Universidade Federal de Juiz de Fora
Departamento de Estatística
Curso de Estatística

Laura Leal Nunes

Aplicação do modelo de Regressão Logística para apoio à decisão de crédito

Juiz de Fora
2011

Laura Leal Nunes

Aplicação do modelo de Regressão Logística para apoio à decisão de crédito

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a obtenção do grau de Bacharel em Estatística.

Orientador: Camila Borelli Zeller

Doutora em Estatística - Universidade Estadual de Campinas

Juiz de Fora

2011

Nunes, Laura

Aplicação do modelo de Regressão Logística para apoio à decisão de crédito / Laura Nunes - 2011

57.p

1.Modelagem 2. Análise multivariada. I.Título.

CDU N/A

Laura Leal Nunes

Aplicação do modelo de Regressão Logística para apoio à decisão de crédito

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a obtenção do grau de Bacharel em Estatística.

Aprovado em

BANCA EXAMINADORA

Camila Borelli Zeller

Doutora em Estatística - Universidade Estadual de Campinas

Ronaldo Rocha Bastos

Doutor em Urban and Regional Planning - Liverpool University

Lupércio França Bessegato

Doutor em Estatística - Universidade Federal de Minas Gerais

À Deus e minha família.

Resumo

O sistema financeiro, a partir da aceleração do processo de globalização, passou a sofrer a influência dos mais diferentes fatores. Diante deste processo, no caso específico das instituições financeiras, surgiu a necessidade de analisar seus clientes para obter uma melhor seleção dos mesmos, com o objetivo de minimizar os seus riscos de inadimplência. Os métodos tradicionais de decisão para fornecer crédito a um indivíduo (ou uma empresa) em particular utilizavam julgamento humano, baseados em experiências de decisões anteriores. Entretanto, com o aumento de demanda de crédito resultante das pressões econômicas, aliadas a uma maior competição comercial e ao florescimento de novas tecnologias computacionais, têm-se conduzido ao desenvolvimento de sofisticadas técnicas estatísticas capazes de distinguir os proponentes a crédito como bom ou mau pagador - os sistemas de modelos de *Credit Scoring*. Dentre os muitos diferentes modelos citados na literatura e utilizados na prática, consta o tradicional modelo de regressão logística.

A construção de modelos de *Credit Scoring* está inserida no contexto de Data Mining, que compreende o processo de exploração, seleção e modelagem de grandes quantidades de dados para descobrir regularidades ou relações entre variáveis e, o modelo de regressão logística é uma das técnicas de classificação que destaca-se neste contexto.

Neste trabalho, serão analisados os resultados do *Credit Scoring* para segregar em grupos indivíduos (ou empresas) dignos e não dignos de crédito, fazendo uso do modelo de regressão logística inserido no contexto de Data Mining.

Palavras-chave: **Credit Scoring**, **Data Mining**, **Regressão Logística**.

Abstract

The financial system, from the acceleration of globalization, has come under the influence of many different factors. Thus, specifically, with financial institutions, the need to analyze clients to get a better selection of them, in order to minimize the risks of default has risen. Traditional methods of decision making to provide credit to an individual (or firm) used in particular professional expertises, based on experiences of previous decisions. However, the increased demand for credit resulting from economic pressures, combined with increased commercial competition and the blossoming of new computer technologies, have led to the development of sophisticated statistical techniques capable of distinguishing the applicants for credit as good or bad credit - the model systems of *Credit Scoring*. Among the many different models cited in the literature and used in practice, the importance of traditional logistic regression model can be acknowledge.

The construction of models of *Credit Scoring* is inserted in the context of data mining, which includes the process of exploration, selection, and modeling large amounts of data to discover regularities or relationships between variables and the logistic regression model is a technique of classification that stands out in this context.

This paper will consider the results of *Credit Scoring* to discriminate individuals into groups (or companies) worthy and not worthy of credit, using the logistic regression model within the context of Data Mining.

Keywords: **Credit Scoring**, **Data Mining**, **Logistic Regression**.

Agradecimentos

À Deus e a Nossa Senhora por estarem sempre presentes, me fazendo ter certeza de que mesmo longe da minha família, eu não estava sozinha. Eu não teria chegado até aqui se não fosse eu receber todos os dias força, tranquilidade e muita paz para lidar com as dificuldades.

Aos meus pais, que sempre me apoiaram e acreditaram em mim. Obrigada pelo carinho e pela paciência. Às minhas gordinhas, Ana Luisa e Gabriela, pelos momentos de alegria quando estamos juntas e também pelo apoio.

Aos meus inesquecíveis amigos da faculdade. À Priscila, por ter sido minha irmã durante todo esse tempo. À Carolina, pelo apoio, carinho, companheirismo e muitos momentos engraçadissimos. Ao Thiago, Samuel e Iago, pelas várias vezes que saimos juntos e nos divertimos bastante. Às meninas, Sarah, Raquel e Leiliane, por terem se mostrado sempre dispostas a ajudar.

Ao Bruno, meu beloved friend, que me aturou por todo esse tempo, nossos dias de iniciação científica foram os melhores! Ao Luis meu amigo irmão, pelas nossas conversas e pelo melhor abraço de todos. Ao Roberto, pelo carinho, atenção e pelos conselhos que com certeza fizeram diferença.

Ao Victor, por ter um coração maravilhoso! Obrigada pelo carinho, apoio e companheirismo.

Aos professores do departamento de Estatística, em especial à professora Camila, que mesmo com tão pouco tempo para a realização deste trabalho, se mostrou disposta a me orientar. Obrigada pela paciência e maravilhosa orientação. Ao professor Ronaldo, pelos 2 anos que trabalhamos juntos em Análise de Correspondência e ao professor Clécio, pelo apoio e paciência, suas aulas fizeram muita diferença.

Aos professores Lupércio e Ronaldo por terem aceitado fazer parte da minha banca.

"O homem planeja seu caminho, mas é Deus quem lhe dirige os passos".

Provérbios 16,9

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Motivação

A concessão de crédito por parte de instituições financeiras desempenha um papel fundamental no desenvolvimento de uma economia, em decorrência da dinâmica que introduz no processo econômico, seja uma oportunidade para as empresas aumentarem seus níveis de produção ou como estímulo para consumo dos indivíduos. Mudanças no cenário financeiro mundial, a partir dos anos 90, fizeram com que as instituições financeiras se preocupassem cada vez mais com o risco de crédito, que está associado à possibilidade do credor incorrer em perdas caso as obrigações assumidas por um tomador não sejam liquidadas nas condições pactuadas.

Os métodos tradicionais de decisões sobre concessão de crédito a clientes (empresas ou pessoas) eram fundamentados, até o início do século XX, em julgamentos humanos a partir de experiências do julgador em decisões anteriores e eram, portanto, bastante subjetivos e de agilidade insuficiente para grandes mercados de crédito. As pressões econômicas decorrentes da elevada demanda por crédito, a grande competição comercial do setor e o surgimento de novas tecnologias computacionais levaram ao desenvolvimento de modelos estatísticos para decisões de concessão de crédito, procurando torná-las mais objetivas e rápidas além de diminuir as perdas das carteiras de crédito. Estes modelos estatísticos são conhecidos na literatura como *Credit Scoring*. Dentre os muitos autores que já abordaram estes modelos, destacamos alguns, por exemplo, Hand & Henley (1997), Caouette et al.(1998), Vasconcellos (2002) e Chaia(2003).

É interessante ressaltar que a construção de modelos de *Credit Scoring* está inserida no contexto de Data Mining. Segundo Thomas et al. (2002), Data Mining tem a base de suas metodologias e técnicas estatísticas originadas em um problema de *Credit Scoring*, porém seu conceito vem sendo aplicado de forma mais abrangente. Data Mining pode ser definido como um processo para a descoberta de padrões e tendências em grandes conjuntos de dados visando formar conhecimento sobre um determinado fenômeno de interesse. O conceito de Data Mining está muito relacionado com a construção de

modelos, que é uma das principais formas de construir conhecimento sobre um evento ou fenômeno de interesse estabelecendo assim os fatores relacionados com os mesmos (Abreu, 2004).

Data Mining pode ser usada para vários tipos de tarefas, que são classificadas em preditivas e descritivas. Tarefas preditivas fazem uso de variáveis (características de interesse conhecidas) existentes na base de dados para predizer valores desconhecidos ou futuros e tarefas descritivas são voltadas para a busca e apresentação de padrões que descrevem os dados (Santos, 2006).

A classificação, uma das tarefas preditivas, busca identificar o conjunto mínimo das características conhecidas de um determinado conjunto de dados que seja suficiente para predizer uma característica desconhecida (Santos, 2006). O objetivo é predizer as características dos dados futuros, com base nos dados disponíveis. No caso de um problema de *Credit Scoring*, o interesse é conhecer os fatores relacionados ao risco de crédito de indivíduos interessados nos serviços prestados pelas instituições financeiras e, o objetivo é estimar a probabilidade de se observar o evento de interesse, como por exemplo a inadimplência, a fim de que o maior número possível de classificações (predições) corretas tanto de adimplentes, indivíduos que cumprem com suas obrigações financeiras, quanto de clientes inadimplentes, indivíduos que não cumprem com suas obrigações financeiras, possam ser obtidas.

1.2 Caracterização do Problema e Justificativa

Uma das principais dificuldades para a concessão de crédito é a identificação do risco de inadimplência ou não cumprimento das obrigações por parte dos proponentes do crédito. Isto decorre da dificuldade em estimar-se adequadamente a probabilidade dos proponentes virem a ser inadimplentes.

"Os ambientes de análise de crédito são caracterizados pela dinâmica na tomada de decisões e pela grande variedade de informações vindas das mais diversas fontes. Essas informações podem ser muitas vezes imcompletas, ambíguas, parcialmente incorretas ou de relevância dúbia. A forma subjetiva como se dá a análise dessas informações faz com que não se consiga explicar o processo de tomada de decisões embora seja sabido que existem fatores que influenciam essas decisões". (Senger e Caldas Junior, 2001, p.19)

Senger e Caldas Junior (2001) afirmam que as decisões de crédito devem ser criteriosas, pois podem provocar prejuízos às instituições financeiras, além de prejuízos morais aos clientes. Quanto mais amplas e precisas as informações sobre a probabilidade de pagamento, melhores serão as condições para a correta avaliação do risco de cada operação. Com isso, as instituições financeiras terão maior propensão a emprestar, os custos dos empréstimos serão menores e todos os agentes econômicos obterão condições de realizar melhores resultados.

Dessa forma, a justificativa para a realização deste trabalho deve-se principalmente ao interesse e pretensão de trabalhar no contexto de análise de crédito, mais especificamente, risco de crédito. Além disso, este trabalho pode servir de motivação para que outros alunos demonstrem interesse em explorar esta mesma linha temática aqui abordada.

1.3 Objetivos

O principal objetivo deste trabalho é analisar os resultados do *Credit Scoring* para segregar em grupos, indivíduos (ou empresas) dignos (adimplentes) e não dignos (inadimplentes) de crédito, fazendo uso das técnicas de Data Mining e em particular do modelo de regressão logística. Em outras palavras, neste trabalho, estamos interessados em construir um modelo preditivo que apóie efetivamente a decisão sobre o risco de crédito.

Dessa forma, podemos relacionar os seguintes objetivos específicos: (i) Explorar como se planeja e se realiza cada uma das etapas e atividades necessárias para a criação de um modelo preditivo para a decisão de crédito a partir de dados históricos de operações de crédito realizadas por uma instituição financeira; (ii) Avaliar a capacidade preditiva do modelo mediante as medidas de sensibilidade e de especificidade, determinando a área sob a curva característica de operação (ROC - *Receiver Operating Characteristic*).

Apesar de, neste trabalho, a contrução do modelo de *Credit Scoring* ser desenvolvido com o uso da técnica de regressão logística, que é um dos métodos usados frequentemente pelas instituições financeiras, inúmeras técnicas podem ser desenvolvidas para classificar em grupos, indivíduos (ou empresas) dignos e não dignos de crédito, dentre elas se destacam, por exemplo, análise discriminante (Hand et al., 1998), regressão linear (Orgler, 1970), modelos probito (Grablowsky and Talley, 1981), árvores de decisão

(Arminger et al., 1997), redes neurais (West, 2000), análise de sobrevivência (Abreu, 2004) e análise de correspondência (Saporta, 2003).

1.4 Apresentação dos capítulos

Neste capítulo, é apresentada toda a estrutura e objetivos do trabalho, além da justificativa do mesmo.

No capítulo 2, são apresentados aspectos teóricos relativos ao contexto de análise de crédito. Neste capítulo, é realizada uma revisão de literatura sobre os termos que serão abordados ou citados ao longo do trabalho, tais como, crédito, risco de crédito e *Credit Scoring*.

No capítulo 3, é apresentada a metodologia que será aplicada no contexto de análise de crédito, ou seja, apresentamos o modelo de regressão logística, incluindo o método de Newton-Raphson para estimação por máxima verossimilhança e a matriz informação de Fisher, últil no cálculo dos desvios padrões dos estimadores dos parâmetros do modelo. Além disso, são apresentados aspectos básicos das medidas usualmente empregadas para avaliar a capacidade preditiva do modelo em estudo, em termos de sensibilidade, especificidade e da área sob a curva característica de operação.

No capítulo 4, apresentamos a aplicação da metodologia discutida em um conjunto de dados reais. Finalmente, no capítulo 5, são apresentadas as considerações finais do tema discutido.

2 Revisão de literatura

2.1 Data Mining (Mineração de Dados)

A descoberta de conhecimento em base de dados é um campo de pesquisa que tem crescido rapidamente e cujo desenvolvimento tem sido dirigido ao benefício de necessidades práticas, sociais e econômicas. Considerando, em geral, que as bases contêm muitos dados, tornou-se necessário o desenvolvimento de processos de análise automática, como o processo de Data Mining (Rezende et al. 2003).

Segundo Fayyad et al. (1996), Data Mining é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados. Seus objetivos principais são a predição e a descrição.

Os dois principais tipos de tarefas para predição são classificação e regressão. A diferença básica é que enquanto a classificação prediz valores discretos (classes), a regressão modela funções contínuas. A classificação, tipo de tarefa que estamos interessados, consiste na predição de um valor categórico como, por exemplo, predizer se o cliente é digno de crédito (bom pagador) ou se ele é não digno (mau pagador). Uma vez estabelecida a tarefa, existe uma variedade de algoritmos para executá-la, dentre eles estão os algoritmos indutores de árvores de decisão ou regra de produção, modelos lineares, modelos não-lineares (redes neurais artificiais e regressão logística) e modelos de dependência probabilística (redes bayesianas).

Na modelagem da predição de inadimplência são utilizados um ou mais algoritmos de aprendizagem com o propósito de identificar um classificador que seja mais apropriado para o relacionamento entre o conjunto de variáveis (dados contábeis) e o rótulo da classe dos dados de entrada (indivíduos ou empresas dignos e não dignos de crédito). Esse modelo gerado pelo algoritmo deve se adaptar bem aos dados de entrada e classificar corretamente os rótulos de classes de registros desconhecidos. Portanto, o objetivo é construir modelos com boa capacidade de generalização, isto é, modelos que possam predizer com precisão os rótulos de classes de registros não conhecidos previamente (Han & Kamber, 2006). No nosso caso, iremos utilizar apenas um algoritmo, o

modelo de regressão logística.

Segundo Rezende et al. (2003), o processo de Data Mining é um ciclo dividido em três grandes etapas: pré-processamento, extração de padrões e pós-processamento, precedidas pela fase de identificação do problema, e sucedidas pela fase de utilização do conhecimento.

2.1.1 Identificação do problema

Nesta etapa, o estudo do domínio da aplicação e a definição dos objetivos a serem alcançados no processo de Data Mining são identificados. A análise inicial para definição dos principais objetivos e restrições, o conhecimento sobre o domínio, fornece um subsídio para todas as etapas em seqüência.

2.1.2 Pré-processamento

Normalmente, os dados disponíveis para análise não estão em um formato adequado para serem utilizados, em razão de limitações de memória ou tempo de processamento, muitas vezes não é possível a aplicação direta dos algoritmos de extração de padrões aos dados. Dessa maneira, torna-se necessária a aplicação de métodos para tratamento, limpeza e redução do volume de dados.

Redução do volume de dados está relacionada à obtenção de uma amostra aleatória, redução do número de variáveis (ajuste de modelos estatísticos) e redução do número de categorias (categorização) de uma variável. Limpeza é a correção de possíveis problemas advindos do processo de coleta (erros de digitação, erros de leitura, dentre outros) e, por último, tratamento da base de dados é deixá-la pronta para ser trabalhada.

2.1.3 Extração de padrões

A etapa de extração de padrões é direcionada ao cumprimento dos objetivos definidos anteriormente, ou seja, nesta etapa é feita a escolha da tarefa de Data Mining conforme os objetivos desejáveis para a solução procurada (Motta, 2004). Ela consiste na aplicação dos algoritmos de Data Mining escolhidos para a extração dos padrões embutidos nos dados.

2.1.4 Pós-processamento

Nesta etapa, são avaliados o desempenho e a qualidade dos padrões extraídos, bem como verificada a facilidade de interpretação dessas regras (Motta, 2004).

As medidas para avaliação do conhecimento podem ser denominadas medidas de desempenho, dentre elas estão precisão, erro, confiança negativa, sensibilidade, especificidade, cobertura, suporte, satisfação, velocidade e tempo de aprendizado (Lavrac et al., 1999).

Neste trabalho, estamos interessados, no contexto de risco de crédito, portanto nas medidas de desempenho sensibilidade, que é a proporção de acerto na predição de clientes adimplentes nos casos em que eles de fato são adimplentes, e especificidade, que é a proporção de acerto na predição de clientes inadimplentes nos casos em que eles de fato são inadimplentes, em particular. Abaixo alguns conceitos importantes relacionados a risco de crédito que serão abordados ao longo do texto.

2.2 Crédito

Dentre os diversos conceitos para o termo crédito, é importante conhecer primeiro seu sentido etimológico. A palavra *crédito* vem do latim *creditu* significando acredito ou confio. Para Silva (2000), crédito representa a entrega do bem presente mediante uma promessa de pagamento. Segundo Santos (2003), crédito refere-se à troca de um valor presente por uma promessa de reembolso futura, não necessariamente certa, em virtude do fator risco.

A oferta de crédito por parte de instituições financeiras desempenha o papel importante de impulsionar a atividade econômica, pois, disponibiliza recursos financeiros às empresas e pessoas físicas para que possam financiar suas necessidades permanentes e eventuais. Conforme Perera (1998), a história do crédito demonstra que sua evolução acompanhou o próprio desenvolvimento econômico da sociedade procurando desenvolver instrumentos necessários para satisfação das necessidades e anseios da humanidade. Quando acontece uma concessão de recursos, a instituição financeira passa a possuir o chamado risco de crédito. Risco é um elemento inerente ao crédito e inseparável deste. Assim, não existe uma operação de crédito sem risco.

2.3 Riscos de crédito

Caouette et al. (1998) assim definem risco de crédito: se o crédito constitui-se na expectativa de entrada de uma determinada quantia no caixa dos credores, em data futura, então o risco de crédito é a probabilidade de que esta expectativa não se cumpra. O risco está associado à possibilidade do credor incorrer em perdas caso as obrigações assumidas por um tomador não sejam liquidadas ou pela deterioração da qualidade de crédito do tomador. A deterioração da qualidade de crédito não resulta em uma perda imediata para o credor e sim no incremento da probabilidade de que um evento de *default* venha a ocorrer. O evento de *default*, neste trabalho, é a perda total esperada, ou seja, o não pagamento da dívida por parte do tomador.

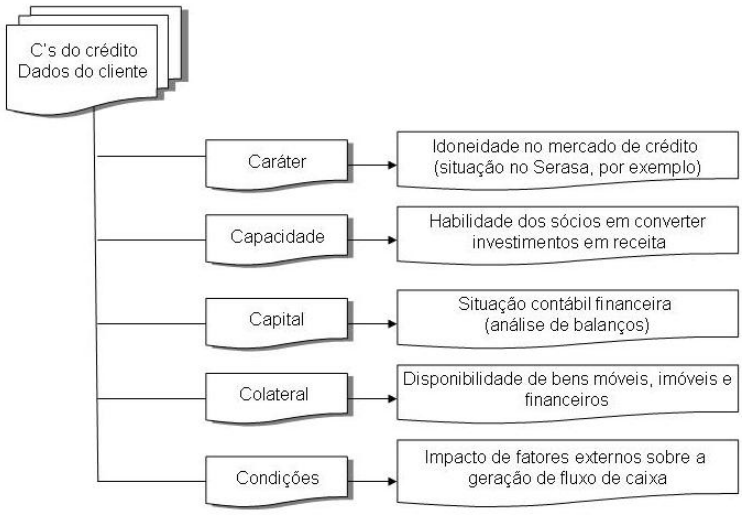
Para a minimização do risco de crédito destaca-se cada vez mais a importância à gestão do risco, baseada em procedimentos subjetivos (análise caso a caso) e objetivos (análise estatística).

2.3.1 Análise subjetiva

Para Santos (2003), a análise subjetiva, ou caso a caso, é baseada na experiência adquirida dos analistas de crédito, no conhecimento técnico, no bom senso e na disponibilidade de informações (internas e externas) que lhes possibilitem diagnosticar se o cliente possui idoneidade e capacidade de gerar receita para honrar o pagamento das parcelas dos financiamentos, por exemplo, a análise dos 5 C's do crédito (Caráter, Capacidade, Capital, Colateral, Condições) que constitui-se na principal ferramenta da análise subjetiva, sendo ele o modelo mais tradicional de organização das informações sobre a possibilidade de pagamento de um cliente. Os conceitos de cada componente do modelo estão resumidos na Figura 2.1.

Os métodos de tomada de decisões baseadas em critérios subjetivos têm perdido espaço nas atividades de crédito, e torna-se cada vez maior a busca por instrumentos mais eficazes para gerenciamento da exposição ao risco de crédito. Estes instrumentos estão relacionados à análise objetiva do crédito.

Figura 2.1: C's do crédito



Fonte: SANTOS, 2003:44

2.3.2 Análise objetiva

A análise objetiva é amparada em pontuações estatísticas de risco com a finalidade de apurar resultados que atestem a capacidade de pagamento dos proponentes de crédito. Segundo Santos (2003), a pontuação de crédito é um instrumento estatístico desenvolvido para que o analista avalie a probabilidade de que determinado cliente venha a tornar-se inadimplente no futuro. Rosenberg & Gleit (1994) comentam que há muitas vantagens na utilização de métodos quantitativos em gerenciamento de crédito, destacando-se os benefícios resultantes da otimização no processo de tomada de decisão. Dentre as técnicas de análise objetiva do risco de crédito, destacam-se os modelos de *Credit Scoring*.

Credit Scoring

A nomenclatura *Credit Scoring* é usada para descrever o processo formal de estimar as probabilidades de candidatos a crédito atrasarem ou não o pagamento do compromisso financeiro que pretendem assumir (Vasconcellos, 2002). Trata-se de modelos que, a partir do histórico de concessões de crédito efetuadas por uma instituição financeira, identificam através de técnicas estatísticas, as principais variáveis que indicam na capacidade do cliente em pagar o crédito, permitindo que estes sejam classificados em grupos distintos e, como conseqüência, a decisão sobre aceitação ou não da concessão do crédito.

O objetivo é desenvolver uma ferramenta para classificar novas operações de

crédito de acordo com suas possibilidades de serem operações boas ou ruins conforme o critério escolhido, que usualmente é relacionado às probabilidades de inadimplência no pagamentos das prestações (Vasconcellos, 2002).

Quando um cliente solicitar um crédito, o mesmo deverá fornecer suas variáveis cadastrais e financeiras que, unidas às variáveis da operação, poderão lhe gerar um *score*. Esse *score* poderá, então, ser utilizado na decisão de conceder ou não o crédito ao cliente, a partir do momento que se define um *score* de corte, acima do qual o pedido do cliente será aceito.

Há vários métodos para a escolha do ponto de corte, dentre eles estão o uso da programação linear inteira (Dantas & DeSouza, 2007) e a análise de perda e ganho em função da rejeição de clientes dignos de crédito (Chaia, 2003), por exemplo. Neste trabalho, será feito por meio da curva ROC *(Receiver Operating Characteristic)* que constitui uma técnica útil para avaliar modelos de risco de crédito e está baseada nos conceitos de sensibilidade e especificidade.

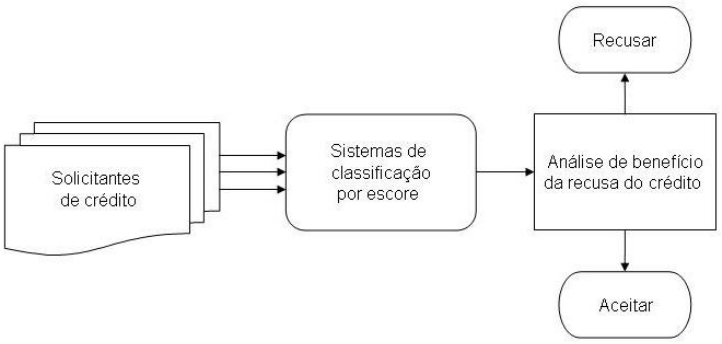
Os modelos de *Credit Scoring* podem ser aplicados tanto à análise de crédito de pessoas físicas quanto empresas. Para pessoas físicas são utilizadas informações cadastrais e de comportamento. Já quando para empresas, são utilizadas demonstrações contábeis. Segundo Saunders (2000), os modelos de *Credit Scoring* podem ser encontrados em praticamente todas as formas de análise de crédito, desde avaliações para concessão de crédito direto ao consumidor até empréstimos comerciais.

Segundo Sicsu (1999), o desenvolvimento de um modelo de *Credit Scoring* compreende as seguintes etapas:

- Planejamento e definições;
- Identificação de variáveis potenciais;
- Planejamento amostral;
- Determinação do escore: aplicação da metodologia estatística;
- Determinação do ponto de corte ou faixas de escore;
- Determinação de regra de decisão;
- Validação e verificação de performance do modelo estatístico;

A figura 2.2 ilustra o método de concessão de crédito por modelos de *Credit Scoring*.

Figura 2.2: Método de concessão de crédito usando os modelos de *Credit Scoring*



Fonte: Chaia (2003, p. 30)

Os modelos de *Credit Scoring* são divididos em duas categorias: modelos de aprovação de crédito e modelos de escoragem comportamental, conhecido como *Behavioural Scoring* (Saunders, 2000). Os modelos de *Credit Scoring* dão suporte à tomada de decisão sobre a concessão ou não do crédito, já os modelos *Behavioural Scoring* auxiliam na administração dos créditos já existentes, ou seja, aqueles clientes que já possuem uma relação creditícia com a instituição. Desta forma, enquanto o objetivo dos modelos de aprovação de crédito é estimar a probabilidade de um novo cliente se tornar inadimplente, os modelos de escoragem comportamental objetivam estimar a probabilidade de inadimplência de um cliente que já possui crédito com a instituição.

Assim, como qualquer modelo de aprovação de crédito, os modelos de *Credit Scoring* possuem suas vantagens e desvantagens. Caouette et. al (1998) e Parkinson & Ochs (1998) resumem as principais vantagens dos modelos de *Credit Scoring*, dadas por:

- Consistência: são modelos que utilizam a experiência da instituição e servem para administrar objetivamente os créditos dos clientes já existentes e dos novos solicitantes;
- Facilidade: os modelos tendem a ser simples e de fácil interpretação;
- Melhor organização da informação do crédito: a sistematização e organização das informações contribuem para a melhoria do processo de concessão do crédito;

- Redução da metodologia subjetiva: o uso do método quantitativo com regras claras e bem definidas contribui para a diminuição do subjetivismo na avaliação do risco de crédito;

- Maior eficiência do processo: o uso de modelos *Credit Scoring* na concessão de crédito direciona os esforços dos analistas, acarretando redução de tempo e maior eficiência a este processo.

Os autores, além das vantagens também resumem as principais desvantagens, dadas por:

- Custo de desenvolvimento: desenvolver um sistema *Credit Scoring* pode acarretar custos, não somente com o sistema em si, mas também com o suporte necessário para a sua construção, como por exemplo, profissionais capacitados, equipamentos, coleta de informações necessárias para o desenvolvimento do modelo, dentre outros;

- Excesso de confiança nos modelos: algumas estatísticas podem superestimar a eficiência dos modelos, fazendo com que usuários menos experientes, considerem tais modelos perfeitos, não criticando seus resultados;

- Falta de dados oportunos: se o modelo necessita de dados que não foram informados, pode haver problemas na sua utilização, gerando resultados diferentes dos esperados. Além da falta de algumas informações necessárias, faz-se necessário também a qualidade e fidedignidade das informações disponíveis, uma vez que elas representam o insumo principal dos modelos;

- Interpretação equivocada dos escores: o uso inadequado do sistema de *Credit Scoring* devido à falta de treinamento e aprendizagem de como utilizar suas informações pode ocasionar problemas sérios às instituições.

Conforme Carmona & Amorin Neto (2002), enquanto os modelos de aprovação de crédito se preocupam apenas com a concessão e o volume de crédito, os modelos de escoragem comportamental podem ser utilizados para estabelecer os limites de crédito, cobrança preventiva, dentre outras estratégias. Inserido no contexto dos modelos de escoragem comportamental, está o *rating* de crédito, que será abordado a seguir.

2.4 *Rating* de crédito

Rating é uma classificação de risco de crédito que pode ser atribuída a um país, a uma empresa, a uma pessoa, a um título ou a uma operação de crédito. Os *ratings* são elaborados por agências de *rating* de crédito cuja especialidade é avaliar risco de crédito. Segundo a Standard e Poor's, uma das principais agências de *rating* mundiais, *rating* de crédito são opiniões sobre risco de crédito, opiniões sobre a capacidade e a vontade de um emissor de honrar suas obrigações financeiras, integralmente e no prazo determinado. Além disso, podem refletir a qualidade de crédito de um título de dívida individual e a probabilidade relativa de inadimplência dessa emissão.

Cada agência aplica sua própria metodologia para medir a qualidade de crédito e usa uma escala específica para publicar opiniões de *rating*. Normalmente, os *ratings* são expressos por meio de letras que variam, por exemplo, de "AAA"a "D"para comunicar a opinião da agência sobre o nível relativo de risco de crédito. No nosso país, as instituições financeiras são obrigadas a ter um sistema de classificação de risco, uma vez que a Resolução 2682/1999 do Conselho Monetário Nacional (CMN) determina que as operações de crédito concedidas pelas referidas instituições devam ser classificadas em níveis de risco, que varia segundo uma escala com nove classes entre AA e H, sendo "AA"o *rating* excelente e "H"o *rating* péssimo, que são os tomadores classificados exclusivamente segundo a perspectiva de perda.

Em geral, os *ratings* não são recomendações para comprar, manter ou vender, ou mesmo uma medida do valor do ativo, nem tem a intenção de sinalizar a adequação de um investimento. Eles apenas tratam de um aspecto da decisão de investimento, a qualidade de crédito, e em alguns casos, podem mostrar o grau de recuperação esperado de investimento em uma situação de *default*. Uma vez que eventos e desenvolvimentos futuros não são previsíveis, a atribuição de um *rating* de crédito não é uma ciência exata, consiste em expressar opiniões relativas sobre a qualidade de crédito de um emissor ou qualidade de crédito de uma emissão em particular, indo da mais forte a mais fraca, dentro de um universo de risco de crédito (Standard & Poor's, 2008).

A tabela abaixo ilustra a escala de *rating* adotada pela Standard & Poor's.

Tabela 2.1: Resumo da escala de *rating* adotada pela Standard & Poor's

Grau de investimento	AAA	Capacidade extremamente forte para honrar compromissos
	AA	Capacidade muito forte para honrar compromissos
	A	Forte capacidade para honrar compromissos financeiros, porém é de alguma forma suscetível a condições econômicas adversas e a mudanças circunstanciais
	BBB	Capacidade adequada para honrar compromissos, porém mais sujeito a condições econômicas adversas
	BBB-	Considerado o nível mais baixo da categoria de grau de investimento pelos participantes do mercado
Grau especulativo	BB+	Considerado o nível mais alto da categoria de grau especulativo pelos participantes do mercado
	BB	Menos vulnerável no curto prazo, porém enfrenta atualmente grande suscetibilidade a condições adversas de negócios, financeiras e econômicas
	B	Mais vulnerável a condições adversas de negócios, financeiras e econômicas, porém atualmente apresenta capacidade para honrar compromissos financeiros
	CCC	Atualmente vulnerável e dependente de condições favoráveis de negócios, financeiras e econômicas para honrar seus compromissos financeiros
	CC	Atualmente fortemente vulnerável
	C	Um pedido de falência foi registrado ou ação similar empetrada, porém os pagamentos das obrigações financeiras continuam sendo realizados
	D	Inadimplente em seus compromissos financeiros

Grau de investimento é utilizado para descrever emissores e emissões, pessoas ou empresas que apresentam níveis relativamente elevados de capacidade e qualidade creditícia, já o termo grau especulativo refere-se a títulos de dívida em que o emissor, pessoa ou empresa atualmente tem a capacidade de honrar, porém enfrenta incertezas, tais como circunstâncias financeiras ou de negócio que poderiam afetar seu risco de crédito.

Uma maneira de monitorar o risco das carteiras de crédito é utilizar um sistema chamado matriz de migração, gerada a partir dos sistemas de classificação de risco. Esses sistemas atribuem uma medida que representa a expectativa de risco de *default* associada ao tomador, essa medida é o *rating* de crédito. As matrizes de migração evidenciam as alterações na qualidade de crédito dos tomadores de recurso ao longo de um determinado período de tempo.

A construção da matriz envolve a seleção de um conjunto de empresas ou indivíduos em uma determinada data (data inicial) e a situação das mesmas em uma segunda data (data final). Para comparar a situação final e inicial utilizam-se os ratings, classificação de risco. Há três possíveis classificações de risco que os tomadores podem migrar durante o período da matriz: (i) Tomador se mantém na mesma classificação de risco; (ii) Tomador migra para classificação de risco melhor (*upgrade*); (iii) Tomador migra para classificação de risco pior (*downgrade*).

A matriz de migração está intimamente ligada aos modelos de escoragem comportamental (*Behavioural Scoring*), pois, como já citado anteriormente, estes modelos auxiliam na administração dos créditos já existentes na instituição financeira e a matriz monitora estes créditos ao longo do tempo, verificando a deterioração ou a melhora na qualidade do crédito.

Quanto maior o risco do tomador, pior será seu *rating*, consequentemente, mais restritivas serão as condições sob as quais a instituição concederá crédito, principalmente em relação a volume, prazo, taxa de juros e garantias.

3 Metodologia

3.1 Introdução

A análise de regressão é uma técnica estatística cujo objetivo consiste em descrever a relação entre uma variável resposta e um conjunto de variáveis explicativas através de um modelo. Em algumas situações, a variável resposta pode ser dicotômica ou binária, isto é, aquela que apresenta duas possibilidades de resposta (sucesso ou fracasso), como por exemplo, a variável de adimplência de uma empresa ou indivíduo, sendo inadimplência o sucesso e adimplência o fracasso. No contexto de análise de crédito, o modelo de regressão logística tem sido muito utilizado, para mais detalhes veja, por exemplo, Brito et al.(2009), Dantas & Desouza (2008), Santos (2008) e trabalhos de Santos & Famá (2006).

3.2 Regressão Logística

De acordo com Hosmer & Lemeshow (1989), o modelo de regressão logística é definido como:

$$\mathbf{Y_i} = \pi(x_i) + \epsilon_i, \tag{3.1}$$

onde ϵ_i é o erro aleatório e assume-se que $Y_i \sim Ber(\pi(x_i))$, ou seja, a variável resposta assume o valor 1 para o acontecimento de interesse (sucesso) e o valor 0 para o acontecimento complementar (fracasso), com probabilidades $\pi(x_i) = P(Y = 1|x_i)$ e $1 - \pi(x_i) = P(Y = 0|x_i)$, respectivamente. A probabilidade de sucesso do modelo de regressão logística é dada por:

$$\pi(x_i) = \pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}}}, \tag{3.2}$$

e a probabilidade de fracasso:

$$1 - \pi(x_i) = 1 - \pi_i = \frac{1}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}}}. \tag{3.3}$$

Em qualquer problema de regressão, a quantidade a ser modelada é o valor esperado da variável resposta dados os valores das variáveis explicativas, ou seja, $E(Y|x_i)$. No modelo

de regressão logística, devido à natureza da variável resposta, temos que

$$0 \leq E(Y|x_i) = 1P(Y_i = 1|x_i) + 0P(Y_i = 0|x_i) = \pi_i \leq 1.$$

Além disso, devido à natureza da variável resposta, temos que a quantidade ϵ_i pode assumir somente um de dois possíveis valores, isto é, $\epsilon_i = 1 - \pi_i$, para $y_i = 1$ ou $\epsilon_i = -\pi_i$ para $y_i = 0$. Dessa forma, segue que ϵ_i tem distribuição com média zero e variância $\pi_i(1 - \pi_i)$.

A transformação de π_i que será central para o estudo do modelo de regressão logística é denominada transformação logito. Esta transformação é definida por:

$$g(x_i) = ln\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}. \tag{3.4}$$

A importância dessa transformação é que o logito é linear em β_0 e β_j, $j = 1, ..., p$, pode ser contínuo e variar de $-\infty$ a $+\infty$, dependendo dos valores assumidos pelas variáveis explicativas. O logito pode ser interpretado como o logaritmo da chances entre π_i e $1 - \pi_i$.

3.2.1 Estimação por máxima verossimilhança

Supondo que (x_i, y_i) seja uma amostra independente com n pares de observações, y_i representa o valor observado da variável resposta dicotômica e x_i é o valor observado da variável explicativa da i-ésima observação em que $i = 1, ..., n$. Para o ajuste do modelo de regressão logística, segundo a equação (3.1 a 3.3), é necessário estimar os parâmetros β_0 e β_j, $j = 1, ..., p$.

O método de máxima verossimilhança é utilizado para estimar os parâmetros. A função de distribuição da probabilidade de Y_i para o modelo de regressão logística com $Y_i \sim Ber(\pi(x_i))$ é dada por:

$$f(y_i, \pi_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}, y_i = 0 \quad ou \quad y_i = 1. \tag{3.5}$$

Uma vez que $Y_1, Y_2, ..., Y_n$ são independentes, a função de verossimilhança é obtida pelo produto dos termos dados na expressão (3.5) e é definida por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}, \tag{3.6}$$

onde denota-se por $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ o vetor de parâmetros desconhecidos. Então, a função de log verossimilhança é dada por:

$$l(\boldsymbol{\beta}) = ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{n} y_i ln[\pi(x_i)] + (1 - y_i) ln[1 - \pi(x_i)]. \tag{3.7}$$

O príncipio da máxima verossimilhança é obter o valor de $\boldsymbol{\beta}$ que maximize $L(\boldsymbol{\beta})$ ou equivalentemente $l(\boldsymbol{\beta})$. Dessa forma, deriva-se $l(\boldsymbol{\beta})$ em relação a cada parâmetro, obtendo o seguinte sistema de equações:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(x_i)] = 0, \tag{3.8}$$

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} [y_i - \pi(x_i)] = 0, \tag{3.9}$$

onde $\pi(x_i)$ é dada pela expressão (3.2).

Note que as equações em (3.8) e (3.9) são não lineares em β_0 e $\beta_j, j = 1, \ldots, j$, assim são necessários métodos iterativos para resolução do sistema de equações. Para encontrar o valor de $\boldsymbol{\beta}$ que maximiza $l(\boldsymbol{\beta})$, foi utilizado o método iterativo de Newton Raphson.

O conjunto de equações iterativas do método de Newton Raphson é assim definido:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{(t+1)} &= \widehat{\boldsymbol{\beta}}^{(t)} + [\mathbf{I}(\widehat{\boldsymbol{\beta}})^{(t)}]^{-1} \mathbf{U}(\widehat{\boldsymbol{\beta}}^{(t)}) \\ &= \widehat{\boldsymbol{\beta}}^{(t)} + [\mathbf{X}^t \widehat{V}^{(t)} \mathbf{X}]^{-1} \mathbf{X}^t (\mathbf{Y} - \widehat{\boldsymbol{\pi}}^{(t)}), \quad t = 0, 1, \ldots \end{aligned} \tag{3.10}$$

sendo que $\widehat{\boldsymbol{\beta}}^{(t)}$ e $\widehat{\boldsymbol{\beta}}^{(t+1)}$ são os estimadores de $\boldsymbol{\beta}$ nos passos t e $t+1$, respectivamente, $\mathbf{U}(\boldsymbol{\beta})$ é o vetor escore e $\mathbf{I}(\boldsymbol{\beta})$ é a matriz de informação de Fisher, que fornece os erros padrão para as estimativas dos parâmetros, onde

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{Y} - \mathbf{X}^T \boldsymbol{\pi} = \mathbf{X}^T (\mathbf{Y} - \boldsymbol{\pi})$$

e

$$\mathbf{I}(\boldsymbol{\beta}) = E\left[-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right] = \mathbf{X}^T V \mathbf{X}.$$

Denota-se por $\mathbf{X}$ a matriz de dimensão $n \times (p + 1)$ que contém os valores das variáveis explicativas, $\widehat{V}$ uma matriz diagonal de dimensão $n \times n$ que contém os elementos $\widehat{\pi}_i(1 - \widehat{\pi}_i)$,

$Y = (y_1, y_2, ..., y_n)^T$ e $\pi = (\pi_1, \pi_2, ..., \pi_n)$, ou seja

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} e \quad \widehat{\boldsymbol{V}} = \begin{bmatrix} \widehat{\pi}_1(1-\widehat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \widehat{\pi}_2(1-\widehat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widehat{\pi}_n(1-\widehat{\pi}_n) \end{bmatrix}.$$

As iterações são repetidas até que uma regra de convergência adequada seja satisfeita. Como critério de convergência, podemos utilizar $\parallel \widehat{\beta}^{(t+1)} - \widehat{\beta}^{(t)} \parallel < \epsilon$ ou $\parallel \frac{\widehat{\beta}^{(t+1)} - \widehat{\beta}^{(t)}}{\beta^{(t)}} \parallel < \epsilon$, onde $\parallel a \parallel$ indica a norma do vetor a e $\epsilon > 0$. Veja Allison (1999) para mais detalhes sobre critérios de convergência. Valores iniciais são necessários para implementar este algoritmo. Em geral, considera-se $\boldsymbol{\beta}^{(0)} = \mathbf{0_{(p+1)}}$, vetor de zeros de dimensão $(p+1) \times 1$ (Souza, 2006).

A partir dos parâmetros estimados pode-se predizer a probabilidade de novos candidatos a crédito, por exemplo, serem inadimplentes, sendo esses valores preditos utilizados para a aprovação ou não da concessão do crédito.

Além da utilização das estimativas dos parâmetros na predição do potencial de risco de novos candidatos a crédito, os estimadores dos parâmetros nos fornecem também a informação, através do seu nível de significância, quais variáveis explicativas estão mais associadas com o evento que está sendo modelado, no caso a inadimplência ou adimplência, ajudando assim na compreensão e interpretação do mesmo (Abreu, 2004).

3.2.2 Testes de Hipóteses e Intervalos de Confiança

Intervalos de confiança assintóticos e testes de hipóteses para o vetor de parâmetros $\boldsymbol{\beta}$ podem ser obtidos, considerando que o estimador de máxima verossimilhança, $\widehat{\boldsymbol{\beta}}$, tem aproximadamente distribuição $N_p(\boldsymbol{\beta}, \mathbf{I}^{-1}(\boldsymbol{\beta}))$. Na prática, geralmente $\mathbf{I}(\boldsymbol{\beta})$ é desconhecida e a substituimos por $\mathbf{I}(\widehat{\beta})$ que é a matriz $\mathbf{I}(\boldsymbol{\beta})$ avaliada na estimativa de máxima verossimilhança de $\boldsymbol{\beta}$.

Testes de hipóteses

Usualmente no modelo de regressão logística são feitos dois conjuntos de testes de hipóteses com finalidades distintas. O primeiro é feito para a escolha do modelo, onde o objetivo é testar se uma variável explicativa ou um conjunto delas têm coeficiente igual à zero.

Testar a significância de um conjunto de variáveis explicativas consiste essencialmente em verificar se o modelo que as inclui revela o melhor ajuste que o modelo em que elas não são incluidas (Carballo, 2002). Depois de escolhido o modelo, outro conjunto de testes de hipóteses pode ser utilizado na verificação da qualidade do ajuste, visando aferir a qualidade de ajuste do modelo.

Para o primeiro conjunto de testes de hipóteses, referente à escolha do modelo, existem várias estatísticas para testar a hipótese de que os coeficientes $\beta_i's$, correspondentes às q variáveis retiradas do modelo, são iguais a zero (Demétrio, 2002). Estaremos interessados em trabalhar, por exemplo, com duas dessas estatísticas. O interesse está em testar as hipóteses:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = q_0 \quad vs \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq q_0,$$

onde $\mathbf{C}$ é uma matriz de dimensão $l \times (p+1)$, l é o posto de $\mathbf{C}$ e q_0 é um vetor conhecido de dimensão l. Logo, as estatísticas Razão de Verossimilhança e Wald podem ser escritas como

$$\Lambda = -2ln\left[\frac{L(\widetilde{\boldsymbol{\beta}})}{L(\widehat{\boldsymbol{\beta}})}\right] = 2[l(\widehat{\boldsymbol{\beta}}) - l(\widetilde{\boldsymbol{\beta}})], \tag{3.11}$$

$$W = (\mathbf{C}\widehat{\boldsymbol{\beta}} - q_0)^T(\mathbf{C}\mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{C}^T)^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}} - q_0), \tag{3.12}$$

onde $\widehat{\boldsymbol{\beta}}$ é o vetor dos estimadores dos parâmetros do modelo de regressão logística com todas as variáveis explicativas (modelo saturado) e $\widetilde{\boldsymbol{\beta}}$ é o vetor dos estimadores dos parâmetros do modelo de regressão logística, quando as q variáveis são retiradas (modelo ajustado).

Essas duas estatísticas são assintoticamente equivalentes e convergem em distribuição para uma qui-quadrado com os graus de liberdade dado pela diferença entre o número de parâmetros do modelo irrestrito e número de parâmetros do modelo restrito (sob a hipótese nula), isto é, Λ e $W \sim \chi^2_l$. Por exemplo, considere que sob o modelo de regressão logística, onde $\boldsymbol{\beta} = (\beta_0, \beta_1)$, temos interesse em testar as seguintes hipóteses $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$.

Note que o teste de hipóteses acima pode ser reescrito como $H_0 : \mathbf{C}\boldsymbol{\beta} = 0$ vs $H_1 : \mathbf{C}\boldsymbol{\beta} \neq 0$, onde $C = (0 \quad 1)$ e $q_0 = 0$. Dessa forma, sob esse teste de hipóteses, as estatísticas de Wald e Razão de Verossimilhança convergem em distribuição para uma qui-quadrado com 1 grau de liberdade.

O segundo conjunto de testes de hipóteses tem por objetivo verificar a qualidade do ajuste do modelo, ou seja, a adequação global do mesmo. Segundo Souza (2006), o processo de ajuste de um modelo consiste em propor ao mesmo um pequeno número de parâmetros, de tal forma que resuma toda a informação da amostra. Dado um conjunto de n observações, um modelo de até n parâmetros pode ser ajustado, sendo denominado modelo saturado, sendo que este indica toda a variação ao componente sistemático e reproduzindo exatamente os dados. Este modelo, apesar de ser não informativo, uma vez que não resume os dados, somente os reproduz, serve como base para medir a discrepância de um modelo intermediário de p parâmetros, por exemplo.

Existem muitas estatísticas para medir esta discrepância, das quais a mais utilizada é a estatística *Deviance*, baseada na função de verossimilhança e proposta por Nelder e Wedderburn (1972). Ela é definida como:

$$Deviance = -2ln\left[\frac{L(\hat{\beta}_0, ..., \hat{\beta}_p)}{L(y_1, ..., y_n)}\right], \tag{3.13}$$

onde $L(y_1, ..., y_n)$ corresponde à verossimilhança do modelo saturado e $L(\hat{\beta}_0, ..., \hat{\beta}_p)$ corresponde à verossimilhança do modelo ajustado. A *Deviance* pode ser escrita como:

$$\begin{aligned} D &= -2\sum_{i=1}^{n}[y_i log(\widehat{\pi}_i) + (1-y_i)log(1-\widehat{\pi}_i)] - \sum_{i=1}^{n}[y_i log(y_i) + (1-y_i)log(1-y_i)] \\ &= -2\sum_{i=1}^{n} y_i[log(\widehat{\pi}_i) - log(y_i)] + (1-y_i)[log(1-\widehat{\pi}_i) - log(1-y_i)] \\ &= -2\sum_{i=1}^{n}\left[y_i ln\left(\frac{\widehat{\pi}_i}{y_i}\right) + (1-y_i)ln\left(\frac{1-\widehat{\pi}_i}{1-y_i}\right)\right]. \end{aligned}$$

A *Deviance* sob a hipótese nula tem assintoticamente uma distribuição quiquadrado com $n-p$ graus de liberdade. No contexto de adequação do modelo, quanto menor for o valor da *Deviance* melhor é o ajuste.

Intervalos de confiança

Segundo Hosmer & Lemeshow (1989), o intervalo de $100(1-\alpha)\%$ de confiança para β_0 e $\beta_j, j = 1, ..., p$, é dado por:

$$\widehat{\beta}_j \pm z_{1-\alpha/2}\widehat{SE}(\widehat{\beta}_j), \quad j = 1, \ldots, p,$$

e para o intercepto,

$$\widehat{\beta}_0 \pm z_{1-\alpha/2}\widehat{SE}(\widehat{\beta}_0),$$

onde $\widehat{SE}(.)$ denota o desvio padrão estimado, obtido via a matriz informação de Fisher.

3.2.3 Interpretação dos coeficientes

Na regressão logística, o coeficiente de uma variável explicativa representa a inclinação ou a taxa de mudança da função logito para cada incremento de uma unidade no valor dessa variável com as demais variáveis explicativas fixas, podendo ser representada como

$$\beta_j = g(x_i + 1) - g(x_i),$$

onde a função $g(.)$ representa a função descrita em (3.4) e podendo β_j ser interpretada como o logaritmo da razão de chances e $\exp(\beta_j)$ como sendo a própria razão de chances ("*Odds* ratio").

A razão de chances é uma medida que representa o quão mais provável é se observar o evento de interesse para um indivíduo do que para outro, assumido como referência. Assim, pode-se dizer que para os indivíduos com $X_i = x_i+1$ o evento de interesse tem $exp(\beta_1)$ vezes a chance daqueles que assumem $X_i = x_i$. No contexto de *Credit Scoring*, suponha que o interesse seja classificar indivíduos em adimplentes e inadimplentes, sendo a inadimplência o evento de interesse. Suponha a variável dicotômica estado civil, em que $x = 0$ são os indivíduos casados e $x = 1$ indivíduos solteiros. Além disso, suponha que a razão de chances dessa variável seja igual a 4. Assumindo indivíduos casados como referência, isso quer dizer que a chance de um indivíduo solteiro ser inadimplente é quatro vezes a chance de um indivíduo casado. A razão de chances e o logaritmo da razão de chances são mostrados a seguir.

Em situações em que a variável explicativa é dicotômica, a chance de resposta, quando x=1, é definida como $\pi(1)/[1 - \pi(1)]$ da mesma forma, quando x=0, é definida como $\pi(0)/[1 - \pi(0)]$. O logaritmo da razão é dado por:

$$g(1) = ln\pi(1)/[1 - \pi(1)] \quad e \quad g(0) = ln\pi(0)/[1 - \pi(0)].$$

A razão de chances, denotada por ψ é definida por:

$$\psi = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]}.$$

O logaritmo da razão das chances ("*log-odds*") é:

$$ln(\psi) = ln\left[\frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]}\right] = g(1) - g(0).$$

Usando a expressão para o modelo de regressão logística definido em (3.2) e (3.3), a razão de chances é dada por:

$$\psi = \frac{\left(\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right) / \left(\frac{1}{1+\exp(\beta_0+\beta_1)}\right)}{\left(\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right) / \left(\frac{1}{1+\exp(\beta_0)}\right)} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1), \tag{3.14}$$

e o logaritmo da razão de chances é dado por:

$$ln(\psi) = ln[exp(\beta_1)] = \beta_1.$$

O intervalo de confiança, com nível de confiança $100(1-\alpha)\%$ para razão de chances é obtido inicialmente calculando o intervalo para β_1 e aplicando a exponencial, tem-se:

$$exp[\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_1)],$$

em que $\widehat{SE}(\widehat{\beta}_1)$ é o desvio padrão estimado de $\widehat{\beta}_1$.

3.3 Validação do modelo

Um procedimento usual para validação do modelo consiste em dividir o conjunto de observações em duas partes. Uma delas é usada para ajustar o modelo, cujas estimativas dos parâmetros são então usadas para fazer predições para os indivíduos do segundo conjunto (Carballo, 2002). Porém, neste trabalho, para validar o modelo de *Credit Scoring*, vamos utilizar a validação cruzada, porque vamos estar trabalhando com uma base de dados desbalanceada, onde o número de indivíduos adimplentes é maior que o número de inadimplentes.

A validação cruzada é uma técnica de reamostragem que permite que todos os dados da base de dados sejam utilizados para treinamento, que consiste no ajuste do modelo, e teste, que consiste na validação do modelo. Na validação cruzada os dados iniciais são particionados aleatoriamente em K subconjuntos $D_1, D_2, ..., D_K$ aproximadamente iguais em tamanho. Em seguida é ajustado um modelo utilizando $K-1$ subconjuntos e é testado o subconjunto restante. O procedimento anterior é repetido K vezes, utilizando sempre um subconjunto diferente para teste (Han e Kamber, 2006).

3.4 Medidas de desempenho

Algumas medidas são importantes para avaliar o desempenho do modelo. Tradicionalmente, a precisão é a mais comumente utilizada, porém para classificação de classes desbalanceadas, em que há uma classe minoritária, por exemplo, inadimplentes e uma classe predominante, adimplentes, a precisão não é mais uma medida interessante porque a classe minoritária tem muito pouco impacto sobre a precisão em comparação com a da classe predominante (Horta, 2010). Das medidas existentes para lidar com o problema de desequilíbrio de classes citadas por Sun (2008) foram escolhidas: matriz de confusão e área sob a curva ROC.

3.4.1 *Matriz de confusão*

Em um modelo com variável resposta binária, como ocorre no caso de um *Credit Scoring*, se busca classificar os clientes em uma das categorias consideradas, ou seja, em adimplentes ou inadimplentes e obter um bom grau de acerto nessas classificações.

Geralmente, nas amostras de validação, se conhece a resposta dos clientes em relação a sua real condição de crédito, e estabelecendo critérios em que se classifique esses clientes em adimplentes e inadimplentes, torna-se possível comparar essa classificação obtida com a verdadeira condição creditícia dos clientes. Essa comparação é feita através da *matriz de confusão* ou *matriz de erros*. Essa matriz descreve uma tabulação cruzada entre a classificação predita através de um único ponto de corte e a condição real e conhecida de cada individuo, onde os valores da diagonal principal compreendem as classificações corretas e os valores fora dessa diagonal correspondem aos erros de classificação. A tabela 3.1 ilustra uma *matriz de confusão* ou *matriz de erros*.

Tabela 3.1: *Matriz de confusão*

		Predito		
		Inadimplente	Adimplente	Total
Real	Inadimplente	a	b	a+b
	Adimplente	c	d	c+d
Total		a+c	b+d	a+b+c+d

Os dados de entrada dessa matriz têm os seguintes significados:

- a é a quantidade de clientes classificados em inadimplentes dado que eles são inadimplentes.
- b é a quantidade de clientes classificados como adimplentes dado que os clientes são inadimplentes.
- c é a quantidade de clientes classificados em inadimplentes dado que eles são adimplentes.
- d é a quantidade de clientes classificados em adimplentes dado que eles são adimplentes.

Algumas medidas comumente utilizadas em problemas de classificação binária, obtidas a partir de uma matriz de confusão são:

- Acurácia (AC): é a proporção do total de predições corretas.

$$AC = \frac{a+d}{a+b+c+d}$$

- Taxa de positivos-corretos (PC) ou sensibilidade: é a proporção de clientes classificados pelo modelo como adimplentes dado que eles são adimplentes.

$$PC = \frac{d}{c+d}$$

- Taxa de falsos positivos (FP): é a proporção de clientes classificados pelo modelo como adimplentes dado que eles são inadimplentes.

$$FP = \frac{b}{a+b}$$

- Taxa de negativos corretos (NC) ou especificidade: é a proporção de clientes classificados pelo modelo como inadimplentes dado que eles são inadimplentes.

$$NC = \frac{a}{a+b}$$

- Taxa de falsos-negativos (FN): é a proporção de clientes classificados pelo modelo como inadimplentes dado que eles são adimplentes.

$$FN = \frac{c}{c+d}$$

- Precisão (P): é a proporção de clientes que realmente são adimplentes dado que eles foram preditos pelo modelo como adimplentes (PV_+) e a proporção de clientes que realmente são inadimplentes dado que eles foram preditos pelo modelo como inadimplentes ($PV\_$), sendo, respectivamente:

$$PV_+ = \frac{d}{b+d} \quad e \quad PV\_ = \frac{a}{a+c}$$

3.4.2 Curva ROC (*Receiver Operating Characteristic*)

A Curva ROC constitui uma técnica útil para avaliar a capacidade preditiva dos modelos de risco de crédito e está baseada nos conceitos da sensibilidade e da especificidade.

Para a construção da Curva ROC, são calculadas a sensibilidade e a especificidade considerando diferentes pontos de corte do modelo, e a curva é obtida registrando em um gráfico "sensibilidade" *versus* "1-especificidade". Usualmente, pode ser usada para auxiliar na decisão do melhor ponto de corte que, em geral, é o ponto que maximiza tanto a sensibilidade quanto a especificidade, estando localizado no "ombro" da curva, ou próximo a ele.

Usualmente, na classificação binária, utiliza-se como limiar de referência 0.5. No contexto de inadimplência de empresas e concessão de crédito, se a probabilidade estimada para uma determinada empresa ou pessoa física é maior que 0.5, ela é classificada como adimplente, e o crédito é concedido. Caso contrário, ela é designada como inadimplente e o crédito é recusado.

A área sob a Curva ROC mede a capacidade de discriminação do modelo, quanto maior a área sob a curva, maior é a porcentagem de acerto nas classificações dos indivíduos adimplentes e inadimplentes. Uma maneira de interpretar a área sob a curva é através da estatística U de Mann-Whitney. Assumimos que n_1 denota o número de indivíduos inadimplentes ($Y = 1$) e n_0 denota o número de indivíduos adimplentes. Então são criados $n_1 \text{x} n_0$ pares: cada indivíduo inadimplente é pareado com um indivíduo adimplente. Destes $n_1 \text{x} n_0$ pares, determina-se a proporção de vezes em que o indivíduo inadimplente teve a maior das duas probabilidades. A proporção de pares classificados corretamente corresponde à estatística U de Mann-Whitney. Esta proporção mostra-se igual a área sob a curva ROC.

Segundo Hosmer & Lemeshow (1989,p. 162) a regra geral é:

- Se $ROC = 0.5$: sugere que não houve discriminação.
- Se $0.7 \leq ROC \leq 0.8$: considerado discriminação aceitável.
- Se $0.8 \leq ROC \leq 0.9$: considerado excelente discriminação.
- Se $ROC \geq 0.9$: considerado discriminação pendente.

Modelos de alto poder discriminatório concentram-se no canto superior esquerdo da curva ROC, pois à medida que a sensibilidade aumenta há pouca perda de especificidade. Já os modelos com menor poder discriminatório aproximam-se da diagonal. Esta diagonal revela a relação entre as taxas de resultados da sensibilidade (verdadeiros-positivos) com o complementar da especificidade (falsos-positivos) que seria obtida pelo modelo sem informação da inadimplência, por exemplo, se o analista de crédito jogasse uma moeda (Carballo, 2002).

4 Aplicação

4.1 Descrição dos Dados

Para ilustrar o desenvolvimento de um modelo de *Credit Scoring*, será utilizado um conjunto de dados reais de um banco do sul da Alemanha, que consiste de 1000 observações de 1000 requerentes de crédito. São consideradas 21 variáveis explicativas e a variável resposta dicotômica confiabilidade de crédito, sendo o nosso evento de interesse ($Y = 1$) a adimplência. Os dados foram estratificados em 300 requerentes não dignos de crédito (inadimplentes) e 700 requerentes dignos de crédito (adimplentes). Os dados brutos podem ser obtidos por meio eletrônico no endereço $http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html$.

Variáveis

1. **Crédito:** é a resposta para cada indivíduo, sendo 0= "não digno de crédito"e 1= "digno de crédito"

2. **Balanço:** saldo atual em conta corrente

3. **Estado civil:** estado civil do cliente

4. **Duração:** duração do empréstimo, em meses

5. **Estrangeiro:** se o cliente é estrangeiro

6. **Garantia:** se o cliente apresenta algum fiador ou co-pretendente

7. **Idade:** idade do cliente, em anos

8. **Moradia:** tipo de moradia do cliente

9. **Histórico:** situação do cliente com relação ao pagamento de créditos anteriores

10. **Ocupação:** ocupação do cliente

11. **Créditos adicionais:** se o cliente está ou tem solicitado crédito em outra(s) instituição(ões)

12. **Pessoas:** número de pessoas dependentes

13. **Propósito:** propósito do empréstimo

14. **Créditos anteriores:** número de empréstimos anteriores, incluindo o atual

15. **Proposta de crédito:** quantidade de crédito tomado, em *Deutsche Marks* (moeda oficial Alemã até 1999)

16. **Razão:** parcelas, em % da renda disponível

17. **Recurso:** bens disponíveis de maior valor dados como garantia para o empréstimo

18. **Sexo:** sexo do cliente

19. **Telefone:** se o cliente possui telefone

20. **Tempo emprego:** tempo no emprego atual

21. **Tempo residência:** tempo na residência atual

22. **Valor poupança:** valor depositado em poupança ou de ações, em *Deutsche Marks*

4.2 Construção do modelo

Na construção do modelo de *Credit Scoring* foi adotado o método de validação cruzada, discutido na seção 3.3, de modo que todos os indivíduos da base foram utilizados para treinamento (ajuste) e teste (validação). Dessa maneira, o banco de dados de 1000 requerentes de crédito foi dividido em 4 subgrupos de 250 individuos selecionados aleatoriamente. As amostras de treinamento foram constituídas de 750 indivíduos (75%), ou seja, composta por 3 subgrupos, e as amostras para teste foram constituídas de 250 indivíduos (25%), composta pelo subgrupo restante. A escolha destas proporções nas amostras de treinamento e teste são as mais usuais encontradas na literatura, o que justifica elas serem usadas neste trabalho. O procedimento de treinamento do modelo e teste foi repetido 4 vezes, utilizando sempre um subconjunto diferente para teste. Dado que os subgrupos foram selecionados aleatoriamente, a proporção de indivíduos dignos e não dignos de crédito em cada subgrupo não é fixada. Há alguns autores que trabalharam com estas proporções fixadas, para mais detalhes veja, por exemplo, Vasconcellos (2002) e Abreu (2004). Para desenvolvimento do modelo foi utilizado o software estatístico R 2.10.1.

Para a seleção das variáveis incluídas nos modelos, utilizou-se o procedimento estatístico denominado *stepwise*, o qual permite chegar, através de entradas e saídas das variáveis explicativas no modelo, a um conjunto de variáveis para a discriminação da ocorrência de clientes dignos e não dignos de crédito. Este procedimento permite duas maneiras de seleção de variáveis: pelo nível de significância das mesmas e pelo critério AIC (critérios de informação de Akaike, 1994). Quando pelo nível de significância, a entrada e saida das variáveis é feita utilizando o teste da razão de verossimilhança, discutido na seção 3.2.2. Para minimizar a entrada de variáveis sem efeito prático, os níveis de significância adotados por alguns autores são inferiores ao tradicionalmente utilizado 0.05 (5%). Para mais detalhes, sobre tal procedimento, veja Pereira (2004) e Abreu (2004), por exemplo. Quando pelo critério AIC ($-2l(\beta) + 2p$, onde $l(\beta)$ está definido em (3.7) e p é o número de parâmetros do modelo), um valor baixo para AIC é considerado como representativo de um melhor ajuste e os modelos são selecionados visando a obter um mínimo AIC. No presente trabalho, estaremos utilizando o critério AIC, em função de ser este o adotado pelo software R.

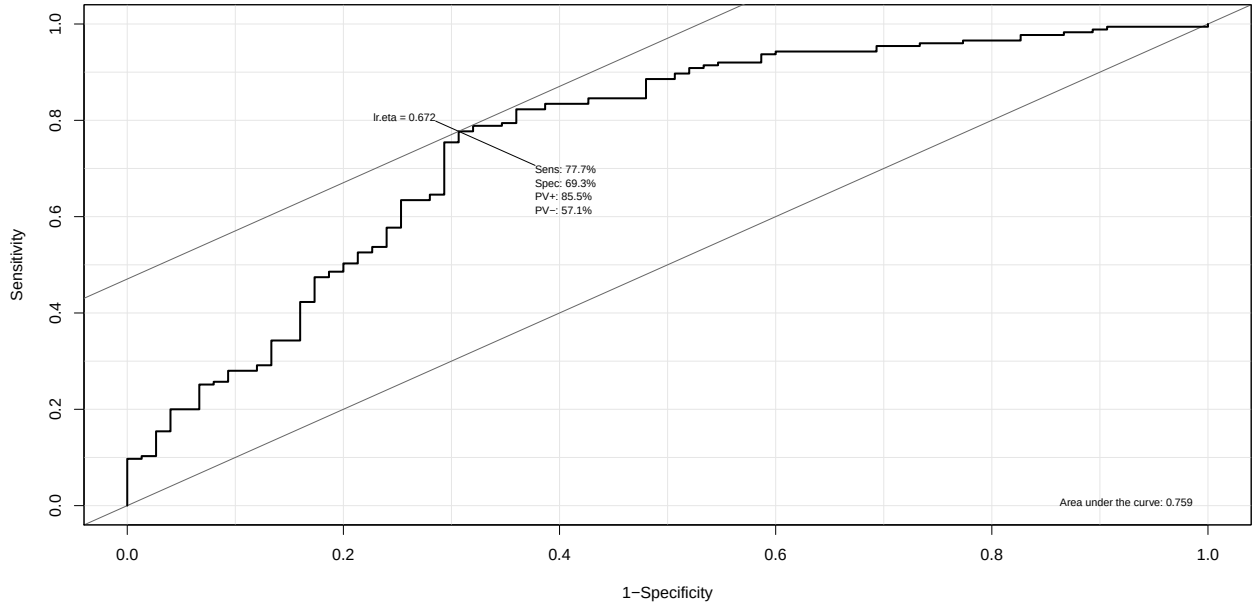
Estaremos interessados, dentre os quatro modelos gerados na validação cruzada, verificar aquele com melhor capacidade de classificação, ou seja, que obteve maior porcentagem de acerto na classificação dos clientes dignos de crédito quando eles realmente são dignos (sensibilidade), e dos não dignos de crédito quando eles realmente não são dignos (especificidade). Além dessas duas medidas, outra medida fundamental é a taxa de falsos positivos, que é a proporção de clientes não dignos classificados como dignos pelo modelo. Essa taxa implica em custo para a instituição financeira, portanto, quanto menor for essa taxa, melhor é o modelo.

Para cada amostra de teste foram feitas a Curva ROC e a *Matriz de confusão* e através dessas medidas foi estabelecido qual o modelo com maior capacidade preditiva. É importante ressaltar que a comparação entre as matrizes de confusão dos modelos só é possível porque a proporção de indivíduos adimplentes e inadimplentes nas amostras de teste são similares. A seguir, um resumo dos modelos gerados pela validação cruzada e a capacidade preditiva de cada um.

4.2.1 Modelo 1

A amostra de treinamento do modelo 1 é constituida por 525 indivíduos dignos de crédito e 225 não dignos de crédito e a amostra de teste formada por 175 dignos de crédito e 75 não dignos. Para verificar a capacidade preditiva do modelo, seguem a curva ROC e a *Matriz de confusão*.

Figura 4.1: Curva ROC - Modelo 1



Sensibilidade = 77,7%. Especificidade = 69,3%. $PV_+ = 85,5\%$ e $PV_- = 57,1\%$. Ir.eta (0,672) corresponde ao ponto de corte do modelo.

De acordo com a Curva ROC, a capacidade de discriminação do modelo foi de 0,759, o que, segundo Hosmer & Lemeshow (1989) é uma discriminação aceitável.

Tabela 4.1: *Matriz de confusão* - Modelo 1

		Predito	
		Não digno	Digno
Real	Não digno	52	23
	Digno	39	136

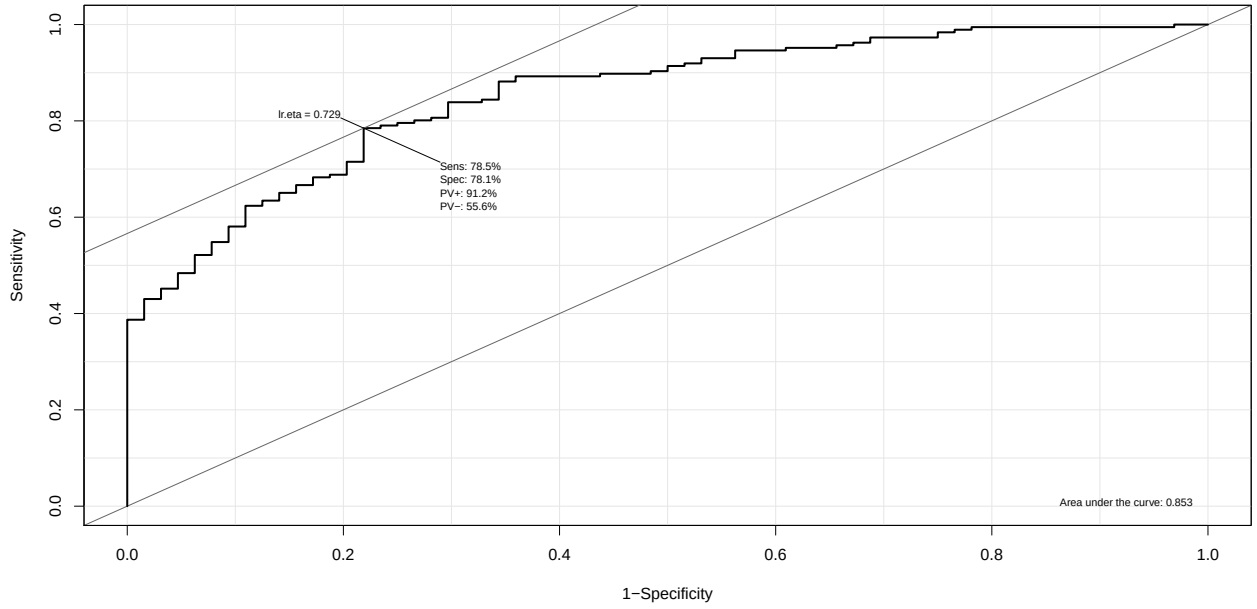
Conforme exibe a *matriz de confusão*, dos 175 indivíduos dignos de crédito, 136 foram classificados corretamente, e dos 75 não dignos, 52 foram classificados corretamente,

o que corresponde, respectivamente, à sensibilidade e especificidade do modelo exibidas na Curva Roc (Figura 4.1). Para este modelo, a taxa de falso-positivo foi de 30,6% e a taxa de falso-negativo foi de 22,2%.

4.2.2 Modelo 2

A amostra de treinamento do modelo 2 é formada por 514 indivíduos dignos de crédito e 236 não dignos de crédito e a amostra de teste formada por 186 dignos de crédito e 64 não dignos.

Figura 4.2: Curva ROC - Modelo 2



Sensibilidade= 78,5%. Especificidade= 78,1%. $PV_+ = 91,2\%$ e $PV_- = 55,6\%$. Ir.eta (0,729) corresponde ao ponto de corte do modelo.

De acordo com a Curva ROC, a capacidade de discriminação do modelo foi de 0,853, o que, segundo Hosmer & Lemeshow (1989) é uma excelente discriminação.

Conforme exibe a *matriz de confusão* do segundo modelo, dos 186 indivíduos dignos de crédito, 146 foram classificados corretamente, e dos 64 não dignos, 50 foram classificados corretamente, o que corresponde, respectivamente, à sensibilidade e especificidade do modelo exibido na Curva Roc (Figura 4.2). Para este modelo, a taxa de falso-positivo foi de 21,8% e a taxa de falso-negativo foi de 21,5%.

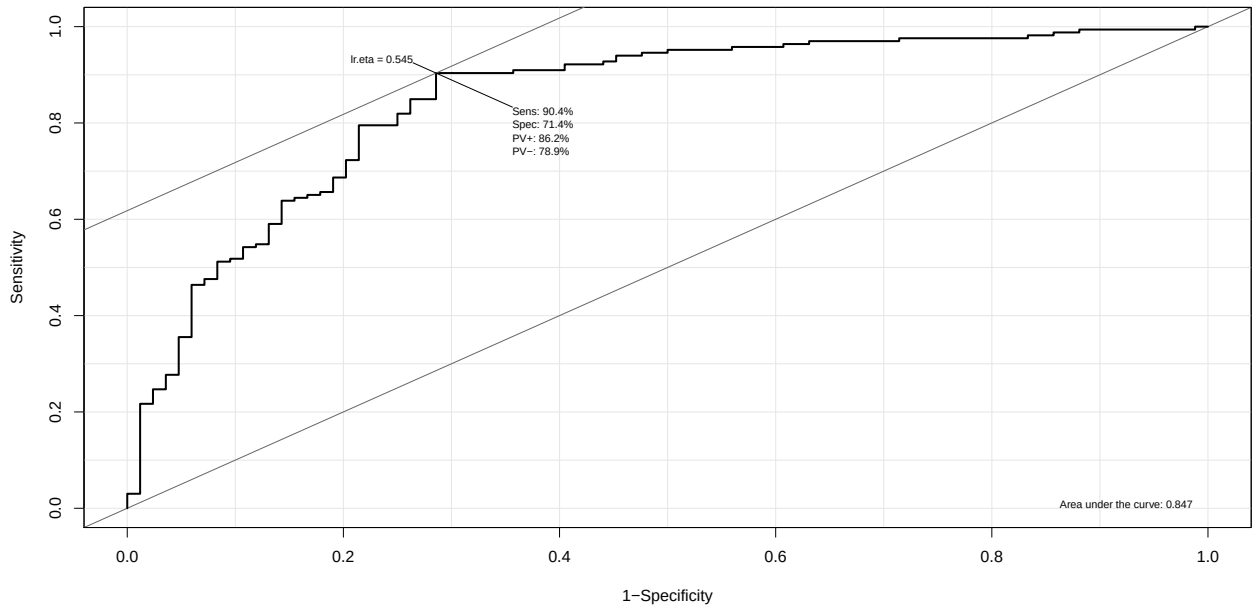
Tabela 4.2: *Matriz de confusão* - Modelo 2

		Predito	
		Não digno	Digno
Real	Não digno	50	14
	Digno	40	146

4.2.3 Modelo 3

A amostra de treinamento do modelo 3 é formada por 534 indivíduos dignos de crédito e 216 não dignos de crédito e a amostra de teste formada por 166 dignos de crédito e 84 não dignos.

Figura 4.3: Curva ROC - Modelo 3



Sensibilidade= 90,4%. Especificidade= 71,4%. $PV_+ = 86,2\%$ e $PV_- = 78,9\%$. Ir.eta (0,545) corresponde ao ponto de corte do modelo.

De acordo com a Curva ROC, a capacidade de discriminação do modelo foi de 0,847, o que, segundo Hosmer & Lemeshow (1989) é uma excelente discriminação.

Tabela 4.3: *Matriz de confusão* - Modelo 3

		Predito	
		Não digno	Digno
Real	Não digno	60	24
	Digno	16	150

Conforme exibe a *matriz de confusão* do terceiro modelo, dos 166 indivíduos dignos de crédito, 150 foram classificados corretamente, e dos 84 não dignos, 60 foram classificados corretamente, o que corresponde, respectivamente, à sensibilidade e especificidade do modelo exibido na Curva Roc (Figura 4.3). Para este modelo, a taxa de falso-positivo foi de 28,6% e a taxa de falso-negativo foi de 9,63%.

4.2.4 Modelo 4

A amostra de treinamento do modelo 4 é formada por 527 indivíduos dignos de crédito e 223 não dignos de crédito e a amostra de teste formada por 173 dignos de crédito e 77 não dignos.

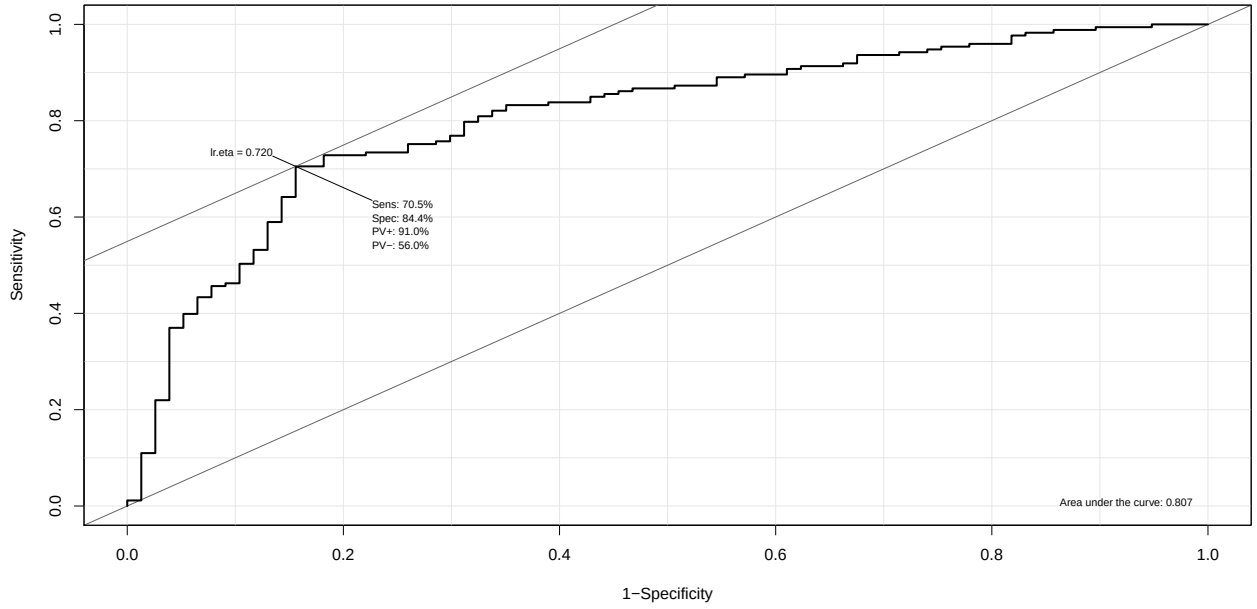
De acordo com a Curva ROC (Figura 4.4), a capacidade de discriminação do modelo foi de 0,807, o que, segundo Hosmer & Lemeshow (1989) é uma excelente discriminação.

Tabela 4.4: *Matriz de confusão* - Modelo 4

		Predito	
		Não digno	Digno
Real	Não digno	65	12
	Digno	51	122

Conforme exibe a *matriz de confusão* do quarto modelo, dos 173 indivíduos dignos de crédito, 122 foram classificados corretamente, e dos 77 não dignos, 65 foram classificados corretamente, o que corresponde, respectivamente, à sensibilidade e especificidade do modelo exibido na Curva Roc. Para este modelo, a taxa de falso-positivo foi de 15,6% e a taxa de falso-negativo foi de 29,48%.

Figura 4.4: Curva ROC - Modelo 4



Sensibilidade= 70,5%. Especificidade= 84,4%. $PV_+ = 91,0\%$ e $PV_- = 56\%$. Ir.eta (0,720) corresponde ao ponto de corte do modelo.

4.3 Resultados

Tabela 4.5: Resumo dos modelos

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Sensitividade	77,70%	78,50%	90,40%	70,50%
Especificidade	69,30%	78,10%	71,40%	84,40%
Área sob a curva	0,759	0,853	0,847	0,807
Falsos-positivos	30,60%	21,80%	28,60%	15,60%

Queremos deixar claro que, apesar de no nosso exemplo o evento de interesse ser a adimplência, na análise de desempenho dos quatro modelos gerados pela validação cruzada, deve-se levar em consideração que para uma instituição financeira concessora de crédito, é mais interessante que o modelo seja eficaz na predição da inadimplência do que da adimplência, já que o erro de aprovar uma operação que se tornará problemática é considerado mais grave que a recusa de uma operação que seria um bom negócio para

a instituição. Diante disso, como já dito anteriormente, nosso interesse maior, além de avaliar a capacidade preditiva do modelo, sensibilidade e especificidade, está em verificar a menor taxa de falsos-positivos.

Verifica-se, que, de uma forma geral, os percentuais de acerto dos modelos em suas classificações ficaram em torno de 80%. O modelo 1 é o que obteve pior desempenho, pois além de possuir a menor capacidade preditiva (0,759), dada pela área abaixo da Curva ROC, obteve maior taxa de falsos-positivos (30,6%). Os modelos 2, 3 e 4, se assemelharam na capacidade preditiva, todos eles maiores que 0,80, o que, de acordo com Hosmer & Lemeshow (1989) é uma excelente discriminação. Apesar da semelhança, o modelo 4 se destaca dos demais, pois além de sua taxa de falsos-positivos ser menor, sua especificidade foi maior, comparada aos outros modelos. Assim, o modelo 4, se apresentou mais eficaz na predição da inadimplência, evento de interesse no contexto de *Credit Scoring*.

4.3.1 Descrição e interpretação do modelo selecionado

Visto ser o modelo 4 aquele que se mostrou mais eficaz, as tabelas a seguir exibem, respectivamente, as 15 variáveis que foram incluídas no modelo final através do método de seleção *stepwise* e os coeficientes estimados para as 15 variáveis e o intercepto, desvio padrão e significância estatística.

Tabela 4.6: Variáveis do modelo

Variável	**Valor**	**Categoria**
Razão	**1**	$\geq$ **35**
	2	$25 \leq ... < 35$
	3	$20 \leq ... < 25$
	4	< 20
Garantia	**1**	**Nenhum**
	2	Co-pretendente
	3	Fiador
Creditos adicionais	**1**	**em outros bancos**
	2	em lojas
	3	nenhum
Propósito	**1**	**Carro novo**
	2	Carro usado
Duração	Contínua (em meses)	**Não categorizada**
Proposta	Contínua (em *Deutche Marks*)	**Não categorizada**
Idade	Contínua (em meses)	**Não categorizada**

*categorias em negrito são os níveis de referência utilizados

Como pode ser observado, além das variáveis categóricas listadas, três variáveis contínuas foram incluídas no modelo, totalizando 15 variáveis explicativas. Verifica-se que a um nível de significância de 0,10, não são todas as categorias das variáveis significativas, porém, consideramos que se pelo menos uma categoria é significativa, as outras também permanecem no modelo.

Tabela 4.7: Variáveis do modelo (cont.)

Variável	**Valor**	**Categoria**
Estado civil	**1**	**Solteiro(a)**
	2	Casado(a)
Estrangeiro	**1**	**Sim**
	2	Não
Tempo residencia	**1**	$<$ 1 **ano**
	2	$1 \leq ... < 4$ anos
	3	$4 \leq ... < 7$ anos
	4	≥ 7 anos
Histórico	**1**	**Ruim**
	2	Bom
Tipo de moradia	**1**	**Gratuito**
	2	Aluguel
	3	Próprio
Tempo emprego	**1**	**Desempregado**
	2	≤ 1 ano
	3	$1 \leq ... < 4$ anos
	4	$4 \leq ... < 7$ anos
	5	≥ 7 anos
Valor poupança	**1**	**Não disponível / Sem poupança**
	2	$<$ 100 DM
	3	$100 \leq ... <$ 500 DM
	4	$500 \leq ... <$ 1000 DM
	5	$\geq$ 1000 DM
Balanço	**1**	**Inativo**
	2	Sem balanço ou em débito
	3	$0a200$ DM
	4	$>$ 200 DM

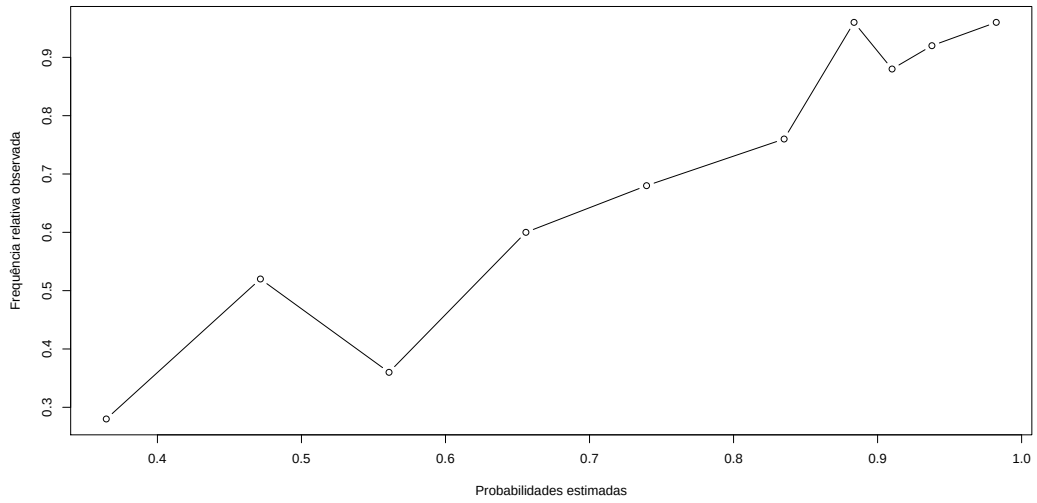
*categorias em negrito são os níveis de referência utilizados

Tabela 4.8: Coeficientes dos parâmetros estimados e significância estatística

Variável	**Categoria**	**Estimativa**	**Desvio padrão**	**p-valor**
Intercepto	-	-0,6620	0,7739	0,392376
Estado civil	2	0,5269	0,2115	0,012733
Estrangeiro	2	1,664	0,7585	0,028272
Tempo de residência	2	-1,073	0,3559	0,002578
	3	-0,6968	0,3839	0,069528
	4	-0,5867	0,3483	0,092068
Histórico	2	0,7854	0,3385	0,020313
Tipo de moradia	2	0,5717	0,2655	0,031302
	3	0,4106	0,3877	0,289455
Tempo emprego	2	-0,5424	0,4257	0,202594
	3	0,01138	0,3909	0,976776
	4	0,7622	0,4454	0,087018
	5	0,02445	0,4073	0,952136
Valor poupança	2	0,1504	0,3255	0,644011
	3	0,3928	0,4426	0,374802
	4	2,110	0,7792	0,006761
	5	0,9700	0,2998	0,001215
Balanço	2	0,5427	0,2470	0,028042
	3	0,8885	0,4225	0,035452
	4	1,823	0,2642	5,17e-12
Razão	2	-0,4638	0,3615	0,199471
	3	-0,7297	0,3933	0,063511
	4	-1,184	0,3525	0,000779
Garantia	2	-0,2212	0,4664	0,635343
	3	0,8898	0,4776	0,062461
Créditos adicionais	2	-0,04402	0,5207	0,932635
	3	0,4939	0,2725	0,069876
Propósito	2	0,7705	0,2045	0,000164
Duração	contínua	-0,03084	0,01028	0,002707
Proposta	contínua	-8,486e-05	4,53e-05	0,061350
Idade	contínua	0,01445	0,01027	0,159564

Uma maneira interessante de verificar o desempenho de uma predição probabilística para um evento binário consiste em construir o gráfico de confiabilidade do modelo. Abaixo, o grafico de confiabilidade do modelo 4.

Figura 4.5: Gráfico de confiabilidade - Modelo 4



Neste gráfico temos no eixo X as probabilidades que foram estimadas pelo modelo, sendo que estas foram divididas em decis, afim de que não seja necessário plotar 250 probabilidades. No eixo Y, temos as frequências relativas observadas, ou seja, dentre as 25 observações do primeiro decil, qual a proporção de clientes que na minha base de dados são adimplentes. O objetivo deste gráfico é verificar se o modelo 4 está sendo realmente coerente na predição dos clientes em adimplentes. Verificamos ser o modelo 4 coerente na predição, pois, observa-se que quanto maior a probabilidade estimada maior é a frequência de clientes adimplentes.

A *Deviance* obtida para o modelo ajustado foi 678,35 com 719 graus de liberdade, o que corresponde a um p-valor menor que 0,0001, logo a hipótese nula de que os coeficientes do modelo sejam iguais a zero é rejeitada, implicando na significância de pelo menos um dos coeficientes exibidos na tabela 4.7 e portanto, concluímos ser o modelo ajustado significativo.

Descrito o modelo e verificado sua adequação global dado pela *Deviance*, vamos agora interpretá-lo. Considerando todas as demais variáveis do modelo fixas, se o cliente tem menos de um ano de emprego, seu score final é diminuído de -0,5424 pontos, ou seja, a chance dele ser adimplente diminui em 0,5424. Se o cliente é estrangeiro, não

terá nenhum acréscimo no seu score final, porém, se o cliente não é estrangeiro terá um acréscimo de 1,664 pontos no seu score final. Note que o score final do cliente terá um acréscimo de 0,01445 pontos a cada ano de envelhecimento do cliente, logo, a chance de adimplência de pessoas mais velhas é maior do que para pessoas jovens. As análises são análogas para as outras variáveis. O escore de cada indivíduo representa a probabilidade dele ser adimplente, logo, tomando como exemplo os indivíduos 1 e 2 com escores 0,7951 e 0,4096, respectivamente, podemos dizer que a probabilidade do indivíduo 1 ser adimplente é maior que a probabilidade de o indivíduo 2 ser adimplente.

Observando os valores da odds ratio na tabela 4.8, podemos dizer para a variável estado civil, que a chance de um indivíduo casado ser digno de crédito é 1,6937 vezes a chance de um indivíduo solteiro. Da mesma forma, a chance de um indivíduo com um bom histórico ser digno de crédito é 2,1932 vezes a chance do indivíduo que possui um histórico ruim. Para a variável tipo de moradia, a chance de um indíviduo que paga aluguel e de um indivíduo que tem casa própria ser adimplente é, respectivamente, 1,7712 e 1,5077 vezes a chance do indivíduo que mora de forma gratuita. Em relação a tempo de emprego, pode-se dizer que a chance de um indivíduo que está a mais de sete anos no emprego atual ser adimplente é 1,0247 vezes a chance de um indivíduo que está desempregado.

Tabela 4.9: Interpretação dos coeficientes

Variável	**Categoria**	**Odds ratio**
Estado civil	2	1,6937
Estrangeiro	2	5,2787
Tempo residência	2	0,3421
	3	0,4941
	4	0,5561
Histórico	2	
Tipo de moradia	2	1,7712
	3	1,5077
Tempo emprego	2	0,5813
	3	1,0114
	4	2,1430
	5	1,0247
Valor poupança	2	1,1623
	3	1,4811
	4	8,2519
	5	2,6378
Balanço	2	1,7206
	3	2,4314
	4	6,1905
Razão	2	0,6288
	3	0,4820
	4	0,3059
Garantia	2	0,8015
	3	2,4347
Créditos adicionais	2	0,9569
	3	1,6387
Propósito	2	2,1608
Duração	contínua	0,9696
Proposta	contínua	0,9999
Idade	contínua	1,0145

5 Conclusões

Modelos de *Credit Scoring* são ferramentas bastante válidas para avaliar a concessão de crédito de uma forma objetiva, tendo em vista que seu desempenho é superior aos métodos de julgamento humano, onde os analistas avaliam se concedem ou não o crédito baseados em critérios subjetivos. O método de análise de concessão de crédito, proposto neste trabalho, mostrou que a implementação do modelo de *Credit Scoring* utilizando o modelo de regressão logística é capaz de classificar corretamente a grande maioria das operações de crédito de uma instituição financeira.

Determinar a inadimplência de uma empresa ou indivíduo apresenta dificuldades inerentes às características do banco de dados. Os dados, geralmente são desbalanceados, onde o número de exemplos da classe de indivíduos ou empresas não dignos (inadimplentes) é bem inferior ao número de exemplos da classe de indivíduos ou empresas classificados como dignos (adimplentes) e, dado que o interesse maior está em obter uma classificação correta para clientes inadimplentes, diante de dados desbalanceados esta classificação é prejudicada.

Diante disso, é preciso obter e organizar um grande número de informações, com um número suficiente de operações de crédito e variáveis relativas às operações, sendo necessário o entendimento da importância de cada variável no processo de discriminação de clientes inadimplentes de adimplentes. Além disso, para gerar um modelo de *Credit Scoring* é necessário que a instituição tenha um sistema computacional de grande capacidade de armazenamento e processamento de dados, pois podem existir diversas linha de crédito para as quais devem ser elaborados modelos individuais que considerem as características de cada linha específica, por exemplo, quando se trata de empresas, estas podem pertencer a setores econômicos distintos, o que pode traduzir em comportamento diferenciado das variáveis para cada setor representado.

Em função da dificuldade em encontrar base de dados para aplicação, pois se trata de informações confidenciais, este trabalho utilizou uma base de dados com um pequeno número de operações de crédito (1000), sendo 300 indivíduos dignos de crédito e 700 indivíduos não dignos, portanto, um banco desbalanceado. A técnica de validação

cruzada para obtenção dos modelos ajustados foi, de certa maneira, uma forma de diminuir o impacto do desbalanceamento da base na capacidade preditiva do modelo, pois permite que todos os indivíduos sejam utilizados tanto para o ajuste quanto para validação.

Referências Bibliográficas

[1] ABREU, H. J. **Aplicação da Análise de Sobrevivência em um problema de Credit Scoring e comparação com a Regressão Logística**. Dissertação (Mestrado em Estatística)-Centro de Ciências Exatas e de Tecnologia, Universidade Federal de São Carlos, São Carlos, 2004.

[2] ARMINGER, G., ENACHE, D. and BONNE, T. (1997). *Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis, and feedforwards neural networks*. Computational Statistics, 12, 293-310.

[3] ALISSON, P. D.;**Logistic regression using the SAS System, theory and application.** SAS Institute, 1999. p. 304.

[4] BRITO, G.A.S; NETO, A.A; CORRAR, L.J. (2009). *Sistema de classificação de risco de crédito: uma aplicação a companhias abertas no Brasil*. Revista Contabilidade & Finanças, USP, São Paulo, vol. 20, p. 28-43.

[5] CAOUETTE, J. B., ALTMAN, E. I. & NARAYANAN, P. *Managing Credit Risk - The next Great Financial Challenge*, New York: John Wiley & Son Inc., 1998.

[6] CARBALLO, M. T. **Predição da Macrossomia Fetal através da Regressão Logística e de Redes Neurais Artificiais**. Dissertação (Bacharel em Estatística)-Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

[7] CARMONA, C. U.; AMORIM NETO, A. Modelagem do Risco de Crédito: Um Estudo do Segmento de Pessoas Físicas em um Banco de Varejo. **Revista Eletrônica de Administração da UFRGS - REAd**. 40 Edição, Porto Alegre, Vol. 10, Jul/ago, 2002. Disponível em ¡http://www.read.adm.ufrgs.br¿.

[8] CHAIA, A. J. **Modelos de gestão do risco de crédito e sua aplicabilidade ao mercado brasileiro**. Dissertação (Mestrado em administração)-Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2003.

[9] CORDEIRO, G. M.; DEMÉTRIO, C. G. B. Modelos Lineares Generalizados e Extensões. São Paulo, 2008.

[10] DANTAS, R. F.; DESOUZA, S.A. (2008). *Modelo de risco e decisão de crédito baseado em estrutura de capital com informação assimétrica*. Pesquisa Operacional, Ceará, n. 2, p. 263-284.

[11] DEMÉTRIO, C. G. B.; *Modelos lineares generalizados em experimentação agronômica*. Piracicaba: CALQ, Departamento Editorial, 2002, 113p.

[12] FAYYAD, U.M., PIATETSKY-SHAPIRO, G. & SMITH, P. "From Data Mining to Knowledge Discovery in Databases", Al Magazine, Vol. 17, No. 3, pp. 37-54, 1996.

[13] GRABLOWSKY, B. J. and TALLEY, W.K. (1981). *Probit and discrimant function for classifying credit applicants: a comparison*. Journal of Economics and Business, 33, 254-261.

[14] HAN J.; KAMBER, M. *Data mining: concepts and techniques*. Morgan Kaufmann, San Francisco, 2 edition, 2006.

[15] HAND, D. J. and HENLEY, D.J. (1997). *Statistical Classification Methods in Consumer Credit Scoring: a Review*. Journal og the Royal Statistical Society, Series A, 160, part 3, 523-541.

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

[17] HORTA, R. A. M. **Uma metodologia de Mineração de Dados para a previsão de insolvência de empresas brasileiras de capital aberto**. Dissertação (Doutorado em Engenharia Civil)-Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010.

[18] HOSMER, D. W. & LEMESHOW, S. Applied logistic regression. New York: John Wiley & Sons, Inc., 1989.

[19] MOTTA, C. G. L. **Sistema Inteligente para Avaliação de Riscos em Vias de Transporte Terrestre**. Dissertação (Mestrado em Ciências em Engenharia Civil)-COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2004.

[20] NELDER, J. A.; WEDDERBURN, R. W. M; Generalized linear models. **Journal of the Royal Statistical Society**, London, v.135, p. 370-384, 1972.

[21] ORGLER, Y. E. (1970). *A credit scring for comercials loans*. Journal of money, Credit and Banking, november, 31-37.

[22] PARKINSON, K. L.; OCHS, J. R. Using credit screening to manage credit risk. **Business Credit**, p.23-27, março, 1998.

[23] PERERA, Luiz Carlos Jacob. *Decisão de Crédito para Grandes Corporações*, Tese (Doutorado em Administração)-FEA/USP, São Paulo: Universidade São Paulo, 1998.

[24] PEREIRA, G. H. A. **Modelos de Risco de Crédito de Clientes: Uma aplicação a dados reais**. Dissertação (Mestrado em Estatística)-Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2004.

[25] REZENDE, S. O., PUGLIESI, J. B., MELANDA E. A. & DE PAULA, M. F."Mineração de Dados", In: *Sistemas Inteligentes: Fundamentos e Aplicações*, Barueri, SP, Brasil, Rezende, S. O. (coord.), Editora Manole Ltda., Cap. 12, pp. 307-336, 2003.

[26] ROSENBERG, E. & GLEIT, A. (1994). Quantitative methods in credit management: a survey. *Operations Research*, 42(4), 589-613.

[27] SANTOS, J. O. *Análise de crédito*empresas e pessoas físicas. 2. ed. São Paulo: Atlas, 2003.

[28] SANTOS, J. O.; FAMÁ, R. (2006). *Avaliação da aplicabilidade de um modelo de credit scoring com variáveis sistêmicas e não-sistêmicas em carteiras de crédito bancário rotativo de pessoas físicas*. Revista Contabilidade & Finanças, USP, São Paulo, n. 44, p. 105-117.

[29] SANTOS, J. O. (2008). *Análise comparativa de métodos para previsão de insolvência em uma carteira de crédito bancário de empresas de médio porte*. Revista de gestão USP, São Paulo, vol. 15, n. 3, p. 11-24.

[30] SAUNDERS, A. *Medindo o risco de crédito-Novas Abordagens para Value at Risk e outros Paradigmas*. Rio de Janeiro: Qualitymark, 2000.

[31] SENGER, L. J.; CALDAS JÚNIOR, J. **Análise de risco de crédito utilizando redes neurais artificiais**; Revista do CCEI/ Universidade da Região da Campanha. v.5 n.8 URCAMP, 2001.

[32] SICSÚ, A.L.**Desenvolvimento de um sistema de Credit Scoring**. São Paulo. Tecnologia de crédito, jan/mar 1998.

[33] SILVA, J. P. *Gestão e Análise de Risco de Crédito*. São Paulo: Atlas, 2000.

[34] SOUZA, E. C. **Análise de influência local no modelo de regressão logística**. Dissertação (Mestrado em Agronomia)-Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2006.

[35] STANDARD & POOR'S. *Corporate rating criteria*. New York: McGraw-Hill Companies Inc., 2008.

[36] SUN JIE, LI HUI. "Data mining method for listed companies financial distress prediction". *Knowledge-Based Systems*, 2008, v. 21, p. 1-5.

[37] THOMAS, L.C.; EDELMAN, D. B.; CROOK, J. N. (2002)*Credit Scoring and Its Applications*, Philadelphia: SIAM.

[38] VASCONCELLOS, M. S. **Proposta de método para análise de concessões de crédito a pessoas físicas**. Dissertação (Mestrado em Economia)-Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2002.

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