme apresentado na Seção 4.2.1

Desta forma, foi escolhida uma árvore com 17 ramos, cuja estrutura se encontra na Figura 5.11. Este modelo apresentou um Índice KS de 0,604 para os dados de treinamento e de 0,341 para os dados de validação conforme os gráficos da Figura 5.12. O Índice de Gini ficou em 0,755 e 0,422 para os dados de treinamento e validação respectivamente, oriundos das curvas ROC da Figura 5.13.

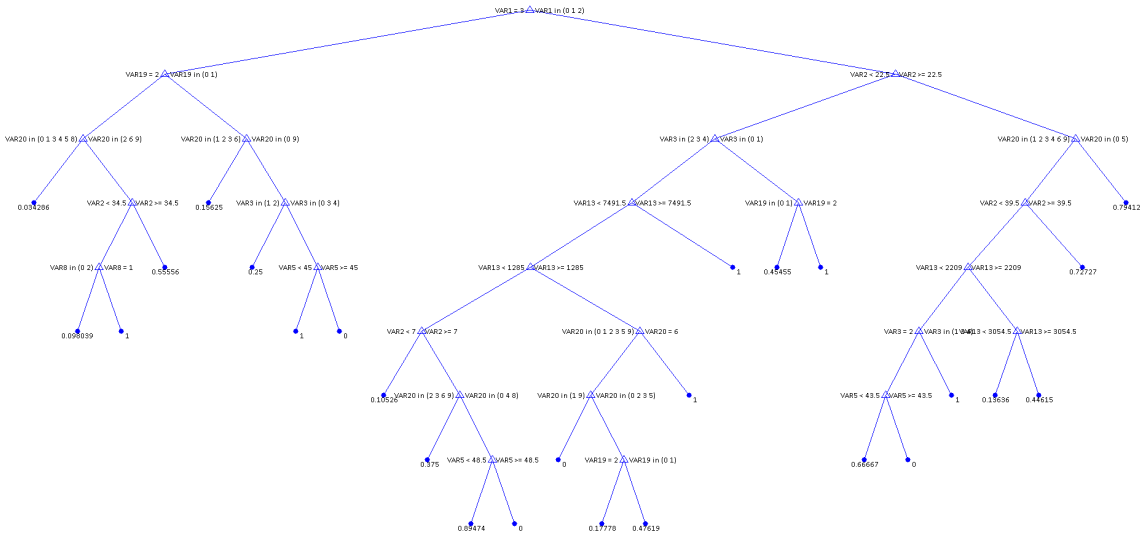


Figura 5.11: Árvore de Decisão podada, contendo 17 ramos.

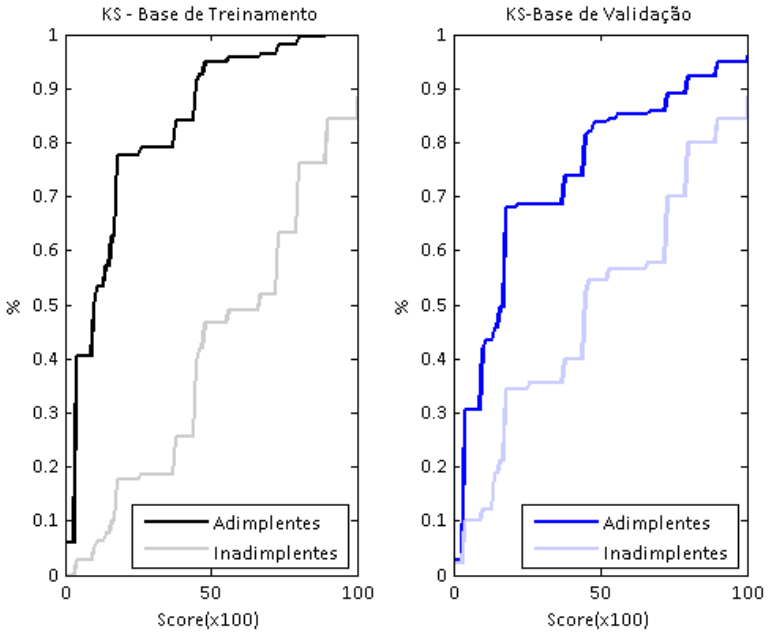


Figura 5.12: Indice KS para modelo de Árvore de Decisão

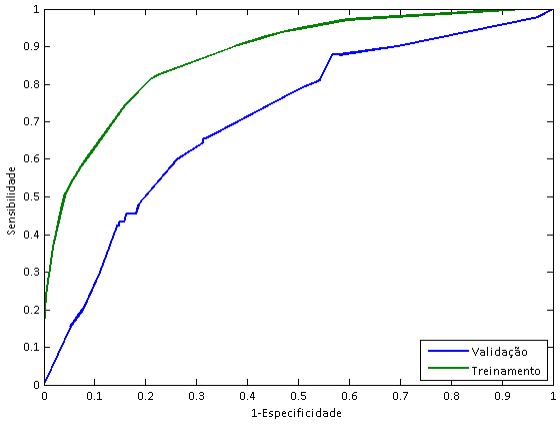


Figura 5.13: Curva ROC para modelo de Árvore de Decisão

A rotina que implementa o modelo de Árvore de Decisão encontra-se no Apêndice B.2.

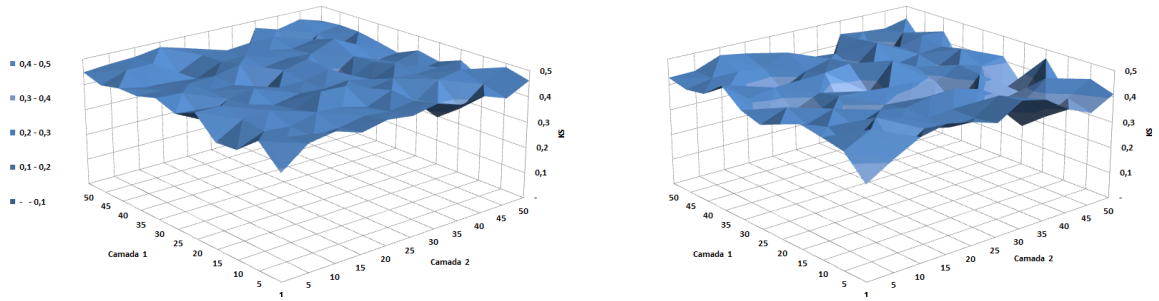
5.2.3 Rede Neural

O último modelo proposto, Rede Neural, foi implementado seguindo o referencial teórico da Seção 4.3. Foram utilizadas as variáveis categóricas transformadas em *dummy´s*. A rede proposta foi composta por 3 camadas. A última camada conterá apenas um neurônio, com a função de ativação sendo a junção logarítmica[7], e para as camadas 1 e 2, foi considerada uma função linear simples. O resultado da rede neural será, dada a função de ativação da ultima camada, um valor entre 0 e 1, ou seja, o *score* do individuo cujo vetor de variáveis foi apresentado na entrada da rede.

Para as duas primeiras camadas, o número de neurônios foi determinado por meio do cálculo dos Índices KS e Gini para diferentes configurações de rede, com um mínimo de 1 neurônio na camada e máximo de 50.

A Figura 5.14 sintetiza os resultados para o Índice KS. O índice se manteve entre 0,4 e 0,5 para os dados de treinamento (Figura 5.14a) independente do número de neurônios, tanto na primeira quanto na segunda camada. Quando olhamos para os dados de validação, Figura 5.14b, verifica-se que apesar de uma maior heterogeneidade dos resultados, eles se mantiveram próximo de 0,4. Ou seja, não se pode concluir que o aumento do numero de neurônios é capaz de promover uma melhora de desempenho.

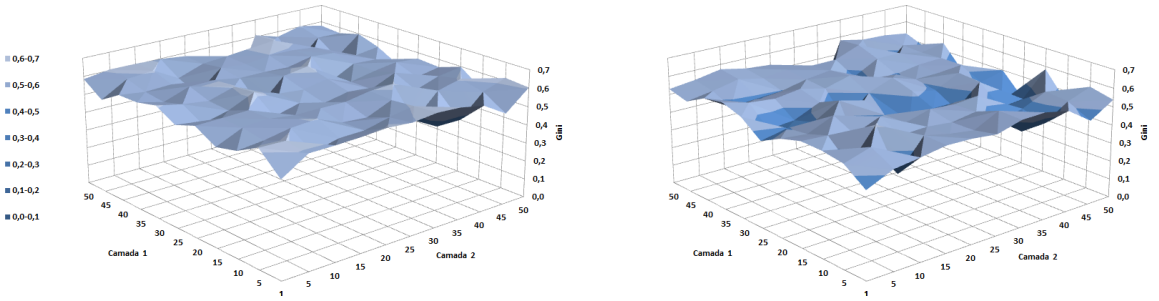
[7] Essa função foi escolhida por ter um domínio [0,1].



(a) Treinamento (b) Validação

Figura 5.14: Índice KS para diferentes configurações de Redes Neurais

A mesma conclusão pode ser tirada quando olhamos para o Índice de Gini. Tanto para os dados de Treinamento quanto de Validação, não há uma melhora de performance com relação ao numero de neurônios. A Figura 5.15 sintetiza os resultados para o Índice de Gini.



(a) Treinamento (b) Validação

Figura 5.15: Índice de Gini para diferentes configurações de Redes Neurais

Considerando os dois Índices, a melhor configuração é uma rede com 25 neurônios na primeira camada e 10 na segunda (e 1 na terceira camada, como dito anteriormente). Está estrutura será a utilizada.

A rede implementada foi capaz de produzir um Índice KS de 0,468 e 0,407 para os dados de treinamento e validação respectivamente, conforme a Figura 5.16. Em relação ao Índice de Gini, foi 0,606 para os dados de treinamento e 0,528 para os de validação, Figura 5.17. Ela foi implementada de acordo com a rotina do Apêndice B.3.

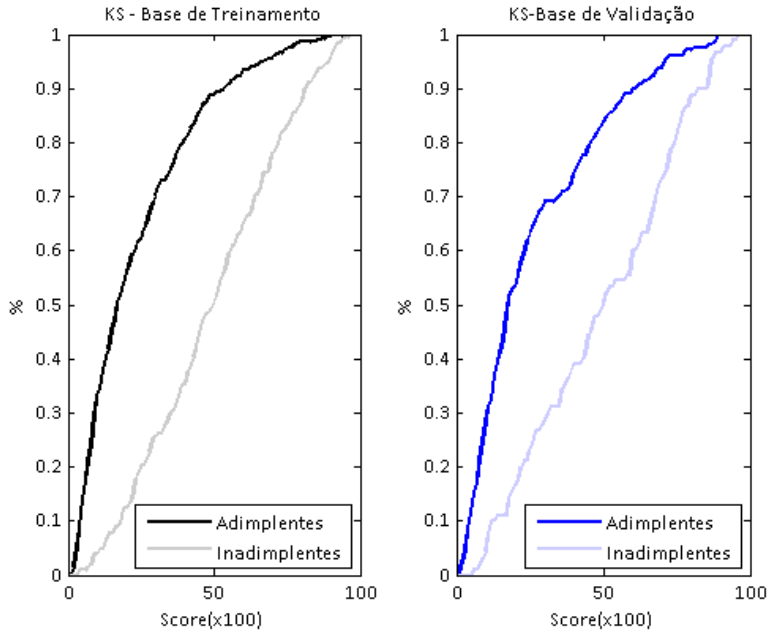


Figura 5.16: Indice KS para Rede Neural

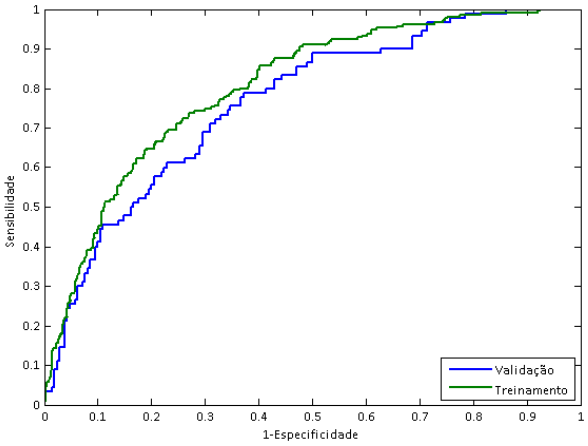


Figura 5.17: Curva ROC para Rede Neural

5.3 Análise dos Resultados

Após a execução dos modelos, conforme descrito na seção anterior, podemos confrontar o resultado de cada um deles, de acordo com os indicadores propostos. O quadro 5.2 sintetiza

os resultados.

Indicadores	Base de Dados	Modelos		
		Regressão Logística	Árvore de Decisão	Rede Neural
KS	Treinamento	0,454422	0,604082	0,468707
	Validação	0,465079	0,341270	0,407937
Gini	Treinamento	0,625870	0,755355	0,606297
	Validação	0,577884	0,422011	0,528466

Tabela 5.2: Quadro de resultados

Além do quadro, a Figura 5.18, é uma representação dos resultados encontrados, facilitando a análise dos dados do quadro 5.2. Não há qualquer indicação na literatura do critério de escolha do melhor modelo. Pode-se tomar a decisão olhando para os números dos indicadores em absoluto, considerando ou não a diferença entre os resultados com os dados de treinamento e validação. Pode-se também considerar, além do valor dos indicadores, a capacidade do modelo em predizer tanto com os dados de treinamento quanto validação, bem como podem ser consideradas uma série de outras características, como tempo de processamento, grau dificuldade de implantação ou mesmo de interpretação do processo de modelagem.

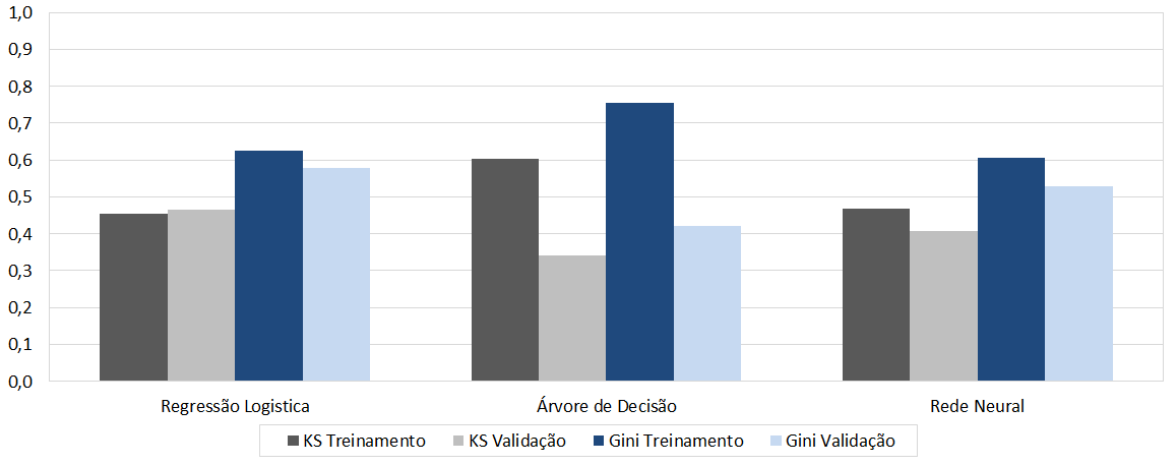


Figura 5.18: Comparativo do desempenho dos modelos

Espera-se que um modelo de *credit score* seja capaz de de ter um bom desempenho não apenas nos dados de treinamento, mas principalmente nos dados de validação. O primeiro diz respeito a clientes adimplentes e inadimplentes observados em um algum momento do passado, onde as decisões já foram tomadas e os resultados da operação (positivos e negativos) já contabilizados. Se as características dos futuros clientes não se tornar muito distinta dos considerados na construção do modelo, espera-se que ele seja capaz de classificar os futuros clientes antes da operação de credito, reduzindo assim a exposição da instituição financeira. Sendo assim, é importante olhar principalmente para o desempenho dos modelos em relação aos dados de validação.

A Árvore de Decisão, apresentou um desempenho considerável apenas para os dados de treinamento, tendo nos dados de validação um comportamento bastante inferior. Isso é fruto do *trade-off* entre construir uma árvore considerando apenas poucas informações e ter uma menor disparidade entre os desempenhos ou incluir mais informações e elevar essa disparidade, conforme foi apontado na Seção 5.2.2. Ao construir um modelo que leva em consideração poucas informações, conforme a Figura 5.10 , poderíamos elevar o desempenho nos os dados de treinamento, mas ficaria-se suscetível ao fato de que uma mudança no perfil futuro dessas poucas informações consideradas levar a resultados inconsistentes.

Quando analisados os desempenhos tanto dos modelos de Regressão Logística quanto da Rede Neural, pode-se concluir que tiveram um desempenho bastante semelhante, com uma ligeira vantagem para o primeiro. Tanto o Índice KS quanto o de Gini apresentaram uma ligeira discrepância entre os resultados com os dados de treinamento e validação na Rede Neural. Essa redução é natural, uma vez que os dados são distintos dos utilizados no treinamento do modelo, possuindo características distintas. Entretanto, essa redução é menor no modelo de Regressão Logística para o Índice de Gini e inexistente quando considerado o

KS.

Além do melhor desempenho da Regressão Logística, um outro fator é favorável a esta técnica. A partir dos coeficientes estimados, conforme a Tabela 5.1, é possível interpretar o efeito provocado por cada uma das variáveis na resposta do modelo. É possível identificar, por exemplo, que a variável VAR13 (Idade) por ter um coeficiente negativo tem um efeito inverso na resposta, ou seja, pessoas com uma idade mais elevada tendem a ter uma probabilidade de inadimplência menor. Esse tipo de análise não é possível de ser feita com o modelo de Redes Neurais, uma vez que o efeito que cada uma das variáveis tem na resposta está de certa forma escondido entre os pesos sinápticos nas camadas intermediarias da rede.

6 Conclusões

Levando-se em consideração outros fatores como a facilidade de interpretação do modelo ou mesmo o tempo de processamento além dos Índices de Gini e KS, podemos afirmar que o modelo de Regressão Logística é o mais indicado para a predição da probabilidade de um individuo vir a se tornar inadimplente. O principal beneficio desse modelo é a sólida teoria já desenvolvida, que permite a construção de intervalos de confiança além de outras características dos coeficientes, como a *odds-ratio*. A análise dos coeficientes, tanto em um sentido econômico quanto estatístico, fornece uma série de informações sobre o modelo, como mensurar a importância de cada variável para o modelo proposto, bem como um possível teste de sensibilidade.

Entretanto, não se pode desconsiderar o poder preditivo das duas outras técnicas. O modelo de Árvore de Decisão claramente tem um bom desempenho para conjuntos de dados em que se dispõe de um reduzido número de variáveis explicativas. Quando esse número aumenta, eleva-se também o problema de super-estimação. Em situações que um número pequeno de variáveis explicativas for suficiente para a estimação, esse modelo pode produzir excelentes resultados. Porém, em um contexto de *credit score*, onde o objetivo e mensurar a probabilidade de um individuo vir a se tornar inadimplente, e o conjunto de indivíduos em uma população tende a ser heterogêneo, demanda-se um número maior de variáveis para que se possa capturar características distintas de uma subpopulação. Por essa razão, o modelo

apresentou um desempenho reduzido.

Por outro lado, o modelo de Redes Neurais apresentou um desempenho satisfatório. Para a estrutura proposta (MLP), o número de neurônios nas camadas intermediarias pareceu não ter uma relação com os resultados produzidos, que ficaram relativamente constantes, conforme puderam ser observados os valores para os Índices de Gini e KS. Entretanto, foi testada apenas uma estrutura, e conforme (HAYKIN, 1999), há uma série de outras configurações que poderiam produzir resultados distintos dos apresentados pela estrutura proposta.

A quantidade de técnicas para a classificação de padrões, que podem ser aplicadas para a classificação de indivíduos em adimplentes e inadimplentes, ou seja, *credit score* teve um desenvolvimento muito grande nos últimos anos. Pode-se citar as técnicas de SVM (do inglês *support vector machine*)[1], Lógica *Fuzzy* [2], Algoritmos Genéticos [3], k-NN[4] (do ingês k-nearest neighbor), Programação Linear [5], Análise de Sobrevivência [6], dentre outras.

Além da possibilidade de se comparar diferentes técnicas, tem-se que levar em consideração a qualidade e os tipos de dados a serem utilizados. No presente trabalho, foram utilizados dados dados sem um índice temporal, ou seja, foi modelada a probabilidade de inadimplência para um único instante do tempo. Porem, deve-se considerar que essa probabilidade pode mudar ao longo do tempo, de forma que um determinado individuo, com um determinado vetor de características apresente diferentes probabilidades ao longo do tempo. Ao invés de se considerar apenas variáveis idiossincráticas do individuo, a inclusão de variáveis macroeconômicas (PIB, desemprego, renda etc.) pode elevar a capacidade preditiva dos modelos de *credit score* se for realizada uma modelagem considerando o tempo, conforme Bellotti

[1] (HUANG et al., 2004),(WORRACHARTDATCHAI, 2007) e (GESTEL et al., 2003)
[2](LAHSANA RAJA N. AINON, 2010), (CEAUSU et al., 2010) e (HOFFMANN et al., 2006)
[3](DESAI et al., 1997) e (LAHSANA RAJA N. AINON, 2010)
[4](ALLEN; ALLEN-DRISCOLL, 2002)
[5](ERENGUC, 1990) e Gochet et al. (1997)
[6](BELLOTTI, 2009),(GIAMBONA, 2006) e (STEPANOVA, 2002)

(2009).

APÊNDICE A – Rotinas SAS

A.1 Manipulação e tratamento dos dados

```
/*Cria Libname*/
LIBNAME GERMAN 'D:\MONO\DATASET';

/*Download e Import da base de dados, diretamente do repositório*/
filename DB url 'http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/
    german/german.data';
PROC IMPORT
  DATAFILE=DB
  OUT=GERMAN.DADOS
  DBMS=DLM REPLACE;
    DELIMITER=" ";
    GETNAMES=NO;
RUN;

/*Renomeia Variável resposta*/
DATA GERMAN.DADOS (DROP = VAR21);
SET GERMAN.DADOS;
  R = VAR21;
RUN;

/*Label das variáveis*/
DATA GERMAN.DADOS;
SET GERMAN.DADOS;
  LABEL   VAR1 = 'Status de conta corrente já existente'
      VAR2 = 'Duração em meses do empréstimo'
      VAR3 = 'Histórico de crédito'
      VAR4 = 'Propósito do empréstimo'
      VAR5 = 'Quantia do Empréstimo'
      VAR6 = 'Contas de poupança/Títulos'
      VAR7 = 'Duração do atual emprego'
      VAR8 = '% da parcela em relação a renda'
      VAR9 = 'Estado civil e sexo'
      VAR10 = 'Outros fiadores/garantidores'
      VAR11 = 'Anos na atual residencia'
      VAR12 = 'Patrimônio'
      VAR13 = 'Idade em anos'
      VAR14 = 'Outros planos de empréstimo'
      VAR15 = 'Status da habitação'
      VAR16 = 'Número de créditos existentes no banco'
      VAR17 = 'Categoria do trabalho'
      VAR18 = 'Número de dependentes'
      VAR19 = 'Tem telefone'
      VAR20 = 'Estrangeiro ou não'
```

```
      R = 'Resposta';
RUN;


DATA GERMAN.DADOS;
SET GERMAN.DADOS;
/*VARIAVEL 1*/
  IF VAR1 = 'A11' THEN VAR1 = '0';
  IF VAR1 = 'A12' THEN VAR1 = '1';
  IF VAR1 = 'A13' THEN VAR1 = '2';
  IF VAR1 = 'A14' THEN VAR1 = '3';

/*VARIAVEL 3*/
  IF VAR3 = 'A30' THEN VAR3 = '0';
  IF VAR3 = 'A31' THEN VAR3 = '1';
  IF VAR3 = 'A32' THEN VAR3 = '2';
  IF VAR3 = 'A33' THEN VAR3 = '3';
  IF VAR3 = 'A34' THEN VAR3 = '4';

/*VARIAVEL 4*/
  IF VAR4 = 'A40' THEN VAR4 = '0';
  IF VAR4 = 'A41' THEN VAR4 = '1';
  IF VAR4 = 'A42' THEN VAR4 = '2';
  IF VAR4 = 'A43' THEN VAR4 = '3';
  IF VAR4 = 'A44' THEN VAR4 = '4';
  IF VAR4 = 'A45' THEN VAR4 = '5';
  IF VAR4 = 'A46' THEN VAR4 = '6';
  IF VAR4 = 'A47' THEN VAR4 = '7';
  IF VAR4 = 'A48' THEN VAR4 = '8';
  IF VAR4 = 'A49' THEN VAR4 = '9';
  IF VAR4 = 'A410' THEN VAR4 = '10';

/*VARIAVEL 6*/
  IF VAR6 = 'A61' THEN VAR6 = '0';
  IF VAR6 = 'A62' THEN VAR6 = '1';
  IF VAR6 = 'A63' THEN VAR6 = '2';
  IF VAR6 = 'A64' THEN VAR6 = '3';
  IF VAR6 = 'A65' THEN VAR6 = '4';

/*VARIAVEL 7*/
  IF VAR7 = 'A71' THEN VAR7 = '0';
  IF VAR7 = 'A72' THEN VAR7 = '1';
  IF VAR7 = 'A73' THEN VAR7 = '2';
  IF VAR7 = 'A74' THEN VAR7 = '3';
  IF VAR7 = 'A75' THEN VAR7 = '4';

/*VARIAVEL 9*/
  IF VAR9 = 'A91' THEN VAR9 = '0';
  IF VAR9 = 'A92' THEN VAR9 = '1';
  IF VAR9 = 'A93' THEN VAR9 = '2';
  IF VAR9 = 'A94' THEN VAR9 = '3';
  IF VAR9 = 'A95' THEN VAR9 = '4';

/*VARIAVEL 10*/
  IF VAR10 = 'A101' THEN VAR10 = '0';
  IF VAR10 = 'A102' THEN VAR10 = '1';
  IF VAR10 = 'A103' THEN VAR10 = '2';

/*VARIAVEL 12*/
  IF VAR12 = 'A121' THEN VAR12 = '0';
  IF VAR12 = 'A122' THEN VAR12 = '1';
  IF VAR12 = 'A123' THEN VAR12 = '2';
  IF VAR12 = 'A124' THEN VAR12 = '3';
```

```
/*VARIAVEL 14*/
  IF VAR14 = 'A141' THEN VAR14 = '0';
  IF VAR14 = 'A142' THEN VAR14 = '1';
  IF VAR14 = 'A143' THEN VAR14 = '2';

/*VARIAVEL 15*/
  IF VAR15 = 'A151' THEN VAR15 = '0';
  IF VAR15 = 'A152' THEN VAR15 = '1';
  IF VAR15 = 'A153' THEN VAR15 = '2';

/*VARIAVEL 17*/
  IF VAR17 = 'A171' THEN VAR17 = '0';
  IF VAR17 = 'A172' THEN VAR17 = '1';
  IF VAR17 = 'A173' THEN VAR17 = '2';
  IF VAR17 = 'A174' THEN VAR17 = '3';

/*VARIAVEL 19*/
  IF VAR19 = 'A191' THEN VAR19 = '0';
  IF VAR19 = 'A192' THEN VAR19 = '1';

/*VARIAVEL 20*/
  IF VAR20 = 'A201' THEN VAR20 = '0';
  IF VAR20 = 'A202' THEN VAR20 = '1';

/*Resposta*/
  IF R = '1' THEN R = '0';
  IF R = '2' THEN R = '1';
RUN;
```

A.2 Stepwise para seleção de variáveis

```
/*Seleciona as variáveis mais significativas*/
PROC PHREG DATA=GERMAN.DADOS OUT=VARIAVEIS_SELECIONADAS;
  CLASS VAR1 VAR3 VAR4 VAR6 VAR7 VAR9 VAR10 VAR12 VAR14 VAR15 VAR17 VAR19 VAR20;
    MODEL R=VAR1-VAR20 /
  SELECTION=STEPWISE
  SLENTRY=1
    SLSTAY=0.5
  DETAILS
;RUN;

/* Cria macro variavel com o vetor de variaveis selecionadas */
PROC TRANSPOSE DATA=VARIAVEIS_SELECIONADAS OUT=VARS;
RUN;


DATA VARS;
SET VARS;
  IF R = . THEN DELETE;
  IF _NAME_ = "_LNLIKE_" THEN DELETE;
  VARIAVEL = SUBSTR(_NAME_,1,LENGTH(_NAME_)-LENGTH("A"||SCAN(_NAME_,3,"A")));
  IF LENGTH(_NAME_) LE 5 THEN VARIAVEL = _NAME_;
RUN;

PROC SQL NOPRINT;
  SELECT DISTINCT VARIAVEL INTO: VARS SEPARATED BY " "
  FROM VARS
QUIT;
PROC SQL NOPRINT;
```

```
  SELECT  COUNT(DISTINCT VARIAVEL) INTO: N_VARS SEPARATED BY " "
  FROM VARS
QUIT;
%PUT VARIAVEIS SELECIONADAS: &VARS.;
%PUT NUMERO DE VARIAVEIS SELECIONADAS: &N_VARS.;
```

A.3 Criação das bases de dados

```
/*Cria base de desenvolvimento*/
PROC SORT DATA=GERMAN.DADOS (KEEP= R &VARS.);
  BY R;
PROC SURVEYSELECT DATA=GERMAN.DADOS
  METHOD = SRS
  SEED=569144001
  OUT=TREINAMENTO
  SAMPRATE=0.7;
  STRATA R;
RUN;

/*BASES DE TREINAMENTO E VALIDAÇÃO*/
DATA TREINAMENTO;
SET TREINAMENTO (KEEP = R &VARS.);
DATA AMOSTRA;
SET TREINAMENTO;
  CHAVE = 1;
DATA AMOSTRA_2;
SET AMOSTRA
  GERMAN.DADOS;
PROC SORT DATA=AMOSTRA_2 NODUPKEY;
  BY &VARS.;
DATA VALIDACAO (DROP=CHAVE);
SET AMOSTRA_2;
  IF CHAVE=1 THEN DELETE;
RUN;

/*Valida bases de desenvolvimento e validacao*/
DATA COMPARAR;
SET TREINAMENTO
  VALIDACAO;
PROC SORT DATA=GERMAN.DADOS;
  BY &VARS.;
PROC SORT DATA=COMPARAR;
  BY &VARS.;
PROC COMPARE BASE=GERMAN.DADOS COMPARE=COMPARAR;
RUN;

DATA FULL;
SET COMPARAR;
RUN;

PROC CONTENTS DATA=GERMAN.DADOS OUT=CONTENTS;
RUN;

PROC SQL NOPRINT;
  SELECT NAME INTO: VARS_CATEG SEPARATED BY " "
  FROM CONTENTS
  WHERE TYPE = 2;
PROC SQL NOPRINT;
  SELECT COUNT(NAME) INTO: N_VARS_CATEG
```

```
  FROM CONTENTS
  WHERE TYPE = 2;
PROC SQL NOPRINT;
  SELECT NAME INTO: VARS_NUM SEPARATED BY " "
  FROM CONTENTS
  WHERE TYPE = 1;
PROC SQL NOPRINT;
  SELECT COUNT(NAME) INTO: N_VARS_NUM
  FROM CONTENTS
  WHERE TYPE = 1;
QUIT;

%PUT VARIAVEIS CATEGORICAS: &VARS_CATEG ====== (&N_VARS_CATEG.);
%PUT VARIAVEIS NUMÉRICAS: &VARS_NUM ====== (&N_VARS_NUM.);

%MACRO DUMMYS(BASE,BASE_OUT);
  DATA &BASE_OUT;
  SET  &BASE.;
    %DO I = 1 %TO &N_VARS_CATEG.;
      %LET SELEC = %SCAN(&VARS_CATEG.,&I.,' ');
      %DO J = 0 %TO 10;
        &SELEC._&J. = 0; IF &SELEC. = &J. THEN &SELEC._&J. = 1;
      %END;
    %END;
  RUN;

  PROC SQL;
  CREATE TABLE PRE AS
  SELECT  SUM(R) AS R
      %DO I = 1 %TO &N_VARS_CATEG.;
        %LET SELEC = %SCAN(&VARS_CATEG.,&I.,' ');
        %DO J = 0 %TO 10;
          ,SUM(&SELEC._&J.*1) AS &SELEC._&J.
        %END;
      %END;
  FROM  &BASE_OUT.;
  QUIT;

  PROC TRANSPOSE DATA=PRE OUT=PRE2 name=variaveis
  prefix=n;
  RUN;

  DATA PRE;
  SET PRE;
    IF n1 = 0 THEN DELETE;
  RUN;

  PROC SQL NOPRINT;
    SELECT variaveis INTO: VARS_DUMMY SEPARATED BY " "
    FROM PRE;
  QUIT;
  %PUT &VARS_DUMMY;

  DATA &BASE_OUT.;
  SET &BASE_OUT. (KEEP =&VARS_NUM &VARS_DUMMY) ;
  RUN;
%MEND;
%DUMMYS(TREINAMENTO,TREINAMENTO_DUMMY);
%DUMMYS(VALIDACAO,VALIDACAO_DUMMY);
%DUMMYS(FULL,FULL_DUMMY);

/*Salva as bases no format DAT, para futura importação nos outros softwares*/
%MACRO EXPORT(BASEIN);
  DATA GERMAN.&BASEIN.;
```

```
116   SET &BASEIN.;
117   RUN;

119   PROC EXPORT DATA=&BASEIN REPLACE
120     OUTFILE="D:\MONO\DATASET\&BASEIN..DAT"
121     DBMS=DLM;
122     DELIMITER=' ';
123   RUN;
124 %MEND;

126 %export(treinamento);
127 %export(treinamento_dummy);
128 %export(validacao);
129 %export(validacao_dummy);
130 %export(full);
131 %export(full_dummy);
```

APÊNDICE B – Rotinas Matlab

B.1 Modelo de Regressão Logistica

```
%Importa bases de dados
Pasta = 'D:\Mono\DataSet\';
treinamento_dummy = importdata(strcat(Pasta,'\treinamento_dummy.dat'),' ',1);
validacao_dummy = importdata(strcat(Pasta,'\validacao_dummy.dat'),' ',1);

treinamento_dummy = treinamento_dummy.data;
validacao_dummy = validacao_dummy.data;

treinamento_r = treinamento_dummy(:,1);
validacao_r = validacao_dummy(:,1);

treinamento_vars = treinamento_dummy;
treinamento_vars(:,1) = [];

validacao_vars = validacao_dummy;
validacao_vars(:,1) = [];

%=====>Regressão Logistica
[Betas, dev,stats] = glmfit(treinamento_vars,[treinamento_r],'binomial','link','logit'
    );

%Cria coluna de 1s
LinhasValid = length(validacao_vars(:,1));
LinhasDesenv = length(treinamento_vars(:,1));

for i=1:LinhasValid;
    Um_Valid(i,1)=1;
end;
for i=1:LinhasDesenv;
    Um_Desenv(i,1)=1;
end;

validacao_vars2 = [Um_Valid validacao_vars];
treinamento_vars2 = [Um_Desenv treinamento_vars];

%Calcula logit para observações de validação
CalcValid = validacao_vars2*Betas;
CalcDesenv = treinamento_vars2*Betas;

for i=1:LinhasValid;
    LogitValid(i,1) = exp(CalcValid(i,1))/(1+exp(CalcValid(i,1)));
end  ;
```

```
for i=1:LinhasDesenv;
    LogitDesenv(i,1) = exp(CalcDesenv(i,1))/(1+exp(CalcDesenv(i,1)));
end  ;

%Cria Histograma (Intervalo: 0-1, em faixas de 0,1)
j=1;
k=1;
for i=1:LinhasDesenv;
    if (treinamento_r(i,1) == 0);
        LogitDesenvBom(j,1) = LogitDesenv(i,1);
        j=j+1;
    end;
    if (treinamento_r(i,1) == 1);
        LogitDesenvMau(k,1) = LogitDesenv(i,1);
        k=k+1;
    end;
end;
NumDesenvBom = j-1;
NumDesenvMau = k-1;
clear i j k;

j=1;
k=1;
for i=1:LinhasValid;
    if (validacao_r(i,1) == 0);
        LogitValidBom(j,1) = LogitValid(i,1);
        j=j+1;
    end;
    if (validacao_r(i,1) == 1);
        LogitValidMau(k,1) = LogitValid(i,1);
        k=k+1;
    end;
end;
NumValidBom = j-1;
NumValidMau = k-1;
clear i j k;

%=====> Histogramas
%Desenvolvimento
Hist_Desenv(1,:)=0:0.01:1;%1ªLinha: Faixas;
Hist_Desenv(2,:)=histc(sort(LogitDesenvBom),Hist_Desenv(1,:));%2ªLinha: Absoluto de
    Bom;
Hist_Desenv(3,:)=histc(sort(LogitDesenvMau),Hist_Desenv(1,:));%3ªLinha: Absoluto de
    Mau;
for i=1:length(Hist_Desenv(1,:));
    Hist_Desenv(4,i)=(Hist_Desenv(2,i))/NumDesenvBom;%4ªLinha: Relativo de Bom
    Hist_Desenv(5,i)=(Hist_Desenv(3,i))/NumDesenvMau;%5ªLinha: Relativo de Mau
end; clear i
Hist_Desenv(6,:)=cumsum(Hist_Desenv(2,:));%6ªLinha: Acum(Absoluto de Bom)
Hist_Desenv(7,:)=cumsum(Hist_Desenv(3,:));%7ªLinha: Acum(Absoluto de Mau)
Hist_Desenv(8,:)=cumsum(Hist_Desenv(4,:));%8ªLinha: Acum(Relativo de Bom)
Hist_Desenv(9,:)=cumsum(Hist_Desenv(5,:));%9ªLinha: Acum(Relativo de Mau)

%Validação
Hist_Valid(1,:)=0:0.01:1;
Hist_Valid(2,:)=histc(sort(LogitValidBom),Hist_Valid(1,:));
Hist_Valid(3,:)=histc(sort(LogitValidMau),Hist_Valid(1,:));
for i=1:length(Hist_Valid(1,:));
    Hist_Valid(4,i)=(Hist_Valid(2,i))/NumValidBom;
    Hist_Valid(5,i)=(Hist_Valid(3,i))/NumValidMau;
end; clear i;
Hist_Valid(6,:)=cumsum(Hist_Valid(2,:));
Hist_Valid(7,:)=cumsum(Hist_Valid(3,:));
Hist_Valid(8,:)=cumsum(Hist_Valid(4,:));
```

```
Hist_Valid(9,:)=cumsum(Hist_Valid(5,:));

%=====> KS
for i=1:length(Hist_Desenv(1,:));
   KS_Desenv=(max(Hist_Desenv(8,:)-Hist_Desenv(9,:)));
end; clear i;
for i=1:length(Hist_Valid(1,:));
   KS_Valid=(max(Hist_Valid(8,:)-Hist_Valid(9,:)));
end; clear i;

subplot(1,2,1);
plot(Hist_Desenv(8,:),'LineWidth',1.5,'Color',[0 0 0]);hold all;
plot(Hist_Desenv(9,:),'LineWidth',1.5,'Color',[.8 .8 .8]);hold all;
axis([0 100 0 1]);
ylabel('%'); xlabel('Score(x100)'); title('KS - Base de Treinamento');
legend('Adimplentes','Inadimplentes','Location','SouthEast');

subplot(1,2,2);
plot(Hist_Valid(8,:),'LineWidth',1.5,'Color',[0 0 1]);hold all;
plot(Hist_Valid(9,:),'LineWidth',1.5,'Color',[.8 .8 1]);hold all;
axis([0 100 0 1]);
ylabel('%'); xlabel('Score(x100)'); title('KS-Base de Validação');
legend('Adimplentes','Inadimplentes','Location','SouthEast');

%=====> ROC Curve
[Xd,Yd,T,AUC_Logit_Desenv] = perfcurve(treinamento_r,LogitDesenv,1);
[Xv,Yv,T,AUC_Logit_Valid] = perfcurve(validacao_r,LogitValid,1);

figure
plot(Xv,Yv,Xd,Yd,'LineWidth',2)
ylabel('Sensibilidade'); xlabel('1-Especificidade');
legend('Validação','Treinamento','Location','SouthEast')

%=====> Gini
Gini_Logit_Desenv = 2*(AUC_Logit_Desenv-0.5);
Gini_Logit_Valid = 2*(AUC_Logit_Valid-0.5);
```

B.2 Árvore de Decisão

```
%Importa bases de dados
Pasta = 'D:\Mono\DataSet\';
treinamento_dummy = importdata(strcat(Pasta,'\treinamento_dummy.dat'),' ',1);
treinamento = importdata(strcat(Pasta,'\treinamento.dat'),' ',1);
validacao_dummy = importdata(strcat(Pasta,'\validacao_dummy.dat'),' ',1);
validacao = importdata(strcat(Pasta,'\validacao.dat'),' ',1);

treinamento_dummy = treinamento_dummy.data;
treinamento = treinamento.data;
validacao_dummy = validacao_dummy.data;
validacao = validacao.data;

treinamento_r = treinamento(:,1);
validacao_r = validacao(:,1);

treinamento_vars = treinamento;
treinamento_vars(:,1) = [];

validacao_vars = validacao;
validacao_vars(:,1) = [];
```

```
nomes = {'VAR1' 'VAR2' 'VAR3' 'VAR20' 'VAR13' 'VAR10' 'VAR8' 'VAR5' 'VAR19' 'VAR14' '
    VAR4'};

%Constrói a árvore
arvore =classregtree(treinamento_vars,treinamento_r,'method','regression','categorical
    ',[1 3 4 7 9 10 11 12],'prune','off','names', nomes);
Tipo=type(arvore);
view(arvore);

%Poda
arvore = treeprune(arvore,'level',29);
view(arvore);

%Resultados encontrados
Arvore_validacao_res = eval(arvore, validacao_vars);
Arvore_treinamento_res = eval(arvore, treinamento_vars);

%Cria Histograma (Intervalo: 0-1, em faixas de 0,1)
LinhasValid = length(validacao_vars(:,1));
LinhasDesenv = length(treinamento_vars(:,1));

%Separa bons e maus da base de desenvolvimento
j=1; k=1;
for i=1:LinhasDesenv;
    if (treinamento_r(i,1) == 0);
        Arvore_Desenv_Bom(j,1) = Arvore_treinamento_res(i,1);
        j=j+1;
    end
    if (treinamento_r(i,1) == 1);
        Arvore_Desenv_Mau(k,1) = Arvore_treinamento_res(i,1);
        k=k+1;
    end
end
NumDesenvBom = j-1;
NumDesenvMau = k-1;
clear i j k;

%Separa bons e maus da base de Validação
j=1; k=1;
for i=1:LinhasValid;
    if (validacao_r(i,1) == 0);
        Arvore_Valid_Bom(j,1) = Arvore_validacao_res(i,1);
        j=j+1;
    end;
    if (validacao_r(i,1) == 1);
        Arvore_Valid_Mau(k,1) = Arvore_validacao_res(i,1);
        k=k+1;
    end;
end
NumValidBom = j-1;
NumValidMau = k-1;
clear i j k;

%=====> Histogramas
%Histograma dos dados de desenvolvimento
Hist_Desenv(1,:)=0:0.01:1; %Faixas
Hist_Desenv(2,:)=histc(sort(Arvore_Desenv_Bom),Hist_Desenv(1,:)); %Absoluto de Bom
Hist_Desenv(3,:)=histc(sort(Arvore_Desenv_Mau),Hist_Desenv(1,:)); %Absoluto de Mau
for i=1:length(Hist_Desenv(1,:));
    Hist_Desenv(4,i)=(Hist_Desenv(2,i))/NumDesenvBom; %Relativo de Bom
    Hist_Desenv(5,i)=(Hist_Desenv(3,i))/NumDesenvMau; %Relativo de Mau
end; clear i;
Hist_Desenv(6,:)=cumsum(Hist_Desenv(2,:)); %Acum(Absoluto de Bom)
```

```
Hist_Desenv(7,:)=cumsum(Hist_Desenv(3,:)); %Acum(Absoluto de Mau)
Hist_Desenv(8,:)=cumsum(Hist_Desenv(4,:)); %Acum(Relativo de Bom)
Hist_Desenv(9,:)=cumsum(Hist_Desenv(5,:)); %Acum(Relativo de Mau)

%Histograma dos dados de validação
Hist_Valid(1,:)=0:0.01:1; %1ªLinha: Faixas
Hist_Valid(2,:)=histc(sort(Arvore_Valid_Bom),Hist_Valid(1,:)); %2ªLinha: Absoluto de
    Bom
Hist_Valid(3,:)=histc(sort(Arvore_Valid_Mau),Hist_Valid(1,:)); %3ªLinha: Absoluto de
    Mau
for i=1:length(Hist_Valid(1,:));
    Hist_Valid(4,i)=(Hist_Valid(2,i))/NumValidBom; %4ªLinha: Relativo de Bom
    Hist_Valid(5,i)=(Hist_Valid(3,i))/NumValidMau; %5ªLinha: Relativo de Mau
end; clear i;
Hist_Valid(6,:)=cumsum(Hist_Valid(2,:)); %6ªLinha: Acum(Absoluto de Bom)
Hist_Valid(7,:)=cumsum(Hist_Valid(3,:)); %7ªLinha: Acum(Absoluto de Mau)
Hist_Valid(8,:)=cumsum(Hist_Valid(4,:)); %8ªLinha: Acum(Relativo de Bom)
Hist_Valid(9,:)=cumsum(Hist_Valid(5,:)); %9ªLinha: Acum(Relativo de Mau)

%=====> KS
for i=1:length(Hist_Desenv(1,:));
   KS_Desenv=(max(Hist_Desenv(8,:)-Hist_Desenv(9,:)));
end; clear i;
for i=1:length(Hist_Valid(1,:));
   KS_Valid=(max(Hist_Valid(8,:)-Hist_Valid(9,:)));
end; clear i;

subplot(1,2,1);
plot(Hist_Desenv(8,:),'LineWidth',1.5,'Color',[0 0 0]);hold all;
plot(Hist_Desenv(9,:),'LineWidth',1.5,'Color',[.8 .8 .8]);hold all;
axis([0 100 0 1]);
ylabel('%'); xlabel('Score(x100)'); title('KS - Base de Treinamento');
legend('Adimplentes','Inadimplentes','Location','SouthEast');

subplot(1,2,2);
plot(Hist_Valid(8,:),'LineWidth',1.5,'Color',[0 0 1]);hold all;
plot(Hist_Valid(9,:),'LineWidth',1.5,'Color',[.8 .8 1]);hold all;
axis([0 100 0 1]);
ylabel('%'); xlabel('Score(x100)'); title('KS-Base de Validação');
legend('Adimplentes','Inadimplentes','Location','SouthEast');

%=====> ROC Curve
[Xd,Yd,~,AUC_Logit_Desenv] = perfcurve(treinamento_r,Arvore_treinamento_res,1);
[Xv,Yv,~,AUC_Logit_Valid] = perfcurve(validacao_r,Arvore_validacao_res,1);

figure
plot(Xv,Yv,Xd,Yd,'LineWidth',2)
ylabel('Sensibilidade'); xlabel('1-Especificidade');
legend('Validação','Treinamento','Location','SouthEast')

%=====> Gini
Gini_Logit_Desenv = 2*(AUC_Logit_Desenv-0.5);
Gini_Logit_Valid = 2*(AUC_Logit_Valid-0.5);
```

B.3 Rede Neural

```
%Importa bases de dados
Pasta = 'D:\Mono\DataSet\';
treinamento_dummy = importdata(strcat(Pasta,'\treinamento_dummy.dat'),' ',1);
```

```
treinamento = importdata(strcat(Pasta,'\treinamento.dat'),' ',1);
validacao_dummy = importdata(strcat(Pasta,'\validacao_dummy.dat'),' ',1);
validacao = importdata(strcat(Pasta,'\validacao.dat'),' ',1);

treinamento_dummy = treinamento_dummy.data;
treinamento = treinamento.data;
validacao_dummy = validacao_dummy.data;
validacao = validacao.data;

treinamento_r = treinamento(:,1);
validacao_r = validacao(:,1);

treinamento_vars = treinamento_dummy;
treinamento_vars(:,1) = [];

validacao_vars = validacao_dummy;
validacao_vars(:,1) = [];

Rede = feedforwardnet;
Rede = init(Rede);

%Numero de Camadas e Neuronios
Rede.numLayers = 3;
Rede.layers{1}.size = 25;
Rede.layers{2}.size = 15;
Rede.layers{3}.size = 1;

%Conexões
Rede.biasConnect = [1;1;1];
Rede.inputConnect = [1;0;0];
Rede.layerConnect = [0 0 0;1 0 0;0 1 0];
Rede.outputConnect = [0 0 1];

%Funções de Treinamento e Tranferencia
Rede.trainFcn = 'trainscg';
Rede.performFcn = 'mse';
Rede.layers{1}.transferFcn = 'purelin'; %Linear
Rede.layers{2}.transferFcn = 'purelin'; %Linear
Rede.layers{3}.transferFcn = 'logsig'; %Log-Sigmoid


%Configurações de Parada
Rede.trainParam.showCommandLine = true; %Exibe informações no prompt
Rede.trainParam.min_grad = 0.0000000001; %Gradiente
Rede.trainParam.max_fail = 1000; %Erro maximo
Rede.trainParam.epochs = 1000; %Iterações
Rede.trainParam.goal = 0; %Objetivo

[Rede,treino] = train(Rede, transpose(treinamento_vars), transpose(treinamento_r));

%Resultados encontrados
Rede_validacao_res = Rede(transpose(validacao_vars));
Rede_treinamento_res = Rede(transpose(treinamento_vars));

Rede_validacao_res=transpose(Rede_validacao_res);
Rede_treinamento_res=transpose(Rede_treinamento_res);

%Cria Histograma (Intervalo: 0-1, em faixas de 0,1)
LinhasValid = length(validacao_vars(:,1));
LinhasDesenv = length(treinamento_vars(:,1));

%Separa bons e maus da base de desenvolvimento
j=1; k=1;
for i=1:LinhasDesenv;
```

```
68      if (treinamento_r(i,1) == 0);
69          Rede_Desenv_Bom(j,1) = Rede_treinamento_res(i,1);
70          j=j+1;
71      end
72      if (treinamento_r(i,1) == 1);
73          Rede_Desenv_Mau(k,1) = Rede_treinamento_res(i,1);
74          k=k+1;
75      end
76  end
77  NumDesenvBom = j-1;
78  NumDesenvMau = k-1;
79  clear i j k;

81  %Separa bons e maus da base de Validação
82  j=1; k=1;
83  for i=1:LinhasValid;
84      if (validacao_r(i,1) == 0);
85          Rede_Valid_Bom(j,1) = Rede_validacao_res(i,1);
86          j=j+1;
87      end;
88      if (validacao_r(i,1) == 1);
89          Rede_Valid_Mau(k,1) = Rede_validacao_res(i,1);
90          k=k+1;
91      end;
92  end
93  NumValidBom = j-1;
94  NumValidMau = k-1;
95  clear i j k;

97  %=====> Histogramas
98  %Histograma dos dados de desenvolvimento
99  Hist_treinamento_rede(1,:)=0:0.01:1; %1ªLinha: Faixas
100 Hist_treinamento_rede(2,:)=histc(sort(Rede_Desenv_Bom),Hist_treinamento_rede(1,:)); %2
        ªLinha: Absoluto de Bom
101 Hist_treinamento_rede(3,:)=histc(sort(Rede_Desenv_Mau),Hist_treinamento_rede(1,:)); %3
        ªLinha: Absoluto de Mau
102 for i=1:length(Hist_treinamento_rede(1,:));
103     Hist_treinamento_rede(4,i)=(Hist_treinamento_rede(2,i))/NumDesenvBom; %4ªLinha:
            Relativo de Bom
104     Hist_treinamento_rede(5,i)=(Hist_treinamento_rede(3,i))/NumDesenvMau; %5ªLinha:
            Relativo de Mau
105 end; clear i;
106 Hist_treinamento_rede(6,:)=cumsum(Hist_treinamento_rede(2,:)); %6ªLinha: Acum(Absoluto
        de Bom)
107 Hist_treinamento_rede(7,:)=cumsum(Hist_treinamento_rede(3,:)); %7ªLinha: Acum(Absoluto
        de Mau)
108 Hist_treinamento_rede(8,:)=cumsum(Hist_treinamento_rede(4,:)); %8ªLinha: Acum(Relativo
        de Bom)
109 Hist_treinamento_rede(9,:)=cumsum(Hist_treinamento_rede(5,:)); %9ªLinha: Acum(Relativo
        de Mau)

111 %Histograma dos dados de validação
112 Hist_validacao_rede(1,:)=0:0.01:1; %1ªLinha: Faixas
113 Hist_validacao_rede(2,:)=histc(sort(Rede_Valid_Bom),Hist_validacao_rede(1,:)); %2
        ªLinha: Absoluto de Bom
114 Hist_validacao_rede(3,:)=histc(sort(Rede_Valid_Mau),Hist_validacao_rede(1,:)); %3
        ªLinha: Absoluto de Mau
115 for i=1:length(Hist_validacao_rede(1,:));
116     Hist_validacao_rede(4,i)=(Hist_validacao_rede(2,i))/NumValidBom; %4ªLinha:
            Relativo de Bom
117     Hist_validacao_rede(5,i)=(Hist_validacao_rede(3,i))/NumValidMau; %5ªLinha:
            Relativo de Mau
118 end; clear i;
```

```
Hist_validacao_rede(6,:)=cumsum(Hist_validacao_rede(2,:)); %6ªLinha: Acum(Absoluto de
    Bom)
Hist_validacao_rede(7,:)=cumsum(Hist_validacao_rede(3,:)); %7ªLinha: Acum(Absoluto de
    Mau)
Hist_validacao_rede(8,:)=cumsum(Hist_validacao_rede(4,:)); %8ªLinha: Acum(Relativo de
    Bom)
Hist_validacao_rede(9,:)=cumsum(Hist_validacao_rede(5,:)); %9ªLinha: Acum(Relativo de
    Mau)

%=====> KS
for i=1:length(Hist_treinamento_rede(1,:));
   KS_treinamento_rede=(max(Hist_treinamento_rede(8,:)-Hist_treinamento_rede(9,:)));
end; clear i;
for i=1:length(Hist_validacao_rede(1,:))
   KS_validacao_rede=(max(Hist_validacao_rede(8,:)-Hist_validacao_rede(9,:)));

end; clear i;

subplot(1,2,1)
plot(Hist_treinamento_rede(8,:),'LineWidth',1.5,'Color',[0 0 0]);hold all;
plot(Hist_treinamento_rede(9,:),'LineWidth',1.5,'Color',[.8 .8 .8]);
axis([0 100 0 1]);
ylabel('%'); xlabel('Score(x100)'); title('KS - Base de Treinamento');
legend('Adimplentes','Inadimplentes','Location','SouthEast');

subplot(1,2,2)
plot(Hist_validacao_rede(8,:),'LineWidth',1.5,'Color',[0 0 1]);hold all;
plot(Hist_validacao_rede(9,:),'LineWidth',1.5,'Color',[.8 .8 1]);
axis([0 100 0 1]);
ylabel('%'); xlabel('Score(x100)'); title('KS-Base de Validação');
legend('Adimplentes','Inadimplentes','Location','SouthEast');

%=====> ROC Curve
[Xd,Yd,T,AUC_Rede_Desenv] = perfcurve(treinamento_r,Rede_treinamento_res,1);
[Xv,Yv,T,AUC_Rede_Valid] = perfcurve(validacao_r,Rede_validacao_res,1);

plot(Xv,Yv,Xd,Yd,'LineWidth',2)
ylabel('Sensibilidade'); xlabel('1-Especificidade');
legend('Validação','Treinamento','Location','SouthEast')

%=====> Gini
Gini_Rede_Desenv = 2*(AUC_Rede_Desenv-0.5);
Gini_Rede_Valid = 2*(AUC_Rede_Valid-0.5);
```

Referências Bibliográficas

ALLEN, C. M.; ALLEN-DRISCOLL, C. Dynamic predictive credit scoring with multi-dimensional analysis of nearest neighbors. 2002.

BEALE MARTIN T. HAGAN, H. B. D. M. H. *Neural Network Toolbox 7 - User´s Guide*. [S.l.]: The MathWorks, Inc., 2010.

BELLOTTI, J. C. T. Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, v. 60, p. 1699–1707, 2009.

BREIMAN, L. et al. *Classification and Regression Trees*. [S.l.]: Wadsworth and Brooks, 1984.

CEAUSU, G. et al. Neuro-fuzzy classifiers for credit scoring. *Recent advances in management, marketing, finances*, p. 132–136, 2010.

CROUHY DAN GALAI, R. M. M. *The Essentials of Risk Management*. [S.l.]: McGraw-Hill, 2005.

DESAI, V. S. et al. Credit scoring models in the credit-union environment using neural networks and genetic algorithms. *Journal of Mathematics Appled in Business & Industry*, v. 8, p. 323–346, 1997.

DRAPER, N. R. D. N. R. *Applied Regression Analysis*. [S.l.]: Wiley Series in Probability and Statistics, 1998.

ERENGUC, G. J. K. S. S. Survey of mathematical programming models and experimental results for linear discriminant analysis. *Managerial and decision economics*, v. 11, p. 215–225, 1990.

FENSTERSTOCK, A. Credit scoring and the next step. *Business Credit*, p. 46–48, 2005.

FISHER, R. The use of multiple measurements in taxonomic problems. v. 7, 1936.

FRANK, A.; ASUNCION, A. Uci-machine learning repository. 2010. Disponível em: <http://archive.ics.uci.edu/ml>.

GAYLER, R. W. Classifier tecnology and the illusion of progress - credit scoring. *Statistical Science*, v. 21, p. 19–23, 2006.

GESTEL, T. V. et al. A support vector machine approach to credit scoring. 2003.

GIAMBONA, V. L. L. F. Survival models and credit scoring: some evidence from italian banking system. 2006.

GOCHET, W. et al. Multigroup dicsriminant analysis using linear programming. *Journal of operational research society*, v. 45, p. 213–225, 1997.

GREENE, W. H. *Econometric Analysis*. [S.l.]: Prentice Hall, 2002.

HAND, D. J. Statistical classification methods in consumer credit scoring: A review. *Royal Statistical Society*, v. 160, p. 523–541, 1997.

HANSSON, S. O. Seven myths of risk. *Risk Management: an international Journal*, v. 7, p. 7–17, 2005.

HASTIE, R. T. J. F. T. *The elements os Statistical Learning*. [S.l.]: Springer, 2009.

HAYKIN, S. *Neural Networks A comprehensive Foundation*. [S.l.: s.n.], 1999.

HOFFMANN, F. et al. Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. *European journal of operational research*, v. 177, p. 540–555, 2006.

HOFFMANN, R. *Análise de Regressão: Uma introdução a econometria*. [S.l.]: Editora HUCITEC, 2006.

HOLTON, G. Defining risk. *Financial Analysts Journal*, v. 60, p. 19–25, 2004.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. [S.l.]: Wiley Series in probability and statistics, 2000.

HUANG, Z. et al. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decison Support Systems*, v. 37, p. 543–558, 2004.

KANTARDZIC, M. *Data Mining: Concepts, Models, Methods, and Algorithms*. [S.l.]: Wiley-IEEE Press, 2002.

KEYNES, J. M. The general theory of employment. *The Quarterly Journal of Economics*, p. 209–223, 1937.

KLEIN, D. G. K. M. *Logistic Regression - A self-learning text*. [S.l.]: Springer, 2010.

KNIGHT, F. H. *Risk, Uncertainty and Profit*. [S.l.]: Houghton Mifflin, 1921.

KRUGMAN, P. *Microeconomics*. [S.l.]: Worth Publishers, 2004.

LAHSANA RAJA N. AINON, T. Y. W. A. Enhancement of transparency and accuracy of credit scoring models through genetic fuzzy classifier. *Maejo International Journal of Science and Technology*, v. 4, p. 136–158, 2010.

LAHSASNA, A. Credit scoring models using soft computing methods: A survey. *The International Arab Journal of Information Technology*, v. 7, p. 115–123, 2010.

LAI, K. Neural network metalearning for credit scoring. *ICIC*, p. 403–408, 2006.

LEWIS, E. M. *Introduction to Credit Scoring*. [S.l.]: Athena Press, 1992.

LIMA, J. C. *A importancia de conhecer a perda esperada para fins de gereciamento do risco de crédito*. [S.l.]: Revista BNDES, 2008.

MEYER, P. L. *Probabilidade, aplicações à Estatística*. [S.l.: s.n.], 1983.

ROSS, S. *Probabilidade - Um Curso Moderno com Aplicações*. [S.l.]: Bookman, 2010.

SICSÚ, A. L. *Credit Scoring: desenvolvimento, implantação, acompanhamento*. [S.l.]: Blucher, 2010.

SILVA, I. N. da; SPATTI, D. H.; FLAUZINO, R. A. *Redes neurais artificiais: para engenharia e ciências aplicadas*. [S.l.: s.n.], 2010.

SILVA, J. P. *Gestão e Análise de Risco de Crédito*. [S.l.]: Editora Atlas SA, 2000.

STEPANOVA, M. Survival analysis methods for personal loan data. *Operational Research*, v. 50, p. 277–289, 2002.

THOMAS, L. C. *Credit Scoring and its applications*. [S.l.]: Society for Industrial Mathematics, 2002.

THOMAS, L. C. *Consumer Credit Models: Pricing, Profit and Portfolios*. [S.l.]: Oxford University Press, 2009.

WORRACHARTDATCHAI, P. S. U. Credit scoring using least squares support vector machine based on data of thai financial institutions. *International Conference on Advanced Communication Technology*, v. 3, p. 2067–2070, 2007.

YAO, P. Feature selection based on svm for credit scoring. *International Conference on Computional Intelligence and Neural Computing*, p. 44–47, 2009.

ZWEIG, G. C. M. H. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *The American Association for Clinical Chemistry*, 1993.