UNIVERSIDADE FEDERAL DO CEARÁ
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE
DEPARTAMENTO DE ADMINISTRAÇÃO
CURSO DE CIÊNCIAS ATUARIAIS

EMANUELLE SEVERINO PINHEIRO

APLICAÇÃO DE TÉCNICAS DE *DATA MINING* NO SUPORTE À TOMADA DE DECISÕES DE UMA EMPRESA DE CARTÃO DE CRÉDITO

FORTALEZA
2014

EMANUELLE SEVERINO PINHEIRO

APLICAÇÃO DE TÉCNICAS DE *DATA MINING* NO SUPORTE À TOMADA DE DECISÕES DE UMA EMPRESA DE CARTÃO DE CRÉDITO

Monografia ao Curso de Ciências Atuariais do Departamento de Administração da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Ciências Atuariais.

Orientador: Prof.ª. Drª. Silvia Maria Dias Pedro Rebouças

FORTALEZA

2014

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca da Faculdade de Economia, Administração, Atuária e Contabilidade

P718a Pinheiro, Emanuelle Severino.
Aplicação de técnicas de data mining no suporte à tomada de decisões de uma empresa de cartão de crédito / Emanuelle Severino Pinheiro. – 2014.
72 f. : il. color., enc. ; 30 cm.

Monografia (graduação) – Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Curso de Ciências Atuariais, Fortaleza, 2014.
Orientação: Profa. Dra. Silvia Maria Dias Pedro Rebouças.

1. Exploração de dados. 2. Processo decisório. I. Título.

CDD 368.01

EMANUELLE SEVERINO PINHEIRO

APLICAÇÃO DE TÉCNICAS DE *DATA MINING* NO SUPORTE À TOMADA DE DECISÕES DE UMA EMPRESA DE CARTÃO DE CRÉDITO

Monografia ao Curso de Ciências Atuariais do Departamento de Administração da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Ciências Atuariais.

Aprovada em: ___/___/_____.

BANCA EXAMINADORA

_____

Prof.ª Dr.ª Silvia Maria Dias Pedro Rebouças (Orientador)

Universidade Federal do Ceará (UFC)

_____

Prof.ª Luciana Moura Reinaldo

Universidade Federal do Ceará (UFC)

_____

Prof. Sérgio César de Paula Cardoso

Universidade Federal do Ceará (UFC)

À Deus.

Aos meus pais, e minha irmã.

À minha querida prima (*in memorian*).

AGRADECIMENTO

Primeiramente aos meus pais, por toda a sabedoria e suporte que foram dados a mim, fundamentais para a busca de todos os meus propósitos.

Ao meu amigo, parceiro e namorado, por sua ajuda constante sempre que preciso e pela sua paciência e aceitação de minhas ausências durante a conclusão de ciclo desta etapa.

À minha orientadora, Silvia Pedro Rebouças, pela sua paciência, dedicação e ajuda possibilitando a realização deste trabalho.

À minha querida professora Luciana Reinaldo, pelo empenho e assistência em um momento crucial, me dando mais segurança neste desafio. Demonstrando o verdadeiro significado de educadora.

Aos meus amigos, Davi Emanuel e Benicio, que me ampararam em diferentes situações, que entenderam e apoiaram este objetivo. Aos amigos de curso, Roberto, Larissa, Rossana e Joana por sua amizade, companheirismo em todo esse tempo de curso, pelas alegrias, tristezas e dores compartilhas, tornando esses 5 anos inesquecíveis, principalmente ao apoio essencial para fazer essa conquista possível.

Ao Rômulo, que atendeu aos meus pedidos de ajuda, sempre prestativo.

Aos meus colegas de trabalho, especialmente ao meu gestor, Thiago, por sua total assistência, paciência e compreensão nessa última etapa de curso, ao Leandro e Matheus M, que ajudaram a finalizar este trabalho.

“Talvez não tenhamos conseguido fazer o melhor, mas lutamos para que o melhor fosse feito. Não somos o que deveríamos ser, não somos o que iremos ser.. mas graças a Deus, não somos o que éramos.”

Martin Luther King

RESUMO

O presente estudo tem como objetivo a aplicação de dois métodos de classificação distintos para a construção de modelos de previsão de desuso de cartão de crédito. As técnicas utilizadas foram Regressão Logística e Árvores de Classificação. Para esse fim, foram utilizadas informações da base de uma empresa de cartão de crédito, compondo uma amostra com 363.172 registros. Como objetivos secundários, essa pesquisa buscou avaliar os indicadores de maiores influencias na análise, comparar os resultados obtidos para indicar o de melhor desempenho. Os resultados obtidos para os dois modelos indicam que é possível indicar com margem aceitável de certeza quais as características podem tornar os clientes desta empresa inativos. De acordo com os critérios escolhidos para a medição da melhor performance entre os modelos não houve uma conclusão muito precisa, pois cada um dos modelos apresentou melhor resultado em um critério especifico. Porém, ambos os modelos apresentaram resultado uniforme para a variável mais importante para o modelo a variante que relaciona a quantidade de compras no último mês com a quantidade de compras nos últimos três meses.

Palavras-chave: *Data Mining*. Modelagem. Classificação de clientes. *Business Intelligence.*

Abstract

This study is based on the application of two different models construction methods that can predict the credit cards disuse. The used methods are Logistic Regression and Classification Tree. The study uses data from a Credit Card Enterprise, totalizing 363.172 records. The study main goals are to recognize the clients characteristics who have a large influence on the credit card disuse, comparing the results with the adjusted models performance. The results obtained for both methods indicate, with a reasonable certainty margin, which characteristics influence the clients to go inactive. Analyzing the criteria used for the best performance, the results are inconclusive on which model to use, because a model can be good for some criteria and can be bad for another criteria. Both models were consistent finding the likelihood ratio between the amount of purchases on the last month and the amount of purchases on the last three months, defining this as the most influential variable for the prediction of credit card disuse.

Keywords: *Data Mining*. Modeling. Classification of clients. *Business Intelligence.*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

Sumário

1 INTRODUÇÃO

Por todo o mundo, o cartão de credito é um dos meios de pagamento mais populares. Permitindo fazer compras físicas e online, pagamentos e levantamentos com segurança e comodidade.

Segundo pesquisa realizada pelo Serviço de Proteção ao Crédito (SPC Brasil, 2013), em junho de 2013, 77% dos brasileiros possuíam pelo menos um cartão de crédito, incluindo cartões de banco e de lojas, e 24% possuíam dois cartões. Conforme os especialistas em finanças pessoais do SPC Brasil (2013), a segurança e a praticidade no momento de parcelar uma compra são os fatores decisivos para que os meios de pagamento eletrônicos sejam cada vez mais utilizados em substituição ao dinheiro.

Ao realizar uma transação comercial com cartão de crédito, o estabelecimento registra a transação, gerando um débito do usuário-consumidor a favor da administradora e um crédito do fornecedor do bem ou serviço contra a administradora.

Porém, ao aceitar uma compra por cartão de crédito, o estabelecimento está aceitando não só a segurança do recebimento do valor daquela compra, como também os gastos embutidos nessa operação, como a taxa de administração, também chamada de taxa de desconto, aluguel das máquinas e a taxa de antecipação.

Para cumprir o crédito que tem para com o fornecedor, a administradora do cartão deve garantir sua capacidade de gerar recursos suficientes através de suas atividades comerciais e conversão de ativos, captando e aprovando de forma independente todos os limites de crédito atribuídos aos clientes, monitorando e gerenciando sua utilização de forma ativa e frequente.

Pode-se concluir então, que a liquidez da administradora do cartão de crédito está diretamente ligada às transações que seus clientes fazem. Logo, é de interesse da empresa o uso contínuo do cartão pelo cliente.

Dentro desse cenário, surge uma questão interessante: como explorar ao máximo o potencial de compra do cliente da empresa?

1.1 Objetivos da Pesquisa

O objetivo deste trabalho é aplicar as técnicas de modelagem, que permitam prever, a partir de características dos clientes e do seu histórico numa empresa de cartão de crédito, se estes são bons ou maus. Neste trabalho consideram-se como bons aqueles clientes que fizeram três ou mais compras em determinado período de análise.

Como objetivos específicos:

a) Comparar os modelos selecionados;

b) Caracterizar os clientes da empresa selecionada;

c) Avaliar a influência de cada variável sócio demográfica, caracterização e histórico na Empresa com o comportamento do cliente;

d) Aplicar modelos de classificação que permitam prever o comportamento do cliente;

e) Comparar o desempenho dos modelos.

1.2 Estrutura da Monografia

Esta monografia é dividida em 5 seções.

A Seção 1 enquadra o contexto de *Data mining* e *Business Intelligence* (BI) concretizado a uma empresa de cartão de crédito e apresenta os objetivos definidos neste trabalho.

Como base para o desenvolvimento do estudo de caso proposto neste trabalho, na Seção 2 é exposto o embasamento teórico sobre o processo de descoberta em bancos de dados, explorando as etapas que compõem alguns métodos que podem ser utilizados na mineração de dados, conhecida também como *Data mining*.

A Seção 3 descreve a metodologia da pesquisa, incluindo uso de classificadores a partir das modelagens em Árvores de Classificação e Regressão Logistica. Também é feita a descrição da população e da amostra, a apresentação das variáveis e a descrição dos métodos utilizados para a análise dos dados, incluindo os métodos para o ajustamento e avaliação do desempenho dos modelos.

A Seção 4 apresenta uma análise descritiva das variáveis, que tem por objetivo caracterizar os clientes e sua relação com a empresa estudada, bem como a estatística para análise de associações, entre a descritiva e multivariada, também os fatores determinantes da classificação dos clientes através das técnicas de Análise Multivariada, pelos métodos de Árvore de Classificação e Regressão (CART) e Regressão Logística.

Por fim, a última seção apresenta as considerações finais deste trabalho.

2 REVISÃO DA LITERATURA

Esta seção apresenta os principais conceitos que envolvem esta pesquisa. Inicialmente se define o termo *Business Intelligence* (BI), em seguida, explica-se *Data mining,* segue-se com a descrição dos métodos de classificação e por fim mostra-se com maior detalhamento as técnicas e as aplicações dos métodos de *Data mining* escolhidos. Bem como uma explicação da estratégia adotada para comparação de desempenho dos modelos.

2.1 *Business Intelligence* (BI)

Angeloni e Reis (2006) definem:

> O conceito de *Business Intelligence* com entendimento de que é Inteligência de Negócios ou Inteligência Empresarial compõe-se de um conjunto de metodologias de gestão implementadas através de ferramentas de software, cuja função é proporcionar ganhos nos processos decisórios gerenciais e da alta administração nas organizações, baseada na capacidade analítica das ferramentas que integram em um só lugar todas as informações necessárias ao processo decisório. Reforça-se que o objetivo do *Business Intelligence* é transformar dados em conhecimento, que suporta o processo decisório com o objetivo de gerar vantagens competitivas.

Para Wanderley (1999), um processo de inteligência de negócios pode propiciar à empresa: antecipar mudanças no mercado; antecipar ações dos competidores; descobrir novos ou potenciais competidores; conhecer as empresas concorrentes; conhecer sobre novas tecnologias, produtos ou processos que tenham impacto no seu negócio; entrar em novos negócios; rever suas próprias práticas e auxiliar na implementação de novas ferramentas gerenciais.

Outro conceito muito utilizado para o BI é o de Barbieri que define BI como:

> Um guarda-chuva conceitual, visto que se dedica à captura de dados, informações e conhecimentos que permitam às empresas competirem com maior eficiência em uma abordagem evolutiva de modelagem de dados, capazes de promover a estruturação de informações em depósitos retrospectivos e históricos, permitindo sua modelagem por ferramentas analíticas. Seu conceito é abrangente e envolve todos os recursos necessários para o processamento e disponibilização da informação ao usuário. (ANGELONI e REIS, 2006).

Sob o ponto de vista empresarial, a informação deve ser encarada como um bem patrimonial da empresa, devendo a mesma ser utilizada de uma maneira estratégica, para que possa atender e atingir rapidamente os objetivos, metas e desafios traçados pela alta gerência de um negócio. “Esta velocidade de mudança faz com que qualquer negócio possa aproveitar uma oportunidade de competição de mercado, sabendo que as informações estratégicas, táticas e operacionais estão disponíveis a qualquer momento para a tomada de decisões.” (MACHADO; ABREU, 2004 *apud* SOUZA, 2007).

O conceito de BI, em síntese, passa pelo desafio da disponibilização de ferramentas e dados para que o nível gerencial de uma organização possa detectar tendências, e tomar decisões eficientes no tempo correto. Com a competitividade crescente em praticamente todos os segmentos do mercado, encontrar um diferencial é fundamental para qualquer empresa, assim como desenvolver estratégias eficazes e tomar decisões inteligentes em menor tempo.

Ansoff (1977) define a decisão estratégica como a que se preocupa principalmente com problemas externos. As decisões táticas preocupam-se com a estruturação dos recursos da empresa. As decisões operacionais visam maximizar a eficiência do processo de conversão dos recursos e a rentabilidade das operações correntes. Embora distintas todas as decisões interagem entre si, são interdependentes e complementares.

A busca pelo controle das variáveis envolvidas para maximização dos resultados gera a necessidade de acompanhamento e gestão das diversas áreas da empresa. Geralmente, as variáveis envolvidas nas áreas produtivas estão correlacionadas, o que impossibilita o controle individual destas variáveis. Diante disso, pode-se dizer que a tomada de decisão inicia-se pelo armazenamento das informações de toda a empresa. A análise deste banco de dados torna-se elemento obrigatório na tomada de decisão.

Laudon e Laudon (2001, *apud* Souza e Neiverth, 2006) destacam que a revolução do conhecimento e da informação começou na virada do século XX e evoluiu gradativamente. A história do *Business Intelligence*, surgiu por volta da década de 50, quando os computadores da época deixaram de ocupar salas inteiras e passaram a armazenar os dados capazes de auxiliar a tomada de decisões.

Naquela época, contudo, os recursos de *hardware* e *software* eram limitados, e a eficiência de transformação de dados em informações ainda não era satisfatória. Foi apenas por volta da década de 70, quando houve uma grande evolução nas formas de armazenamento de dados, que foi possível reunir as informações em um único espaço. Essa reunião era feita através da tecnologia de "Sistema Gerenciador de Banco de Dados" (SGDB), o que fazia com que as ferramentas de BI da época pudessem oferecer aos gestores as informações pretendidas.

Segundo Primak (2008), o termo *Business Intelligence*® surgiu apenas na década de 80, cunhado pela empresa *Gartner Group*. O mesmo autor definiu o *Business Intelligence*® "como o processo inteligente de coleta, organização, análise, compartilhamento e monitoração de dados contidos em *Data Warehouse* e/ou *Data Mart*, gerando informações para o suporte à tomada de decisões no ambiente de negócios".

O setor corporativo, ainda segundo o autor, passou a se interessar pelas soluções de BI de forma mais expressiva, principalmente por final dos anos de 1996, quando o conceito começou a ser espalhado como processo de evolução do *Executive Information Systems* (EIS).

Com a evolução da tecnologia, o termo *Business Intelligence* ganhou mais popularidade embutido com uma série de ferramentas, como planilhas eletrônicas, geradores de consultas e de relatórios, *Data Mart, Data mining*, que têm como objetivo, segundo Primark (2006), facilitar e agilizar a atividade comercial, dinamizar a capacidade de tomar decisões e refinar estratégias de relacionamento com os devidos clientes, respondendo as necessidades da corporação.

2.2 Data mining

Num ambiente extremamente mutável, torna-se necessária à aplicação de técnicas e ferramentas que agilizem o processo de extração de informações relevantes de grandes volumes de dados. O *Data mining* veio preencher essa lacuna na necessidade de análise que ultrapassa a habilidade e a capacidade humana (AMO, 2003).

O núcleo central do processo de prospecção de conhecimento é composto pelos métodos de mineração de dados (*Data mining*). O *Data mining* é o suporte ideal para a gestão de negócios, coletando e reunindo todos os dados da organização, assim como indicadores e métricas da performance da empresa, e transformando-os em informação.

Primak (2008) menciona que o *Data mining*

> (...) Está mais relacionado com processos de análise de inferência do que com os de análise dimensional de dados, representando assim uma forma de busca de informação baseada em algoritmos que objetivam o reconhecimento de padrões escondidos nos dados e não necessariamente revelados pelas outras abordagens analíticas, como o OLAP (cubo).

Fayyad (1996) descreve o *Data mining* como o processo não trivial de identificar, em dados, padrões válidos, novos, e potencialmente úteis e compreensíveis. O *Data mining* está relacionado à aplicação de algoritmos que, mediante limitações computacionais, são capazes de reproduzir uma relação particular de padrões a partir de grandes massas de dados.

A expressão *Data mining* surgiu pela primeira vez em 1990 em comunidades de bases de dados. A mineração de dados é a etapa de análise do processo conhecido como *Knowledge Discovery in Databases* (KDD), sendo a sua tradução literal "Descoberta de Conhecimento em Bases de Dado".

A partir do século XXI houve um aprimoramento nas ferramentas de *software* com oferecimento de informações precisas e no momento correto para alinhar ações de melhoria de desempenho das empresas.

Em termos gerais, segundo Elmasri e Navathe (2002), as técnicas de *Data mining* compreendem os seguintes propósitos:

a) Previsão - pode mostrar como certos atributos dentro dos dados irão comportar-se no futuro.
b) Identificação - Padrões de dados podem ser utilizados para identificar a existência de um item, um evento ou uma atividade.
c) Classificação - o *Data mining* pode repartir os dados de modo que diferentes classes possam ser identificadas com base em combinações de parâmetros.
d) Otimização - aperfeiçoar o uso de recursos limitados e maximizar variáveis de resultado como vendas ou lucros sob um determinado conjunto de restrições.

Segundo Tarapanoff *et al* (2001), Elmasri e Navathe (2002), e Amo (2003) o conhecimento descoberto durante a fase de *Data mining* pode ser descrito de acordo com cinco tarefas: análise de regras de associação, classificação e predição, análise de padrões sequenciais, análise de *clusters* e análise de *outliers*. O Quadro 1 detalha as cinco tarefas do *Data mining* e seus conceitos.

Quadro 1 – Tarefas do *Data mining(continua)*

Tarefa	**Conceito**
Análise de regras de associação	Padrão da forma X → Y, onde X e Y são conjuntos de valores. Aplica-se, por exemplo, nos casos em que se deseja estudar preferências, visando criar oportunidades para formação de grupos de consumidores.
Classificação e predição	O processo de criar modelos (funções) que descrevem e distinguem classes ou conceitos, baseados em dados conhecidos, com o propósito de utilizar estes modelos para predizerem a classe de objetos que ainda não foram classificados.

Quadro 1 – Tarefas do *Data mining(continuação)*

Tarefa	Conceito
Análise de Padrões Sequenciais	Estuda uma expressão da forma $\{I_{1,}, \ldots., I_n\}$, onde cada I_i é um conjunto de itens. Estes conjuntos estão alinhados de forma cronológica para explicar se um comportamento particular em um dado momento pode ter como consequência outro comportamento ou sequência de comportamentos.
Análise de *Clusters* (agrupamentos)	A análise de *clusters* trabalha sobre dados onde as classes não estão definidas. Os registros são agrupados em função de suas similaridades básicas.
A análise de *outliers*	Estuda os dados que não apresentam o comportamento da maioria(exceções). Muitos métodos de mineração descartam estes *outliers* como sendo ruído indesejado.

Fonte: Elaborada pela autora

2.2.1 Classificação

Classificação é um método de mineração de dados cujo objetivo é classificar elementos de um conjunto de dados em diferentes classes, esse método pode ser utilizado em aplicações que incluem diagnósticos médicos, avaliações de risco em empréstimos, detecção de fraudes, etc. Basicamente, o que é feito matematicamente, é reconhecer as principais fontes de variação de muitas variáveis e tornar as informações interpretáveis.

Quando a variável dependente é do tipo dicotômica, como é o caso deste trabalho, recomenda-se utilizar técnicas de estatísticas de análise multivariada como árvores de decisão, regressão logística, redes neurais, vizinhos mais próximos e análise discriminante.

Desta forma, utilizou-se para este trabalho as técnicas de regressão logística e árvores de classificação, para traçar um perfil de clientes que são mais propensos a deixar de comprar utilizando o cartão de crédito na Empresa X.

2.3 Regressão Logística

Segundo Corrar, Paulo e Dias Filho (2007), a técnica de regressão logística foi desenvolvida por volta da década de 1960, em resposta ao desafio de explicar a ocorrência de determinados fenômenos quando a variável dependente fosse de natureza binária.

A regressão logística é uma técnica estatística utilizada para descrever o comportamento entre uma variável dependente binária e variáveis independentes métricas ou não métricas. Ou seja, destina-se a investigar o efeito das variáveis pelas quais os indivíduos, estão expostos sobre a probabilidade de ocorrência de determinado evento de interesse.

Segundo Frota (2011), o uso da regressão logística tem estado presente nas duas últimas décadas para estimar a probabilidade de eventos dicotômicos, com aplicações em economia, medicina, análise de risco e tomadas de decisão.

Segundo Fávero, Belfiore *et al* (2009), a vantagem da regressão logística diante das outras técnicas reside na flexibilidade de seus pressupostos, o que amplia sua aplicabilidade.

A função logística, $f(Z) = \frac{1}{1 - e^{-(z)}}$, assume valores entre 0 e 1, para qualquer Z entre -∞ e +∞. Assim, a popularidade desta técnica advém não apenas da possibilidade de prever a ocorrência de eventos de interesse, mas também da capacidade de apresentar a probabilidade de sua ocorrência.

A função logística só foi reconhecida pelo mundo acadêmico-científico em 1920, a partir de estudo desenvolvido por Pearls e Reed a respeito do crescimento da população norte americano. Em artigo publicado na revista *Proceedings*, da *Belgian Royal Academy, Pierre François Verhulst* (1945) definiu uma função para tratar do crescimento exponencial da população e a nomeou função logística, devido ao diagrama da curva ser parecido com a curva logarítmica atualmente denominada de exponencial.

A regressão logística tem sido uma das principais ferramentas na modelagem estatística de dados, sendo largamente utilizada em diversos tipos de problema. Paula (2002) explica:

> Mesmo quando a resposta não é originalmente binária, alguns pesquisadores têm dicotomizado a variável resposta de modo que a probabilidade de sucesso possa ser modelada por intermédio da regressão logística. Tudo isso se deve, principalmente, à facilidade de interpretação dos parâmetros de um modelo logístico e também pela possibilidade do uso desse tipo de metodologia em análise com objetivo de discriminação.

Os modelos logísticos surgiram da necessidade de modelos mais satisfatórios para dados qualitativos. O modelo de regressão logística é o principal modelo em que a variável resposta assume resposta binária. A regressão logística é semelhante à regressão linear. Em ambos os casos utiliza-se uma ou mais variáveis explicativas (X) para predizer o valor da variável resposta (Y). Adota-se usualmente o valor 1 como aquele que se pretende relacionar ao acontecimento de interesse, sucesso e 0 ao "fracasso".

Foram encontrados exemplos de aplicação da regressão logística combinada ou comparada com outro tipo de classificação em várias áreas durante a pesquisa deste trabalho.

Palmuti e Picchiai (2012) aplicaram a técnica a uma amostra de empreendedores de uma instituição de crédito popular. O modelo dessa pesquisa considerou como variável dependente a "qualidade do crédito", que possui duas categorias: adimplente e inadimplente, sendo inadimplente o cliente que se encontrava em atraso há um período superior a 30 dias na data da coleta dos dados. O modelo final teve taxa de acertos de 87,4%. Dessa forma, o modelo de Palmuti e Picchiai (2012) observou que as variáveis do perfil não possuem influência na qualidade do crédito, mas, sim, variáveis relacionadas diretamente ao risco (juros, valor da prestação, renda etc.).

A classificação de clientes também tem sido alvo de pesquisas com aplicação de regressão logística. Camargos, Araújo e Camargos (2012) de forma análoga, desenvolveram um modelo de regressão logística para a classificação de clientes de uma instituição bancária pública do estado de Minas Gerais. A variável dependente qualidade do crédito considerou como inadimplentes os clientes com atraso superior a 90 dias no pagamento de pelo menos uma das parcelas do financiamento; os adimplentes, por sua vez, foram considerados aqueles que não estavam atrasados em nenhuma parcela do financiamento. O modelo final de Camargos, Araújo e Camargos (2012) mostrou taxa de acerto geral obtida de 67,7%, enquanto que para os adimplentes foi de 68,7% e para os inadimplentes foi de 49,7%, considerando o ponto de corte de 0,06.

A regressão logística tem sido aplicada também em pesquisas de prevenção de insolvência/inadimplência. Casa Nova *et al* (2013) realizou um estudo comparando diferentes técnicas para avaliação de insolvência/inadimplência de empresas de pequeno e médio porte dos setores industrial e de comércio. Foram aplicadas as técnicas de análise envoltória de dados, redes neurais e regressão logística, comparando-se os resultados de classificação. O modelo de regressão logística no caso das empresas industriais obteve taxa de acerto geral, variando de 62% a 68%, para o modelo que utilizou escores fatoriais; porém, quando se considera a classificação das empresas em adimplentes e inadimplentes, observou-se que em

torno de metade das empresas inadimplentes foram classificadas corretamente (43% em 2001, 53% em 2002 e 57% em 2003). No caso do setor comercial, o modelo que utilizou os indicadores originais obteve melhor taxa de acertos, variando de 62% a 68% nos anos de 2001 a 2003. Em relação às empresas más pagadoras, observou-se taxa de acerto variando de 46% a 58% nos anos de 2001 a 2003. A autora destaca que, no caso das empresas comerciais, "o modelo logístico não acrescentou informação relevante à decisão, pois se todas as empresas fossem classificadas como adimplentes, a taxa de acerto seria de 54%".

Para a classificação de empresas também se pode aplicar algum modelo regressivo. Soares, Coutinho e Camargos (2012) realizaram um estudo similar ao de Casa Nova (2013), construindo um modelo de crédito por meio de um modelo logístico multinominal para classificar empresas em quatro classes com base em indicadores contábeis. A taxa de acertos geral foi de 59,7%, com acertos de 40% para a classe 4 de rating; 88% para a classe 3; 3,5% para a classe 2; e 0% para a classe 1.

Selau e Ribeiro (2009 91 *apud* GONÇALVES et al, 2013, p 148) realizaram um estudo com clientes de cartão de crédito de uma rede de farmácias no estado do Rio Grande do Sul, considerando uma amostra de 11.681 clientes. Foram considerados bons pagadores aqueles que tiveram atrasos de até 30 dias e maus pagadores aqueles que tiveram atrasos superiores há 60 dias. A taxa de acerto geral do modelo foi de 73%.

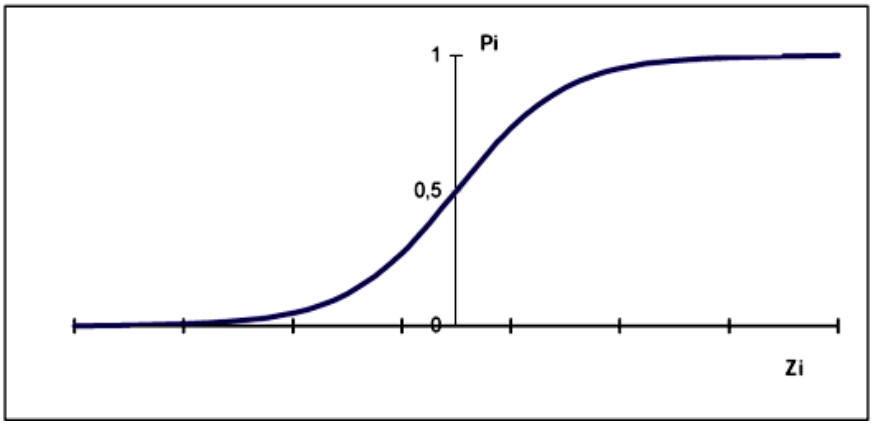
2.3.1. A função Logística

A função logística f(Z) descreve a forma matemática na qual o modelo de logístico se baseia, conforme mostra Equação 1.

$$f(Z) = \frac{1}{1 + e^{-(z)}} \qquad \text{ou} \qquad f(Z) = \frac{e^{Z}}{1 + e^{z}} \qquad (1)$$

A Figura 1 ilustra o comportamento da regressão logística. Observa-se que, independente do valor de *z*, a amplitude de resultados de *f(z)* está entre 0 e 1, ou seja , $0 \leq f(z) \leq 1 \; \forall - \infty \leq z \leq + \infty$.

Figura 1 – Função Logística Acumulada



Fonte: *Scientific Eletronic Library Online*

Função logística:

$$f(Z) = \frac{1}{1+e^{-(Z)}} = \frac{1}{1+e^{-(a+\sum B_i X_i)}} \quad (2)$$

Em termos probabilísticos, temos:

$$P(\mathbf{X}) = f(Y = 1|X_1, X_2, \dots, X_p) = \frac{1}{1+e^{-(\alpha+\sum \beta_i X_i)}} \quad (3)$$

Onde:

a) P(X) = probabilidade condicional de ocorrer o evento Y = 1, conhecido no modelo logístico, dado vetor X**.**

b) $X_1, \dots, X_k$ ou o vetor X : variáveis independentes.

c) α *e* β_i: parâmetros desconhecidos que serão estimados pelo modelo.

O parâmetro α domina P(X) quando X_i *s*ão zero, e β_i ajustam a taxa de modificação de P(X) com relação às variáveis independentes. O modelo logístico pode ser linearizado. Tal fato facilita a obtenção dos parâmetros (α *e* β_i).

Para o caso de variáveis dicotônicas, pode-se eliminar o limite superior $P(x) = 1$, utilizando a razão $\frac{P(x)}{1-P(x)}$. Esta razão é positiva, pois $0 \leq P(x) \leq 1$, porém não há limite superior. Quando $P(x) \to 1$ a razão $\frac{P(x)}{1-P(x)} \to \infty$. O limite inferior pode ser eliminado aplicando o logaritmo natural, ou seja, $\ln(\frac{P(x)}{1-P(x)})$. O resultado está contido no seguinte intervalo $-\infty \leq \ln\frac{P(x)}{1-P(x)} \leq +\infty$ de forma semelhante aos parâmetros (α *e* β_i).

Transforma-se a probabilidade de P(x), então:

$$logit\ P(x) = \ln\frac{P(x)}{1-P(x)} \quad (4)$$

Essa transformação é chamada de *logit (logistic probability unit),* termo criado por Berkson (1944). O termo $\frac{P(x)}{1-P(x)}$, na *logit,* é chamado de *Odds* (chance). Ao substituir $P(x) = \frac{1}{1+e^{-(\alpha+\sum \beta_i X_i)}}$ em 1- P(x) , encontra-se:

$$1 - P(x) = 1 - \frac{1}{1+e^{-(\alpha+\sum \beta_i X_i)}} = \frac{e^{-(\alpha+\sum \beta_i X_i)}}{1+e^{-(\alpha+\sum \beta_i X_i)}} \quad (5)$$

A razão em *Odds* resulta em:

$$\frac{P(x)}{1-P(x)} = \frac{\frac{1}{1+e^{-(\alpha+\sum \beta_i X_i)}}}{\frac{e^{-(\alpha+\sum \beta_i X_i)}}{1++e^{-(\alpha+\sum \beta_i X_i)}}} = e^{(\alpha+\sum \beta_i X_i)} \quad (6)$$

Aplicando a Equação 6, na Equação 4, temos:

$$logit\ P(x) = \ln \frac{P(x)}{1-P(x)} = \ln\left[e^{(\alpha+\sum \beta_i X_i)}\right] = (\alpha + \sum \beta_i, X_i) \quad (7)$$

Um conceito importante é a razão de *Odds* na transformação *logit,* em uma forma no qual, representa a chance de ocorrência do evento de interesse. Com o valor *logit,* as variaveis independentes podem assumir valores positivos e negativos, mas sempre transformada de volta em um valor de probabilidade entre 0 e 1. Contudo, o *logit* jamais pode realmente alcançar 0 ou 1.

Ao aplicar os conceitos acima ao caso "classificação de clientes", tem-se a probabilidade de um cliente deixar ou não de utilizar o seu cartão de credito da Empresa X em função dos valores de x (sexo ou idade, por exemplo).

O diferencial da regressão logística se dá pelo fato das variáveis serem tratadas neste estudo como variáveis individuais ou categorias únicas e são relacionadas com outras variáveis independentes. Neste contexto o desenvolvimento da regressão logística é semelhantes aos demais modelos de regressão, ou seja, encontrar o melhor modelo que relacione a variável dependente com as variáveis independentes. Dito isto, a variável dependente *z* torna-se um índice que agrega as *p* variáveis independentes, que podem ser quantitativas ou qualitativas. Uma variável com *c* categorias pode ser representada com *c*-1 variáveis mudas, uma vez que, se todas assumirem o valor 0, significa que a categoria observada é a categoria de referencia.

2.3.2 Pressupostos do modelo de regressão logística

Um modelo baseado em regressão logística assume que: i) as variáveis independentes são multicolineares, ii) os resíduos são independentes e apresentam distribuição binomial; iii) linearidade e aditividade; iv) proporcionalidade, ou seja, a

contribuição de cada variável $X_i, i = 1, \dots, p$ é proporcional ao seu valor com um fator β_i; v) constância de efeito, ou seja, a contribuição de uma variável independente é constante e independente da contribuição das outras variáveis independentes.

A validação destes pressupostos pode fazer-se agraves da análise dos resíduos. A multicolinearidade pode ser encontrada através do coeficiente de determinação ($T = 1 - R^2$), obtido pela regressão linear múltipla entre cada variável independente e as variáveis independentes restantes do modelo.

2.3.3 Seleção das Variáveis

Neste trabalho, inicialmente, todas as 23 variáveis foram incluídas para construção do modelo. A escolha das variáveis foi feita por intermédio do método *forward stepwise*, que é o mais largamente utilizado em modelos de regressão logística.

O método de seleção *forward stepwise* adiciona-se uma variável de cada vez, selecionando em primeiro lugar aquela que apresentar um valor de correlação mais elevado, em módulo, com a variável resposta, e assim consequentemente, até que o processo pare quando o aumento do coeficiente de determinação, devido à inclusão de uma nova variável explicativa no modelo não é mais importante.

2.3.4 Estimação e Interpretação dos coeficientes

Para estimação do modelo da regressão logística pode-se utilizar o método de máxima verossimilhança, que estima os coeficientes de regressão que maximizam a probabilidade de encontrar as realizações da variável dependente. Uma vez que Y_i segue uma distribuição Bernoulli com parâmetro π_j (probabilidade de Y = 1) e assumindo independência, a função de verossimilhança é dada por:

$$L = P\,(Y_1 = y_1)\,X \dots . X\,P(Y_n = y_n) = \prod_{j=1}^{n} \pi_j^{y_j}\,(1 - \pi_j)^{1-y_j} \qquad (8)$$

A partir do processo de estimação descrito, sabe-se que os coeficientes $(B_0, B_1, \dots, B_n)$ são medidas das variações nas proporções das probabilidades.

A direção da relação reflete mudanças na variável dependente associadas à variável independente. Uma relação positiva significa que um aumento na variável independente é associado com um aumento na probabilidade prevista, e vice-versa para uma relação negativa.

2.3.5 Significância e Qualidade do modelo de regressão logística

Sendo o objetivo avaliar o "bom" ajuste do modelo construído através da regressão logística, pode-se fazê-lo usando representações gráficas dos valores dos resíduos. Este caso permite comparar os resíduos dos vários elementos. Pode-se ainda aplicar testes baseados em estatísticas desses valores, fundamentados no valor da estatística de teste e avaliando a qualidade do ajuste do modelo de uma forma global (MARTINS, 2008).

2.3.5.1 Teste do Rácio de Verossimilhanças

O teste de rácio verossimilhança compara os valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem a variável em questão. A comparação dos observados com os valores preditos é baseada no log da verossimilhança. Para entender melhor essa comparação, é útil pensar em um valor observado da variável resposta também como sendo um valor predito resultante de um modelo saturado. Um modelo saturado é aquele que contém tantos parâmetros quanto observações.

Podemos testar a significância do modelo ajustado com:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ Isto é, o modelo não é estatisticamente significativo.

$H_1: \exists_i: \beta_i \neq 0, i = 1, \ldots, p$, isto é, o modelo é estatisticamente significativo.

Hair *et al* (2005) explicam que "A medida geral de quão bem o modelo se ajusta é dada pelo valor de verossimilhança, (na verdade é -2 vezes o logaritmo do valor da verossimilhança e é chamado de -2LL)". Sendo LL o logaritmo natural de Lo/Lx, conforme mostra Equação 9.

$$Estatística\ de\ teste: -2\ln\frac{L_0}{L_x} \sim X_p^2 \qquad (9)$$

Onde Lo é a verossimilhança do modelo nulo (somente a constante) e o Lx é a verossimilhança do modelo completo. Dito isto, quando menor o valor de -2LL melhor ajustamento do modelo.

2.3.5.2 Teste de significância dos coeficientes do modelo e medidas de ajustamento

A conclusão de que o modelo ajustado é significativo, implica que existe pelo menos uma variável independente com poder explicativo. Para a identificação de qual variável ou quais variáveis independentes tem poder explicativo, é usual recorrer ao teste Wald, que é simplesmente um teste de escore z, onde:

Testa-se as hipóteses:

$H_0: \beta_1 = 0$

$H_1: \beta_i \neq 0$

As estatísticas do Qui-Quadrado de Wald:

$$X^2_{wald_i} = \left(\frac{\widehat{\beta}_i}{se(\widehat{\beta}_i)}\right)^2 \sim X^2_i \quad (10)$$

Os testes dos coeficientes são aproximadamente escores z, os quais são posteriormente elevados ao quadrado, fazendo com que esta estatística tenha distribuição de qui-quadrado. Esse teste é usado para avaliar a significância de cada coeficiente (β) no modelo.

Ainda podem ser calculadas medidas de avaliação da qualidade de ajustamento do modelo, com analogia ao coeficiente de determinação R^2 de um modelo de regressão linear múltipla, tais como:

Cox & Snell R^2

$$R^2_{CS} = 1 - (\frac{L_0}{L_\beta})^{\frac{2}{N}} \quad (11)$$

$$R^2_{CS_{MÁX}} = 1 - (L_0)^{\frac{2}{N}} \quad (12)$$

Nunca atinge valor 1, mesmo quando o ajustamento é perfeito.

Nagelkerke R^2:

$$\widetilde{R^2_N} = \frac{R^2_{CS}}{R^2_{CS_{MÁX}}} \quad (13)$$

Esses métodos podem ser vistos como uma aproximação da alteração na variável dependente, devido à mudanças nas variáveis independentes.

2.4 Árvores de Classificação

Segundo Basgalupp (2010), o algoritmo de Árvores de Classificação e Regressão, foi proposto por Breiman *et al* em 1984 e consiste em uma técnica não paramétrica que induz tanto árvores de classificação, caso a variável dependente seja categórica, quanto árvores de regressão, caso a variável dependente seja contínua.

Para Rebouças (2001), as Árvores de Classificação e Regressão, CART, podem ser consideradas como modelos de regressão não paramétricos, que têm como objetivo estabelecer uma relação entre o vetor de variáveis independente e a variável resposta.

Segundo Taconelli, Zocchi e Dias, (2009 *apud* SOARES, 2013), uma das maiores virtudes da CART é a capacidade de pesquisa de relações entre os dados, mesmo que não

sejam evidentes. O método CART baseia-se na execução de partições binárias sucessivas de uma amostra, com base nos resultados amostrados das variáveis independentes. A classificação dessas subamostras é realizada conforme alguma medida descritiva e a predição de novos elementos, executada por meio da estrutura de classificação.

Para Rebuças, os modelos baseados em árvores possuem características que têm contribuído para o crescimento da sua popularidade, tais como facilidade de interpretação, tratamento de dados ausentes e captura automática de interações (não explícitas) entre as variáveis explicativas. Deste modo, as árvores são bastante úteis na organização da informação e podem ser utilizadas para criar modelos de atribuição de crédito ou de outros riscos financeiros.

A utilização de CART apresenta as seguintes vantagens: não assume nenhuma distribuição particular para os dados; as características ou atributos podem ser qualitativos ou quantitativos; pode construir modelos para qualquer função desde que o número de exemplos de treinamento seja suficiente; possui elevado grau de interpretação.

Segundo Steiner *et al* (2004) para gerar uma árvore de classificação com uma alta taxa de predição é necessário fazer a escolha correta dos atributos que serão usados como teste no agrupamento dos casos. Estes testes devem gerar uma árvore com o menor número possível de subconjuntos. O ideal é escolher os testes de modo que a árvore final seja a menor possível.

Um das técnicas mais utilizadas nos estudos de *Data mining* são as Árvores de Classificação e Regressão. As árvores de classificação podem ser consideradas como modelos de regressão não paramétricas, que tem por objetivo estabelecer alguma relação entre as variáveis independentes e a variável resposta, dependente.

Esses tipos de modelos são purificados conforme sucessivas divisões no conjunto de dados, de modo a tornar os subconjuntos resultantes mais homogêneos. Essas subdivisões são apresentadas através da estrutura de árvore, na qual cada "nó" corresponde a uma divisão.

Após a construção da árvore as variáveis inseridas podem assumir valores contínuos ou categóricos. A árvore resultante deste trabalho é designada árvore de classificação, porque a variável dependente é qualitativa, caso contrário, quando a variável dependente é quantitativa trata-se de um modelo com Árvore de Regressão.

Os componentes de uma árvore de classificação são os nós e as regras de divisão (*splitting rules*). Os nós estão associados aos subconjuntos resultantes da aplicação da regra de divisão ao conjunto de dados. Ao primeiro nó, podemos referir o nome de nó pai ou nó raiz e os nós terminais denominam-se de nós folhas. O algoritmo para a construção da arvore

seleciona um atributo de quebra e divide o conjunto de dados, criando um ramo para cada valor deste atributo. A cada ramo é criado um nó e novamente selecionado um novo critério de quebra. Este processo pretende separar os dados em classes.

O método CART desenvolve-se com o objetivo de maximização da homogeneidade dos nós. A finalidade de um nó terminal puro é em que todos os casos da variável se apresentam com valor igual.

Para facilitar a interpretação da árvore pode-se reduzir sua complexidade, caso esta apresentar uma dimensão elevada. A técnica de poda (*pruning*) reduz sua complexidade sem por em risco a qualidade do ajustamento. Essa técnica retira as regras de decisão menos importantes, sua aplicação é importante para evitar o problema do *overfiting*, ou seja, para evitar o juste excessivo aos dados utilizados para a modelação.

Uma árvore de classificação permite predizer os valores de determinada variável qualitativa (no caso: critério, bom ou mau), ao selecionar as variáveis independentes que melhor se aplicam. Para o ajustamento da árvore, foram incluídas como variáveis independentes todas as variáveis apresentadas na Seção 3.3. Como tratamento de variáveis omissas, utilizou-se da técnica do descarte dos registros e não tendo estes sido utilizados na construção dos modelos.

Foram desenvolvidos vários métodos aplicados na escolha dos atributos e dos testes a serem utilizados, sendo que todos concordam em dois pontos: “uma divisão que mantém as proporções de classes em todas as partições é inútil e uma divisão onde em cada partição todos os exemplos são da mesma classe tem utilidade máxima.” (LEMOS, 2003).

Santos *et al* (2006 *apud* SOARES, 2013, p 32), utilizaram as CART, entre outros modelos, para classificar empresas que estiveram em funcionamento entre 1999 e 2003, todas situadas na região norte de Portugal. Destas, 325 haviam pedido concordata durante o período, enquanto as outras 1.963 permaneceram solventes. Construíram com esses dados quatro modelos utilizando árvores de classificação. Dois modelos consideravam apenas um ano anterior à entrada em insolvência, enquanto os outros dois consideravam toda a informação que precedia o evento, ou seja, três anos. Para todos os modelos o conjunto de dados foi dividido em duas partes de maneira aleatória, sendo uma subamostra usada para o treinamento da árvore, enquanto a outra servia para validá-la.

Rebouças (2001) com o objetivo de desenvolver modelos de análise de risco de crédito (modelos de *Scoring*) para identificação de variáveis que melhor separassem os clientes que cumprem os seus compromissos (Não Devedores), dos que revelam dificuldades no cumprimento dos mesmos (Devedores) aplicou modelos de Árvores de Classificação

CART, Regressão Logística e Redes Neurais. Segundo a autora, o modelo de Regressão Logística apresentou melhores resultados, tanto em termos de qualidade de ajustamento, tanto em relação à capacidade preditiva.

Garcia (2003) empregou árvores de classificação e regressão para descoberta de conhecimento em banco de dados na área de saúde. Como base utilizou a informações pertinentes as autorizações de internações hospitalares (AIHS) emitidas pelos hospitais conveniados ao sistema único de saúde (SUS). Como objetivo buscou-se padrões capazes de determinar situações que levam as AIHs, referentes à acidentes vasculares cerebral (AVC), serem ou não separadas para revisão. Como resultados, a autora, encontrou que os padrões obtidos pelo algoritmo CART eram válidos, após serem confrontados com a prática e que sua aplicação pode ser utilizada para novos casos obtendo resultados confiáveis.

2.4.1 Medidas para selecionar a melhor divisão

Para medir o método de impuridade e decréscimo mínimo na impuridade, em variáveis qualitativas usam-se os métodos de Taxas de Erro de Classificação (*misclassification error*), índice de *Gini* (*Gini Index),* Desviância ou Entropia (*desviance* ou *cross-entropy).* Estas medidas são muito similares, mas o índice de Gini é utilizado quando o algoritmo CART é o escolhido.

O método *Gini* emprega um índice de dispersão estatística proposto em 1912 pelo estatístico italiano Corrado Gini. Este método é definido para uma variável nominal com k categorias, onde $p\ (i/t)$ é a probabilidade a priori da classe *i* se formar no nó *t*. Cada variável pode ser usada diversas vezes ao longo do processo de construção da árvore. Deste modo, este índice contabiliza a proporção de observações em cada classe da variável dependente em um nó relativo ao total, ou seja, o nó raiz. Para uma população 100% pura o índice de Gini seria igual a 1. O método de Gini é definido pela Equação 14:

$$gini_{index}(t) = 1 - \sum_{i=1}^{k} p(i|t)^2 \qquad (14)$$

Para encontrar o valor resultado do índice basta calcular a diferença entre $gini_{index}$ antes e após a divisão. Essa diferença, Gini, é representada pela Equação 15:

$$Gini = gini_{index}(pai) - \sum_{j=1}^{n} \frac{N\,(v_j)}{N} gini_{index}(v_j) \qquad (15)$$

Onde *n* é o número de valores do atributo, ou seja, o número de nós-filhos, *N* é o número total de objetos do nó-pai e $N\left(v_j\right)$ é o número de exemplos associados ao nó-filho v_j. Assim é selecionado o atributo que gerar um maior valor para Gini.

2.4.2 Avaliação de classificadores

Após a construção da árvore é importante avaliá-la. Esta avaliação é realizada através da utilização de dados que não tenham sido usados na primeira amostra para a construção de uma segunda árvore. Esta ação permite mostrar até que ponto a estrutura da árvore ajustada pode ser generalizada a outras observações. Isso pôde ser feito utilizando somente 70% da amostra para construção da árvore modelo, e 30% para árvore de validação.

Esta técnica, denominada de *Houldout,* não requer que seja utilizada duas amostras de dimensões iguais. De acordo com esse método, começa-se por construir a árvore, segundo procedimento descrito anteriormente, utilizando a amostra do modelo. As regras de divisão criadas serão então aplicadas à amostra de validação. A comparação do desempenho do modelo nas duas amostras faz-se a partir dos indicadores de ajustamento, apresentados nas seções mais a frente.

2.5 Tabelas de classificação e curvas de ROC

Para analisar o poder de predição do modelo, é comum a utilização da tabela de classificação. Para elaboração desta tabela, é necessária a escolha do ponto de corte, *c* (*classification cutoff*), onde os valores acima destes pontos indicam a presença do evento de interesse e os valores abaixo, indicam ausência. O Quadro 2 exemplifica uma tabela de classificação.

Quadro 2 – Tabela de Classificação

	Classe Preditiva	
Classe Real	Bom	Mau
Bom	Verdadeiro Bom	Falso Mau
Mau	Falso Bom	Verdadeiro Mau

Fonte: Elaborada pela autora (2014)

Onde:

Verdadeiros “bons” (VB): são os indivíduos que pertencem à classe “bom” e foram corretamente classificados.

Falsos “bons” (FB): são os indivíduos que pertencem à classe mau, porém classificados como bons pelo classificador.

Falsos “maus” (FM): são os indivíduos que pertencem à classe “bom”, porém classificados como “maus” pelo classificador.

Verdadeiros “maus” (VM): são os indivíduos que pertencem à classe “mau” e foram corretamente classificados.

A partir destes indicadores é possível calcular outras medidas de desempenho do classificador, como: acurácia, *recall*, especificidade e precisão.

2.5.1 Acurácia

Estima a probabilidade de o classificador acertar suas previsões. É medida pela Equação 16:

$$Acurácia = \frac{|VB|+|VM|}{|VB|+|VM|+|FB|+|FM|} \quad (16)$$

Pode ser expresso também em termos de taxa de erro, que é o complemento da acurácia, expresso pela Equação 17:

$$Erro = \frac{|FB|+|FM|}{|VB|+|VM|+|FB|+|FM|} \quad (17)$$

2.5.2 Recall e Especificidade

Quando o conjunto de dados contém muitos exemplos de uma mesma classe e poucos de outras, é fácil selecionar sempre a classe majoritária e obter uma boa taxa de acurácia. Então, nesses casos, utilizam-se as medidas se sensibilidade e especificidade como alternativa.

A medida de sensibilidade também é conhecida como *recall,* Equação 18, mede o quão bem um modelo classifica os exemplos positivos:

$$Recall = \frac{|VB|}{|VB|+|FM|} \quad (18)$$

A medida de especificidade estima quão bem um modelo classifica como maus, cujo as classes realmente são maus. Na Equação 19 vemos o cálculo.

$$Especificidade = \frac{|VM|}{|VM|+|FB|} \quad (19)$$

2.5.3 Precisão

Mensura quantos exemplos classificados como bons são realmente pertencentes a classe bom. Essa medida é bastante utilizada quando é preferível que um bom seja classificado como mau, do que o contrario. A Equação 20 mostra o cálculo.

$$Precisão = \frac{|VB|}{|VB|+|FB|} \quad (20)$$

Se para cada *c* conhecido, fosse calculado os indicadores de sensitividade e especificidade, é possível a construção do gráfico conhecido como Curva ROC (*Receiver Operating Characteristic).* Quanto maior a área abaixo da Curva ROC, maior a capacidade do

modelo discriminar o evento de interesse. Além disso, quanto mais próxima a Curva ROC estiver da reta diagonal, pior é o poder discriminatório do modelo.

Segundo Fávero *et al* (2009), uma referência usual em relação à área da Curva ROC é apresentada no Quadro 3.

Quadro 3 – Área abaixo da Curva ROC

Área abaixo da curva ROC	**Interpretação**
Menor ou igual a 0,5	Baixa discriminação
Entre 0,6 e 0,8	Discriminação aceitável
Maior que 0,8	Discriminação excelente

Fonte: Fávero *et al* (2009)

A análise da curva ROC baseia-se na sensibilidade e na especificidade. Para um determinado ponto X, a especificidade é medida pela relação de bons corretamente classificados, e a medida (1 – especificidade), representa os bons classificados como maus. Um gráfico das respostas constitui a curva de ROC.

O cálculo da curva ROC é intuitivo, pois, seja n_1o número de indivíduos com Y = 1 e n_0 o número de indivíduos com Y=0, existem $n_1 * n_0$pares em que os indivíduos Y=1 podem ser combinados com Y=0. Destes pares de números é determinada a proporção das vezes em que os indivíduos Y=1 têm a maior probabilidade.

Quando se considera um teste onde estão presentes populações com indivíduos que apresentam o fator de interesse e indivíduos que não apresentam o fator de interesse é raro a observação de boa separação entre as populações.

2.6 Kolmogorov- Smirnov

Outra medida de ajuste de qualidade de ajuste do modelo é o KS (Kolmogorov – Sminorv), que mede o grau de segregação dos dois grupos (bom e mau).

Lilliefors (1967) descreve o procedimento para testar se um conjunto de N observações deriva de uma distribuição normal. O autor apresenta um teste de hipótese para a medida D que é calculada através da Equação 21:

$$D = máx\,|F(x) - S_N(x)| \qquad (21)$$

Onde $S_N(x)$ é a função distribuição acumulada da amostra, e $F(x)$ é a função distribuição acumulada normal com média e variância igual a amostra. Para a validação dos modelos propostos, pode-se utilizar, conforme descrito por Joseph (2005) o KS sendo a maior

distancia entre a distribuição acumulada da variável do sucesso e fracasso. Nesse caso, KS será dado por:

$$KS = máx\,|F_S - F_{NS}| \quad (22)$$

Onde F_Sé a função distribuição acumulada dos casos do sucesso, e F_{NS} a função distribuição acumulada dos casos de não sucesso.

O valor resultante do KS pode ser interpretado conforme o apresentado no Quadro 4:

Quadro 4 – Tabela de Qualidade do Ajuste do Modelo (KS)

KS	**Interpretação**
Menor que 30	Baixa discriminação
Entre 30 e 50	Discriminação aceitável
Maior que 50	Discriminação excelente

FONTE: Fávero *et al* (2009)

As estatísticas de KS foram então calculadas, conforme mostra a Equação 22 e a máxima diferença entre as distribuições acumuladas de “bons” e “maus” clientes, nos modelos árvore de classificação e regressão logística foi encontrada.

3 METODOLOGIA

Esta seção tem o objetivo de apresentar e descrever os instrumentos e os procedimentos utilizados para a coleta e análise dos dados, justificando o uso de tais instrumentos e procedimentos com base na fundamentação teórica apresentada. Inicialmente classifica-se o tipo de pesquisa adotada. Em seguida são descritas a população e os critérios de escolha da amostra. Segue-se com uma descrição das variáveis e por fim uma explicação dos métodos de *Data mining* escolhidos.

3.1 Tipo de Pesquisa

De acordo com Rodrigues (2007) a pesquisa realizada para construção desse trabalho é do tipo descritiva, visto que "os fatos são observados, registrados, analisados, classificados e interpretados, sem interferência do pesquisador".

Esta pesquisa ainda é um estudo de caso, visto que os dados foram cedidos por uma empresa de cartão de crédito - que aqui chamou-se de Empresa X – à vista disso, os resultados aqui obtidos não podem ser generalizados.

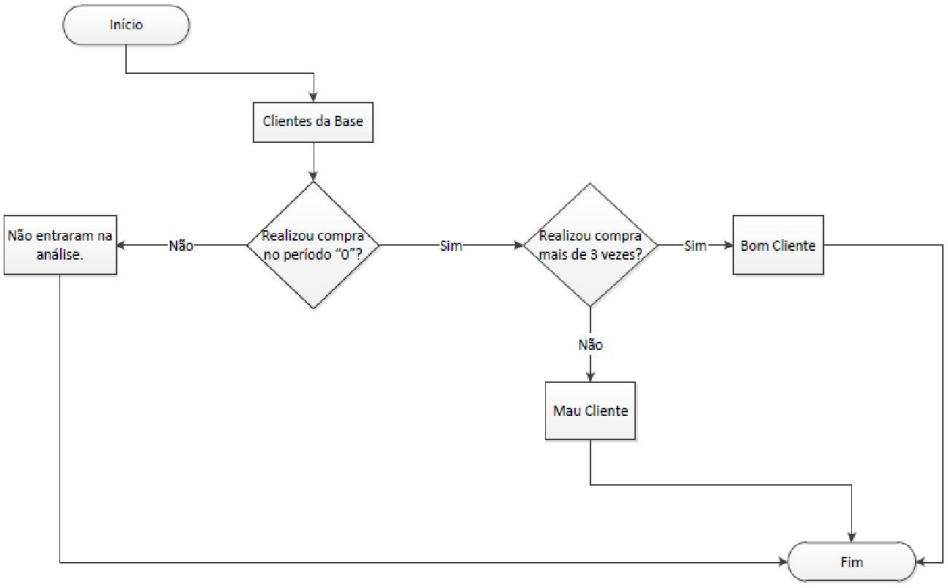
3.2 Os dados

A base de estudo dessa pesquisa contou com observações de clientes, durante seis meses, com inicio do período de observação em Maio de 2012 até Outubro de 2012, resultando, inicialmente, em 524.240 observações.

Os clientes selecionados para análise foram os considerados não inadimplentes no mês de referência, e com no mínimo uma compra nesse mês, e analisou-se seu status três meses depois, caso ainda continuasse não inadimplente nesse terceiro mês, este foi classificado de acordo com seu número de compras nos últimos três meses. Caso o número de compras tenha sido superior ou igual a três, esse cliente foi classificado como "bom cliente", caso inferior a três, "mau cliente".

A Figura 2 mostra o processo de classificação do cliente. Os períodos 0 e 3 apresentados na figura foram a título de exemplo.

Figura 2 – Classificação do cliente



Fonte: Elaborada pela autora (2014)

Os clientes que realizaram pelo menos uma compra no Período 0, foram considerados aptos para a composição da amostra. Esses clientes então, foram analisados 3 períodos a frente, e a seguir classificados como “Bom” ou “Mau” dado o número de compras acumuladas nesses 3 períodos que decorreram. Da seguinte forma:

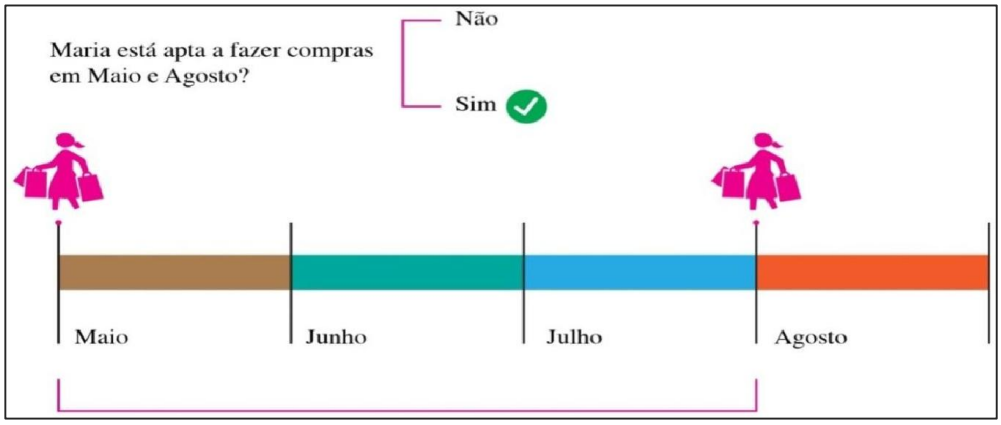
a) Clientes de Maio/12 → analisados em Agosto/12: Se fizeram até 2 compras nos últimos 3 períodos: “Bom” cliente, caso contrário “Mau” cliente.
b) Clientes de Junho/12 → analisados em Setembro/12: Se fizeram até 2 compras nos últimos 3 períodos: “Bom” cliente, caso contrário “Mau” cliente.
c) Clientes de Julho/12 → analisados em Outubro/12: Se fizeram até 2 compras nos últimos 3 períodos: “Bom” cliente, caso contrário “Mau” cliente.
d) Clientes de Agosto/12 → analisados em Novembro/12: Se fizeram até 2 compras nos últimos 3 períodos: “Bom” cliente, caso contrário “Mau” cliente.
e) Clientes de Setembro/12 → analisados em Dezembro/12: Se fizeram até 2 compras nos últimos 3 períodos: “Bom” cliente, caso contrário “Mau” cliente.
f) Clientes de Outubro/12 → analisados em Janeiro/13: Se fizeram até 2 compras nos últimos 3 períodos: “Bom” cliente, caso contrário “Mau” cliente.

Os clientes que estavam ativos no período 0 e entraram em inadimplência em algum período anterior ao período 3, foram desconsiderados. Para os clientes ativos que

estavam em dia com a Empresa X, foram considerados todas as vezes que no Período 0 fizeram pelo menos alguma compra naquele mês e foram classificados diferentemente de acordo com o perfil apresentado naquele período específico.

Para ilustrar o funcionamento básico da classificação dos clientes para composição da amostra, pode-se considerar a Figura 3. De inicio é necessário à escolha ou não do cliente para a composição da amostra. O critério para a entrada na base é se o cliente está apto para fazer compras no período 0 e no período 3. Suponha Maria, uma cliente que pode fazer alguma compra no mês de maio/12 e em agosto/12, de acordo com as premissas da classificação ele pode entrar no modelo de acordo com a resposta da pergunta: "Fez compra?".

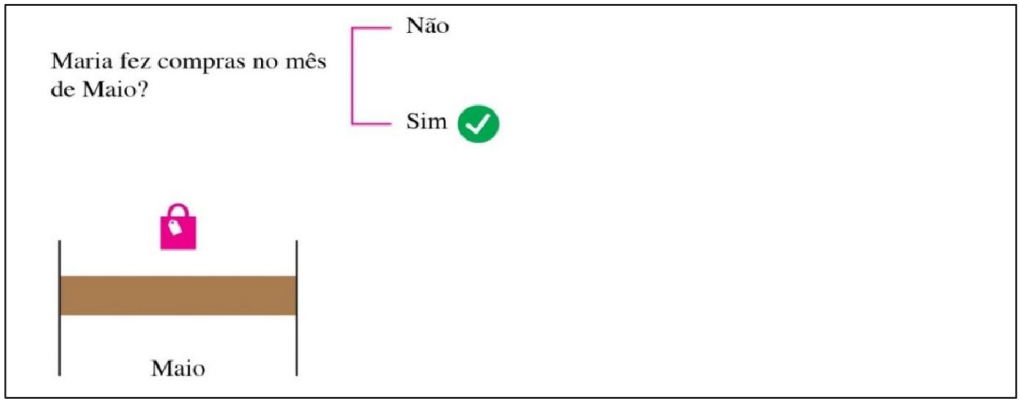
Figura 3 – Verificação de inadimplência da cliente Maria



Fonte: Elaborado pela autora (2014)

Se tiver feito compra, Maria, então, de acordo com as premissas da classificação, já atende o terceiro e último quesito da amostra. A Figura 4 mostra a resposta desse questionamento.

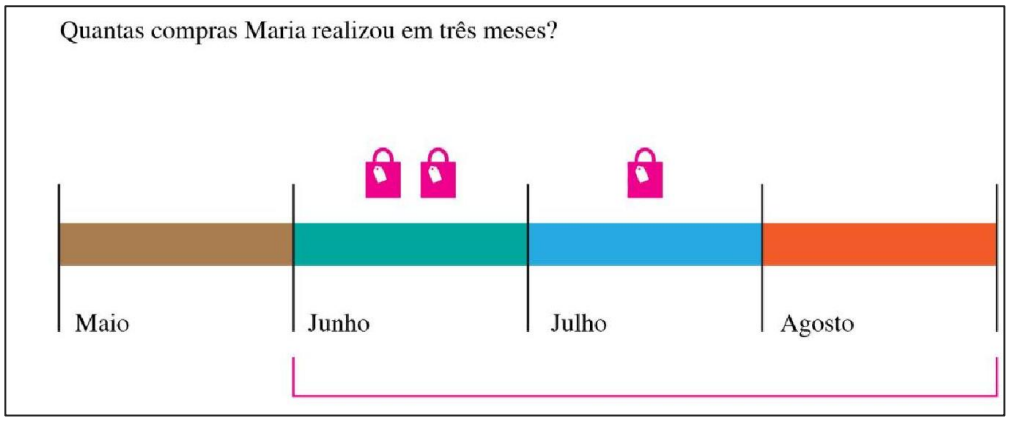
Figura 4 – Quantidade de compras no período 0



Fonte: Elaborado pela autora (2014)

Pode-se então classificar essa cliente, avaliando o somatório da quantidade de compras em três meses. São duas possíveis respostas. A Figura 5 demonstra um possível cenário.

Figura 5 – Quantidade de compras no período de 3 meses

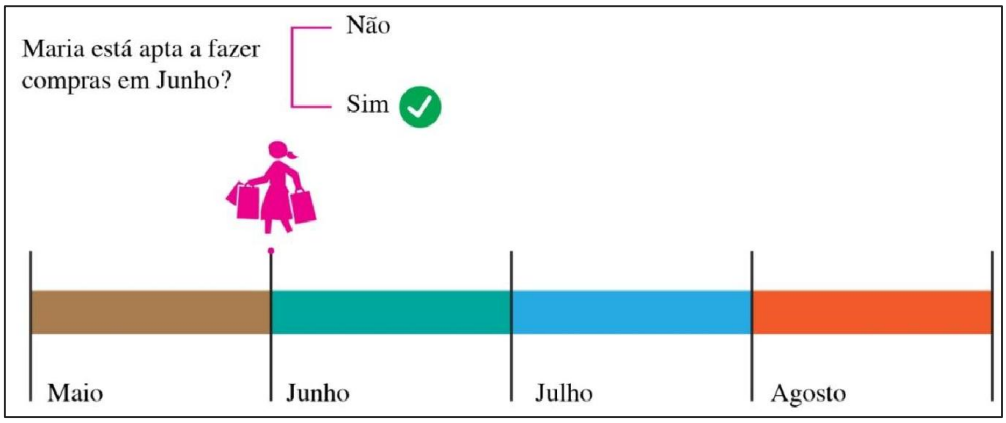


Fonte: Elaborado pela autora (2014)

As “sacolinhas” que estão na Figura 5, representam quantas transações Maria efetuou no período. Como o critério da classificação da variável dependente pode ser denominado bom o cliente que fez três ou mais compras num período de três meses, Maria foi classificada como um “bom” cliente. Ela, no somatório do período de três meses, fez três transações no seu cartão da empresa X.

Como os períodos 0 são consecutivos, um cliente pode entrar mais de uma vez na base do modelo. Porém, ele poderá ter resultados diferentes em cada um. Seguindo o exemplo da Figura 2, Maria, que estava apta a entrar no modelo e foi classificada como “bom” cliente no período de maio/12, novamente foi analisada em Jun/12, conforme mostra a Figura 6.

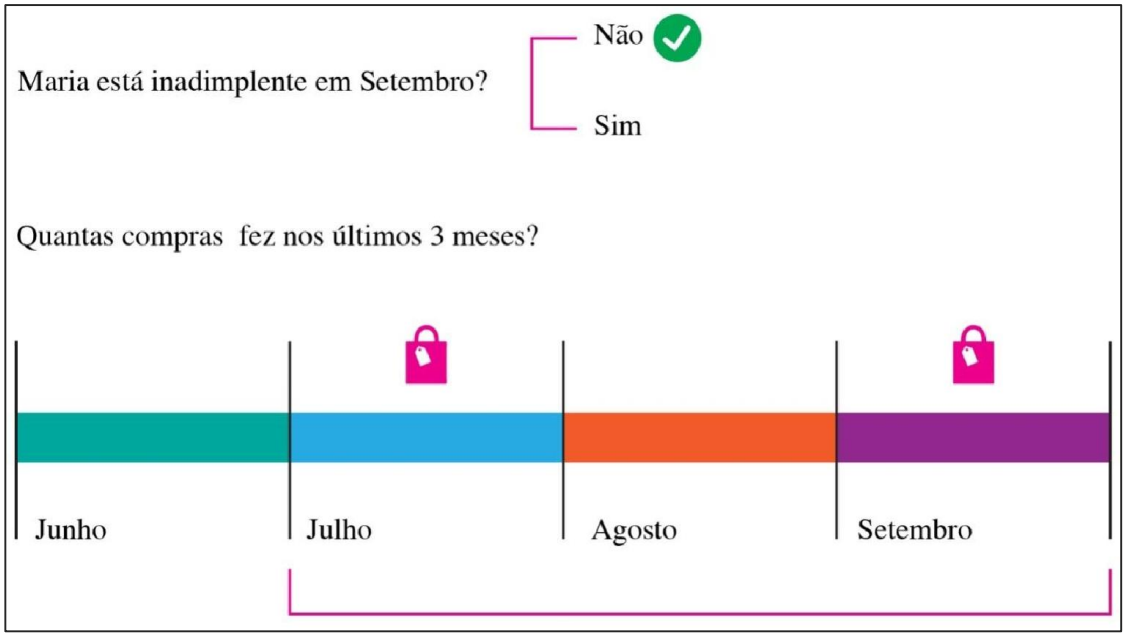
Figura 6 – Situação de inadimplência de Maria em Jun/12



Fonte: Elaborado pela autora (2014)

Conforme a Figura 6, Maria está apta novamente a entrar no modelo, então se avalia o status de inadimplência de Maria em set/12, conforme Figura 7. Maria então está definitivamente no modelo, pela segunda vez, porém, qual sua classificação em Jun/12?

Figura 7 – Maria em set/12 e quantidade de compras em três meses



Fonte: Elaborado pela autora (2014)

Nessa segunda situação, Figura 7, Maria não foi mais classificada como bom cliente, como aconteceu em maio/12, pois seu somatório de transações foi inferior a três. O que pode acontecer de forma igual no terceiro mês (terceiro período 0), caso Maria ainda esteja apta em jul/12, e em out/12 (terceiro período 3) e apresentasse um número diferente de compras durante esses três meses.

3.3 Amostra

A proporção de clientes que foram classificados na amostra inicial como “maus” foi de 37,5%, que para a construção do modelo de Regressão Logística e Arvore de Classificação (CART), de resposta binária mostrou o desempenho na classificação de “clientes bons” bastante satisfatória, porém, muito deficiente na classificação de clientes “maus”.

A solução encontrada para solucionar a alta taxa de classificação errada dos clientes foi a redução da amostra de bons clientes, igualando então, ambas as proporções – 50%/50%. A amostra final de bons clientes foi composta de forma aleatória.

A amostra utilizada neste estudo é composta, então, por 363.172 casos, igualmente dividida entre registros bons e maus.

3.4 Variáveis

Foram consideradas, ainda, 21 variáveis, que podem ser classificadas em sócio demográficas, da relação com a empresa e variáveis relacionadas ao histórico na empresa. Sendo, 8 variáveis sócio demográficas, 7 variáveis da relação com a empresa, e 6 variáveis do histórico do cliente na empresa.

3.4.1 Variáveis sócio demográficas

O Quadro 5 apresenta as variáveis sócio demográficas dos clientes considerados neste trabalho.

Quadro 5 – Variáveis sócio demográficas

Variável	Classificação	Descrição
Sexo	Variável Qualitativa Nominal	2 categorias: Feminino e Masculino
Idade	Variável Quantitativa Continua	Medida em anos
Natureza Ocupação	Variável Qualitativa Nominal	2 categorias: Empregado e Autônomo
Estado Civil	Variável Qualitativa Nominal	2 categorias: Casado e Solteiro
Grau de Instrução	Variável Qualitativa Nominal	5 categorias: Até 1º Grau Completo, Até 2º Grau Completo, Superior Incompleto, Superior Completo e Outros
Se tem dependentes	Variável Qualitativa Nominal	2 categorias: Sim ou Não
Renda comprovada	Variável Quantitativa Continua	Medida em Reais
Tempo de Emprego	Variável Quantitativa Continua	Medida em anos

Fonte: elaborada pela autora (2014)

As variáveis “Estado Civil”, “Natureza Ocupação” e “Dependentes” foram transformadas em binárias. Onde:

- Estado Civil: Inicialmente esta variável possuía 7 classificações. Onde Casado e Companheiro foram agrupados em uma só categoria, chamada de “Casado”, Desquitado, Divorciado, Separado, Solteiro e Viúvo foram agrupados, e renomeados de “Solteiro”.

- Natureza Ocupação: Inicialmente esta variável possuía 5 classificações, Autônomo, Aposentado, Outros, Profissionais Liberais, Empregado. A categoria Autônomo foi isolado e as demais classes foram agrupadas em “Empregados”.

- Dependentes: inicialmente esta variável era quantitativa. Optou-se por transformá-la em qualitativa, tendo-se criado a condição se o cliente possuía adicional ou não.

Essa transformação é proposta por Hunt (1994), em que para uma das possibilidades é atribuído os valores das características eleitas e a outra possibilidade é atribuída aos demais valores. Esta função apresenta a vantagem da simplicidade e inteligibilidade resultante.

3.4.2 Variáveis de caracterização com a Empresa X

A caracterização dos clientes com a Empresa foi relacionada através de variáveis que expressam a individualização do cliente com a empresa. Estão expressas no Quadro 6.

Quadro 6 – Variáveis de caracterização

Variável	Classificação e Descrição
Segmento	Variável Qualitativa Nominal 2 categorias: Supermercado ou não supermercado
Praça	Variável Qualitativa Nominal 6 categorias: Campinas, Fortaleza, Sobral, Juazeiro, Natal e Mossoró
Tempo de cadastros da proposta (cartão)	Variável Quantitativa Continua Medida em meses
Limite Total	Variável Quantitativa Continua Medida em Reais
Possui outro cartão	Variável Qualitativa Nominal 2 categorias: Sim ou Não
Possui adicional	Variável Qualitativa Nominal 2 categorias: Sim ou Não
Tempo para ativação do cartão	Variável Quantitativa Continua Medida em meses

Fonte: elaborada pela autora (2014)

A variável “Segmento” inicialmente não era binária. Optou-se formar um grupo onde todos os produtos que não estão contidos na categoria “Não Supermercado” foram agrupados em uma só classe. E os demais foram classificados como “Supermercado”.

Criaram-se ainda 6 variáveis referentes ao histórico do cliente na empresa. A criação destes indicadores traduz, na perspectiva do Marketing, o que seria o nível direto de utilização do cartão e a relação direta do comportamento desse cliente na empresa.

Todas elas classificam-se como quantitativas, sendo expressas no Quadro 7, e são percentuais.

Quadro 7– Variáveis do histórico do cliente na empresa

Variável	Descrição
Comprometimento do saldo	Relação direta das variáveis: Limite Utilizado/Limite Total do cliente
Quantidade de extratos/Tempo ativado	Relação direta das variáveis: Quantidade de extratos/Tempo ativado
Quantidade de compras no último mês/ Quantidade de compras nos últimos 3 meses	Relação direta das variáveis: Quantidade de compras no último mês/ Quantidade de compras nos últimos 3 meses
Quantidade de compras no último mês/ Quantidade de compras nos últimos 6 meses	Relação direta das variáveis: Quantidade de compras no último mês/ Quantidade de compras nos últimos 6 meses
Quantidade de compras nos últimos 3 meses/ Quantidade de compras nos últimos 6 meses	Relação direta das variáveis: Quantidade de compras no último mês/ Quantidade de compras nos últimos 6 meses
Limite/Renda	Relação direta das variáveis: Limite Total/Renda do cliente

Fonte: Elaborada pela autora (2014)

3.5 Análise dos Dados

A análise dos dados foi realizada com recurso de um *software* de tratamento estatístico: o SPSS - *Statistical Package for Social Science*, versão 21.

Começou-se por caracterizar a amostra através de estatística descritiva, tendo-se construído tabelas de frequências para as variáveis qualitativas e determinado medidas de localização e de dispersão para as quantitativas.

A existência de associação entre a variável dependente (bom ou mau cliente) e cada uma das variáveis independentes consideradas posteriormente nos modelos de

classificação, foi avaliada recorrendo a testes do qui-quadrado, e a testes t, para amostras independentes, consoante a variável independente fosse qualitativa ou quantitativa.

Com o intuito de classificar foram propostos dois modelos, ambos desenvolvidos no *software* SPSS, cuja qualificação do ajuste foi comparada através dos índices apresentados nas seções mais a frente. Foi calculada a influência das variáveis sócio demográficas, de relação com a empresa e das variáveis de histórico do cliente na empresa na variável resposta que pode ser caracterizada por bom e mau cliente.

Sobre a questão de dados ausentes ou inexistentes, optou-se pela exclusão desses clientes da base de dados, por se tratarem de informações com pouca representatividade e sua exclusão não resultam em alterações significativas nas distribuições das respostas.

Após tratamento dos dados, a base considerada contou com 363.172 observações. Foram excluídos todos os dados inexistentes e com erro de digitação. Os dados considerados com erro não puderam receber tratamento pelo fato de já irem para o banco de dados da Empresa X preenchidos por terceiros, impossibilitando a sua correção.

A análise multivariada dos fatores determinantes a avaliação dos clientes em bons e maus foi realizada através de Árvores de Classificação e Regressão (algoritmo CART) e por meio da Regressão Logística.

Por fim, comparou-se as eficiências dos modelos através de 3 métodos distintos. Confrontou-se inicialmente os modelos através dos resultados das medidas calculadas através das tabelas de classificação, sendo estes os métodos de acurácia, *recall*, especificidade, precisão e erro. Posteriormente, utilizou-se do método comparativo da área abaixo da curva de ROC e por fim, recorreu-se ao método do cálculo do indicador de *Kolmogorov- Smirnov* (KS).

4 APRESENTAÇÃO DOS RESULTADOS

Esta seção apresenta os resultados obtidos após a adoção dos procedimentos descritos na seção 3. Inicialmente é exibida a análise descritiva dos dados, posteriormente são expostos os resultados obtidos para cada modelo proposto e posteriormente, comentados e avaliação dos ajustamentos. Por fim, é feita uma comparação do desempenho de cada método a fim de apontar aquele com melhor desempenho.

4.1 Análise descritiva das variáveis sócio demográficas

As Tabelas 1 e 2 ilustram as distribuições das variáveis sócio demográficas: Sexo, Estado Civil, Idade, Natureza da Ocupação, Grau de Instrução, Dependentes, Renda Comprovada e Tempo Empregado.

Para uma melhor apresentação das distribuições, as variáveis qualitativas estão apresentadas na Tabela 1, e as variáveis quantitativas apresentadas na Tabela 2.

Tabela 1 – Frequências das Variáveis Sócio demográficas – qualitativas

Variáveis e respectivas categorias	Número de observações	Porcentagem das observações
Grau de Instrução	Frequência	Porcentual
Até 1º Grau Completo	83.164	22,90%
Até 2º Grau Completo	216.905	59,73%
Outros	27.060	7,45%
Superior Completo	22.983	6,33%
Superior Incompleto	13.060	3,60%
Total	363.172	100,00%
Estado Civil	Frequência	Porcentual
Casado	179.883	49,53%
Solteiro	183.289	50,47%
Total	363.172	100,00%
Sexo	Frequência	Porcentual
F	262.003	72,14%
M	101.169	27,86%
Total	363.172	100,00%
Natureza Ocupação	Frequência	Porcentual
Autônomo	161.524	44,48%
Empregado	201.648	55,52%
Total	363.172	100,00%
Se tem dependentes	Frequência	Porcentual
Sim	259.804	71,54%
Não	103.368	28,46%
Total	363.172	100,00%

Fonte: Elaborada pela autora (2014)

A variável ‘Grau de Instrução’ apresentou maior representatividade na classe ‘Até 2º Grau Completo’, com aproximadamente 60%. Logo seguida pela classe ‘Até 1º Grau completo’, com aproximadamente, 23%. A variável ‘estado civil’ está bem divida, cerca de metade da amostra é casada, e a outra solteira.

O sexo predominante da amostra é o feminino, com 72,14%. Divisão semelhante à da variável ‘Se tem dependentes’, onde a resposta ‘sim’ representa 71,54% do total. Há uma divisão pouco dominante na variável ‘Natureza Ocupação’, 55,50% da amostra pertence à classe “empregado”.

Tabela 2 – Medidas descritivas das variáveis sócio demográficas quantitativas

Variáveis	Mínimo	Máximo	Média	Desvio Padrão
Idade	16	90	39,64	13,81
Renda Comprovada	100	26.000,00	949,20	673,45
Tempo de Emprego	-5,78	72,06	7,22	7,30

FONTE: Elaborada pela autora (2014)

Na variável ‘tempo empregado’, há uma justificativa para o número negativo, pois a empresa X realiza atualizações das informações da sua base. O cálculo do tempo empregado, foi calculada pela diferença entre a data cadastramento em relação a data de admissão, data que pode ter sido alterada posteriormente à data de cadastramento.

Percebe-se que a variável ‘renda comprovada’ apresenta um mínimo de R$ 100,00, pois este é o menor valor disponível para limite da Empresa X. O valor mínimo da variável ‘idade’ é 16 anos, pois essa é a idade mínima para a aprovação do cartão.

4.2 Análise descritiva das Variáveis de caracterização com a empresa

As Tabelas 3 e 4 ilustram a distribuição das variáveis de relacionamento com a empresa: segmento, praça, possui outro cartão, possui adicional, meses de cadastro, limite total e meses para ativação. Para uma melhor apresentação das distribuições, as variáveis qualitativas estão apresentadas na Tabela 3, e as variáveis quantitativas apresentadas na Tabela 4.

Tabela 4– Frequências das variáveis relacionadas à empresa – qualitativas

Variáveis e respectivas categorias	Número de observações	Porcentagem das observações
Segmento	Frequência	Porcentual
NÃO SUPERMERCADO	237.259	65,33%
SUPERMERCADOS	125.913	34,67%
Total	363.172	100,00%
Praça	Frequência	Porcentual
CAMPINA	18	0,00%
FORTALEZA	252.325	69,48%
JUAZEIRO	22.492	6,19%
MOSSORÓ	31.214	8,59%
NATAL	35.037	9,65%
SOBRAL	22.086	6,08%
Total	363.172	100,00%
Possui Outro Cartao	Frequência	Porcentual
NÃO	107.846	29,70%
SIM	255.326	70,30%
Total	363.172	100,00%
Possui adicional	Frequência	Porcentual
NÃO	327.523	90,18%
SIM	35.649	9,82%
Total	363.172	100,00%

Fonte: Elaborada pela autora (2014)

Na variável segmento é justificável a grande concentração na classe 'não supermercado', 65,3%, pois essa classe é um agrupamento de os demais segmentos, diferentes da classe 'supermercado'.

Há, também, uma concentração na variável 'Praça', na cidade de Fortaleza, por esta se tratar da sede da empresa X. Os clientes da empresa X não apresentam na sua grande maioria um dependente no seu cartão, 9,82% possui sim um dependente, porém esse perfil não percebe na variável 'possui outro cartão', já que sua concentração maior é na resposta positiva, 70,30%.

Tabela 4 – Medidas descritivas das variáveis quantitativas relacionadas à empresa

	N	Mínimo	Máximo	Média	Desvio padrão	Variância
Meses de cadastro	363.172	3,03	92,57	23,57	16,28	264,77
Limite Total	363.172	10,00	10.000,00	566,13	481,84	232.174,64
Meses para ativação	363.172	0,00	77,50	1,20	3,56	12,68

Fonte: Elaborada pela autora (2014)

Sobre o tempo de cadastro da proposta do cartão, a média é de 23,5 meses e há uma grande variedade no limite desses clientes. O tempo médio de ativação do cliente da empresa X é em torno de 1 mês após seu cadastro e vale ressaltar que a diferença entre estas

duas variáveis é quando o cliente registra a proposta e quando ele utiliza a primeira vez o seu cartão, para o tempo de cadastro de proposta e meses para a ativação, respectivamente.

4.1.3 Análise descritiva das variáveis do histórico do cliente na empresa

A Tabela 5 ilustra a distribuição das variáveis resultantes: comprometimento do saldo, quantidade de extratos/tempo ativado; quantidade de compras no último mês/ quantidade de compras nos últimos 3 meses; quantidade de compras no último mês/ quantidade de compras nos últimos 6 meses, quantidade de compras nos últimos 3 meses/ quantidade de compras nos últimos 6 meses e limite sobre renda.

Tabela 5– Medidas descritivas das variáveis de histórico

Variável	Mínimo	Máximo	Média	Desvio Padrão
Comprometimento do saldo	-97,00%	103,00%	63,66%	35,19%
Quantidade de extratos/Tempo ativado	1,00%	100,00%	56,45%	24,93%
Quantidade de compras no último mês/ Quantidade de compras nos últimos 3 meses	2,00%	100,00%	52,09%	27,88%
Quantidade de compras no último mês/ Quantidade de compras nos últimos 6 meses	1,00%	100,00%	35,37%	26,99%
Quantidade de compras nos últimos 3 meses/ Quantidade de compras nos últimos 6 meses	3,00%	100,00%	65,46%	23,43%
Limite/Renda	1,00%	500,00%	64,47%	47,81%

FONTE: elaborada pela autora (2014)

A variável que mede a quantidade de extratos em relação ao tempo ativado varia de 0% a 100%. Isso mostra que há clientes na amostra que nunca utilizaram o cartão, e que sempre utilizaram, ou seja, emitiram fatura em todos os meses possíveis.

Percebe-se através da variável resultado da relação da quantidade de compras do último mês com quantidade de compras nos últimos três meses que há clientes que compraram em um período maior que um mês, chegando até em somente 2% no último mês, como há clientes que compraram 100% de suas compras totais no último mês. Resultado semelhante à relação quantidade de compras último mês com quantidade de compras dos últimos seis meses, aonde o percentual de compras do último mês chegou a 1%.

A relação limite com renda, apresentou uma média de 64%. Com mínimo de 1%, e máximo de até 500%. Por mais que a empresa X tenha uma política do limite total não ultrapassar a renda do cliente, caso esse no decorrer do tempo apresente bom histórico, é possível essa situação.

A variável "comprometimento do saldo", que é a relação do limite disponível com o limite total, é possível apresentar valor superior a 100% devido ao *over limit*[1], porém esse limite não está disponível em todo mês, e depende de uma serie de fatores do histórico do cliente.

4.2 Estatística Inferencial

Uma vez que se pretende estudar a associação das variáveis de caracterização sócio-demográfica, de relação com a empresa e de histórico, com o fato do cliente ser bom ou mau (variável dependente), tem-se o interesse aplicar testes de hipóteses que possam avaliar cada uma destas associações. Para as variáveis qualitativas, realizou-se o teste qui-quadrado, cuja hipótese nula é de que não há associação entre as variáveis, registrando-se associações significativas para valores de p inferiores ou iguais ao nível de significância fixado em 5%. Conforme mostra a tabela 6, todos os valores *p* são inferiores a 0,0001, evidenciando associações significativas entre todas as variáveis de caracterização do cliente e o seu comportamento (bom ou mau).

Tabela 6- Teste Qui-Quadrado

Variável	**Valor**	**Graus de liberdade**	**Valor *p***
Segmento	1.756,49	1	0,000
Praça	1.221,83	5	0,000
Sexo	90,86	1	0,000
Natureza Ocupação	90,01	1	0,000
Grau de instrução	75,69	4	0,000
Possui outro cartão	10,28	1	0,000
Possui adicional	1.269	1	0,000
Estado civil	25,96	1	0,000
Se tem dependentes	309,05	1	0,000

Fonte: Elaborada pela autora (2014)

Construiu-se ainda para variáveis qualitativas uma tabela de contingência, para melhor análise das relações das classes das respectivas variáveis com a resposta do critério. A Tabela 7 traz o resultado destas associações.

[1] *Over Limit:* Percentual liberado para a realização de compras do cliente além do limite disponível normal. Varia por empresa, e do histórico de utilização de cada cliente.

Tabela 7 – Tabela de contingência

Variável		Critério	
		% Bom	% Mau
Segmento	Não Supermercado	47,5%	52,5%
	Supermercado	54,8%	45,2%
Praça	Campina	66,7%	33,3%
	Fortaleza	50,7%	49,3%
	Juazeiro	39,9%	60,1%
	Mossoró	48,0%	52,0%
	Natal	53,8%	46,2%
	Sobral	49,4%	50,6%
Sexo	Feminino	49,5%	50,5%
	Masculino	51,3%	48,7%
Ocupação	Autônomo	49,1%	50,9%
	Empregados	50,7%	49,3%
Grau de Instrução	Até 1º Grau Completo	51,1%	48,9%
	Até 2º Grau Completo	49,7%	50,3%
	Outros	49,4%	50,6%
	Superior Completo	50,7%	49,3%
	Superior Incompleto	48,0%	52,0%
Possui outro carrão	Não	49,6%	50,4%
	Sim	50,2%	49,8%
Possui adicional	Não	49,0%	51,0%
	Sim	59,0%	41,0%
Estado Civil	Casado	49,6%	50,4%
	Solteiro	50,4%	49,6%
Se tem dependentes	Sim	49,1%	50,9%
	Não	52,3%	47,7%

Fonte: Elaborada pela autora (2014)

Ao analisar a variável “Segmento” percebe-se que a classe não supermercado apresenta maior percentual de “Mau”, diferentemente da classe supermercado, que apresenta maior representatividade de bons registros.

Na variável “Praça” os registros de Campina são os que apresentam maior diferença, onde apenas 33% estão concentrados na classe “mau”, logo seguida por Juazeiro, mas que de forma oposta, apresentou 60,1% dos seus registros concentrados na classe “mau”. Fortaleza, Mossoró, Natal e Sobral apresentaram divisão semelhante, onde Fortaleza e Natal apresentam pequena representação maior nos registros bons, diferentemente das praças de Mossoró e Sobral.

O sexo feminino apresentou maior concentração dos registros classificados como “maus”, de forma oposta, o sexo masculino apresentou maior classificação como “bons”.

A variável ocupação apresentou na classe “autônomos” maior representatividade de maus registros, enquanto a classe “empregados” apresentou maioria de bons.

A variável grau de instrução apresentou nas classes “Até 1º grau completo” e “superior completo” maior representatividade de bons registros, enquanto as classes de “Até 2º grau completo”, “outros” e “Superior incompleto” apresentaram maioria de maus registros.

Para quem possui outro cartão o percentual maior é de bons registros, enquanto quem não possui outro cartão a maioria é de maus registros. Divisão semelhante para quem possui adicional, que sua maioridade está no registro de bons, enquanto quem não possui adicional apresenta maioria de maus registros.

Para quem é solteiro a maioria dos registros se encontra classificado como “bom”, enquanto para quem está na classe casado a maioria dos registros é de maus.

Caso o registro possua dependente, este tem maior chance de ser classificado como “mau”, de acordo com o critério e caso não possua dependente, este tem maior chance de ser classificado como “bom”.

Para as variáveis quantitativas, realizou-se o teste t para amostras independentes. Para cada variável independente, a hipótese nula é de que a respetiva média é a mesma nos grupos dos bons e no dos maus clientes. Em que se aceita a hipótese nula, com 5% de significância. Conforme mostra a tabela 8, rejeita-se a hipótese nula, e verifica-se que existem diferenças significativas entre as médias das variáveis independentes nos dois grupos definidos pela variável dependente.

Tabela 8 – Teste t para amostras independentes (*continua*)

Variável	Critério	Média	Desvio padrão	valor *p*
Tempo de cadastro	Mau	22,42	15,28	0,00
	Bom	24,72	17,12	
Idade	Mau	39,11	13,77	0,00
	Bom	40,17	13,83	
Tempo de emprego	Mau	7,18	7,37	0,00
	Bom	7,27	7,39	
Limite total	Mau	516,61	456,03	0,00
	Bom	615,66	501,49	
Renda comprovada	Mau	976,34	812,39	0,00
	Bom	922,05	465,54	
Quantidade de extratos/tempo ativado	Mau	0,53	0,25	0,00
	Bom	0,60	0,24	

Tabela 8 – Teste t para amostras independentes (*continuação*)

Variável	Critério	Média	Desvio padrão	valor *p*
Comprometimento do saldo	Mau	0,63	0,38	0,00
	Bom	0,65	0,33	
Quantidade de compras último mês/quantidade de comoras últimos 3 meses	Mau	0,58	0,30	0,00
	Bom	0,46	0,25	
Quantidade de compras último mês/quantidade de comoras últimos 6 meses	Mau	0,41	0,29	0,00
	Bom	0,30	0,24	
Quantidade de compras últimos 3 meses/quantidade de comoras últimos 6 meses	Mau	0,68	0,25	0,00
	Bom	0,63	0,22	
Tempo para ativar	Mau	1,28	3,78	0,00
	Bom	1,12	3,32	
Limite/Renda	Mau	0,58	0,39	0,00
	Bom	0,71	0,55	

Fonte: Elaborada pela autora (2014)

4.3 Árvores de Classificação

O primeiro modelo aplicado no estudo foi o de árvores de classificação e regressão. Foram inclusas as variáveis caracterizadas como sócio demográficas, de caracterização com a empresa e as relativas ao histórico na empresa.

Conforme metodologia descrita na seção 3.5, a base, para a construção do classificador baseado em árvore, foi particionada em duas, onde 70% da amostra foi selecionada para construção do modelo, caracterizando 254.612 registros, e os 30% restantes da amostra, constituíram o grupo de validação, caracterizando 108.560 registros. O critério de divisão dos nós foi o de *Gini*, restringindo a profundidade das divisões a 5 níveis .

As árvores geradas com a amostra de modelação e amostra de validação apresentaram 13 nós terminais, e somente a variável sexo não foi incluída na geração das árvores resultantes. Dos 13 nós terminais, 5 nós apresentam maioria de registros classificados como bons, e 8 apresentam maioria de registros classificados como maus.

As figuras 8 e 9 mostram as árvores resultado dos modelos de modelação e validação, respectivamente.

Observa-se dentro dos retângulos de ambas as figuras a porcentagem de clientes que são classificados como maus e como bons. Abaixo dos nós ficam as variáveis selecionadas para a divisão até que chegue às folhas.

A árvore de modelação, Figura 8, seguiu a seguinte divisão de características:

- As 254.612 observações do nó raiz foram divididos segundo a Quantidade de compras no último mês em relação à quantidade de compras nos últimos 3 meses, em grupos de 204.559 e 50.053 (nó 1 e 2, respectivamente).

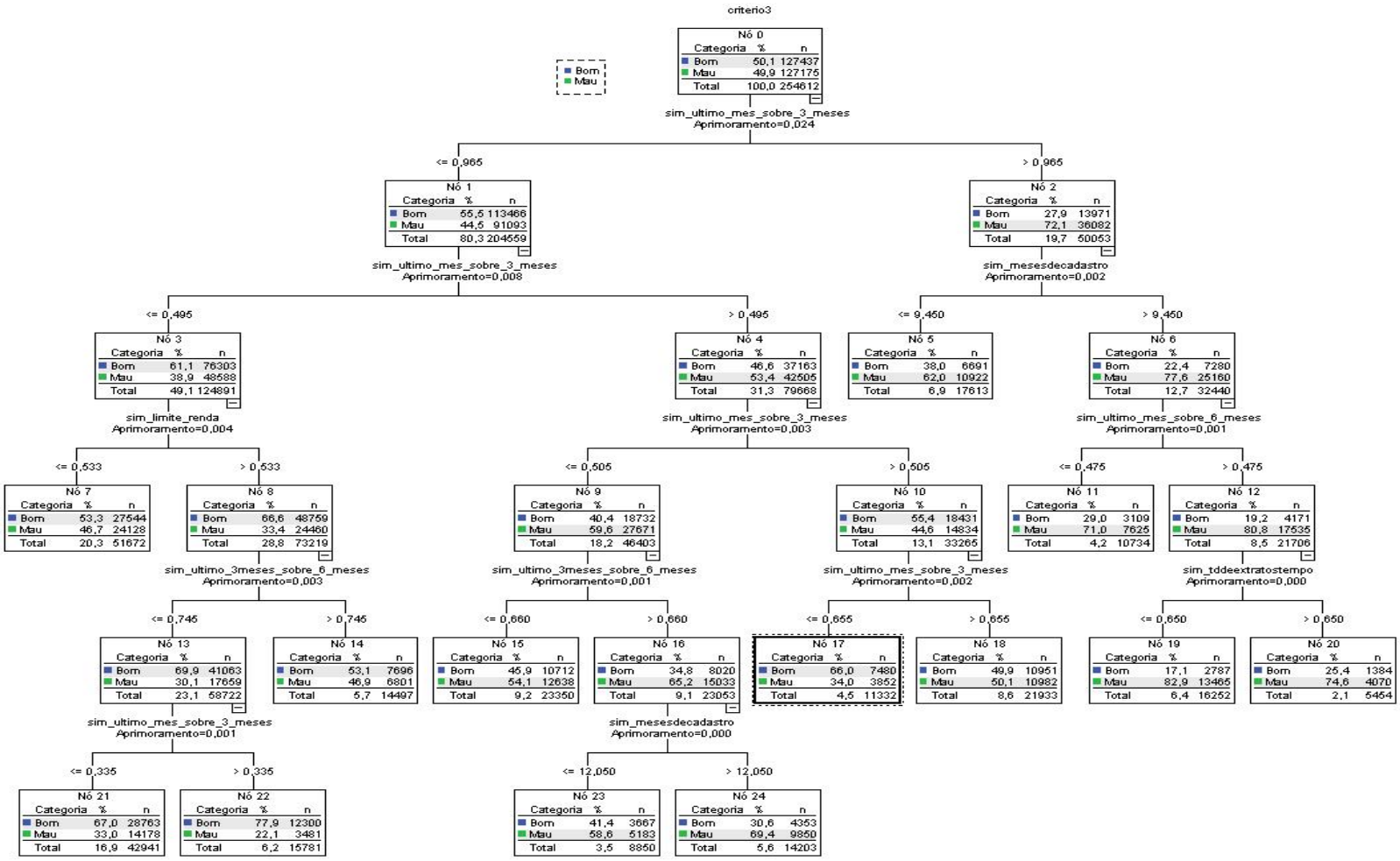
- As informações do nó 3 foram divididas conforme a característica Limite/Renda. Onde os registros foram separados de acordo com o percentual maior ou não que 53,3% desta relação. Os registros que apresentaram valores superiores, formaram o nó 8 (32.140 registros) e os registros que apresentaram valores inferiores, formaram o nó 7 (21.811 registros).

- O nó 8, ramificou-se ainda nos nós 13 e 14, consoante a quantidade de compras nos últimos 3 meses em relação a quantidade de compras nos últimos 6 meses. Aqueles cujo percentual foi inferior a 74,5% formaram o nó 13, que, por sua vez, ramificou-se nos nós terminais 21 e 22.

- Os nós 9 e 10, são folhas do nó 4. Resultam em 1 e 2 nós terminais, respectivamente, após outra ramificação.

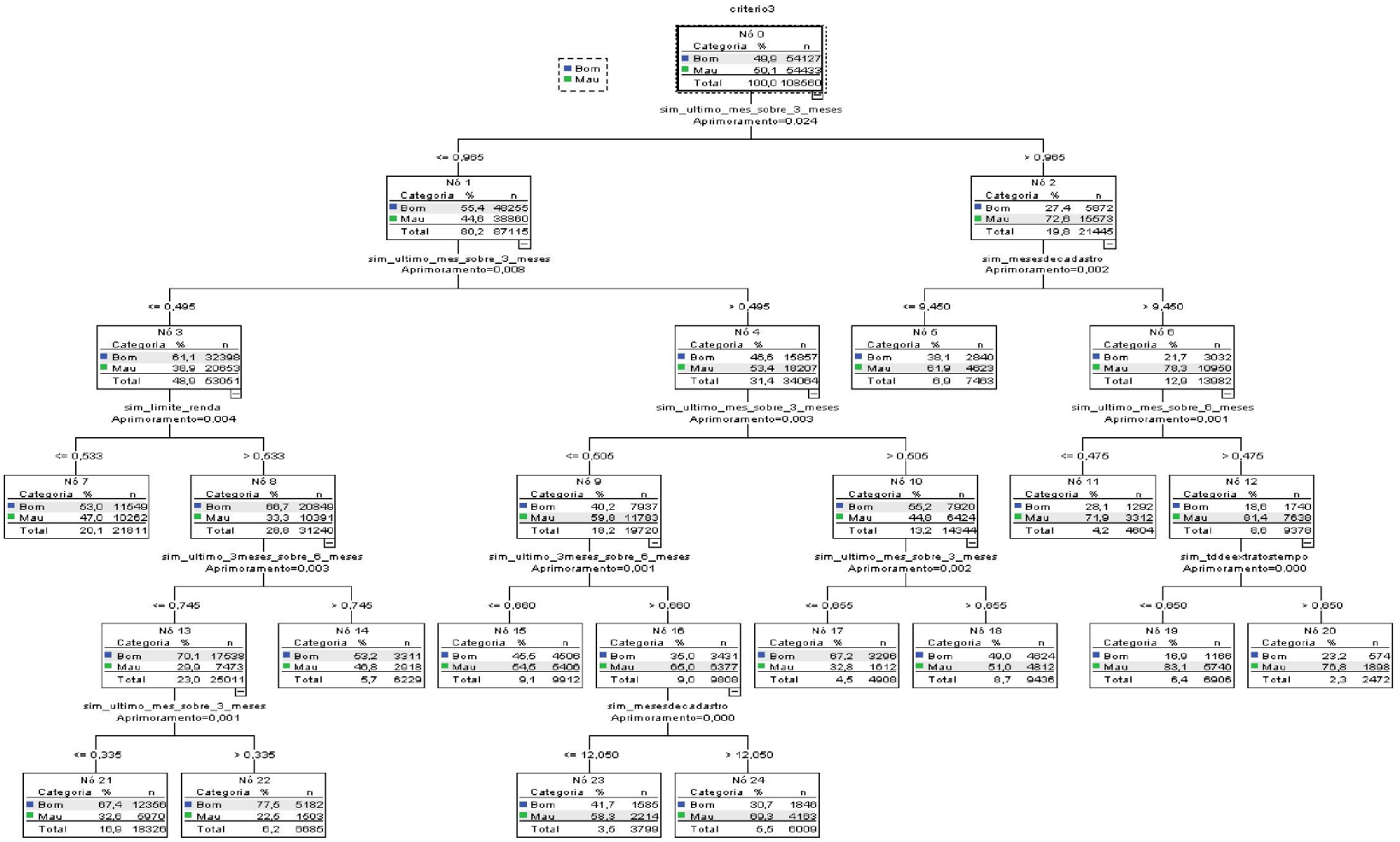
- O nó 6 ramifica-se até resultar em 3 nós terminais (5, 19 e 20), onde em duas destas folhas estão concentradas os maiores percentuais de maus clientes.

Figura 8 – Árvore de Classificação - Amostra de Modelação



Fonte: Elaborada pela autora (2014)

Figura 9 – Árvore de Classificação – Amostra de Validação



Fonte: Elaborada pela autora (2014)

Na Tabela 9, mostra-se o resumo das folhas da árvore de classificação, a quantidade de clientes em cada nó terminal, e o percentual de clientes na classificação bom e mau, para a amostra de treinamento e validação.

Tabela 9 - Resumo das folhas – Árvore *Gini:* modelação e validação

Nó terminal	Quantidade Modelação	Quantidade Validação	Prob (1) - Modelação	Prob (1) - Validação	Classificação Prevista
19	16.252	6.906	0,829	0,831	Mau
20	5.454	2.472	0,746	0,768	Mau
11	10.734	4.604	0,71	0,719	Mau
24	14.203	6.009	0,694	0,693	Mau
5	17.613	7.463	0,62	0,619	Mau
23	8.850	3.799	0,586	0,583	Mau
15	23.350	9.912	0,541	0,545	Mau
18	21.933	9.436	0,501	0,51	Mau
14	14.497	6.229	0,469	0,468	Bom
7	51.672	21.811	0,467	0,47	Bom
17	11.332	4.908	0,34	0,328	Bom
21	42.941	18.326	0,33	0,326	Bom
22	15.781	6.685	0,221	0,225	Bom

Fonte: Elaborada pela autora (2014)

Comparando-se as probabilidades previstas da Tabela 8, entre a amostra de modelação e amostra de validação percebem-se diferenças mínimas nos percentuais. As menores variações são as dos nós 14, 24 e 5, com variação absoluta de 0,001 cada um. A maior variação é encontrada no nó 17, com diferença de 0,012 entre o percentual encontrado na amostra de modelação e amostra de validação.

Pode se encontrar no nó 19 uma alta concentração de “maus” clientes, com 82,9% de concentração na amostra de modelação e 83,1% na amostra de validação. De forma oposta, no nó 22 pode se encontrar uma concentração de 77,9% e 77,5% de bons clientes, na amostra de modelação e validação, respectivamente. No nó 18, porém, há uma divisão estatisticamente igual na representatividade de cada grupo, 50,1% e 51,0%, na amostra de modelação e validação, respectivamente. Porém, ambos classificados como um nó de maus clientes.

Para cada nó, tem-se a apresentação de características mostradas no Quadro 8.

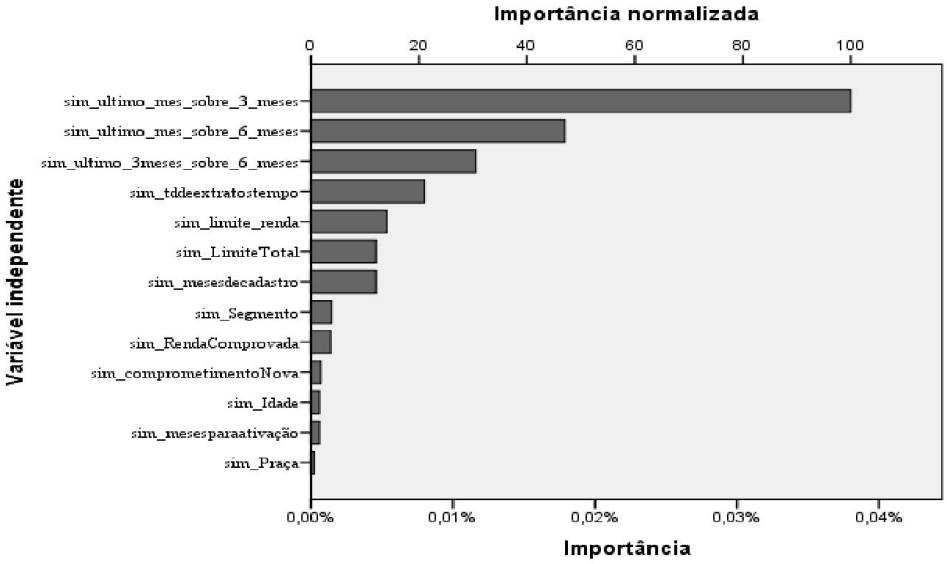
Quadro 8 - Descrição dos nós terminais

Nó	Descrição
19	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: >0,965 Meses de Cadastro: >9,450 Quantidade de compras último mês/Quantidade de Compras últimos 6 meses: >0,475 Quantidade de extratos/ Tempo Ativado: <= 0,65
20	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: >0,965 Meses de Cadastro: >9,450 Quantidade de compras último mês/Quantidade de Compras últimos 6 meses: >0,475 Quantidade de extratos/ Tempo Ativado: > 0,65
11	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: >0,965 Meses de Cadastro: >9,450 Quantidade de compras último mês/Quantidade de Compras últimos 6 meses: <=0,475
24	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: >0,965 Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: >0,495 e <=0,505 Quantidade de compras últimos 3 meses/Quantidade de Compras últimos 6 meses: >0,66 Tempo de cadastro: >12,05
5	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: >0,965 Meses de Cadastro: <=9,45
23	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: >0,965 Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: > 0,495 e <=0,505 Quantidade de compras últimos 3 meses/Quantidade de Compras últimos 6 meses: >0,66 Tempo de cadastro: <=12,05
18	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: <=0,965 Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: >0,655
14	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: <=0,495 Limite Total/ Renda Comprovada: >0,533 Quantidade de compras últimos 3 meses/Quantidade de Compras últimos 6 meses: >0,746
7	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: <=0,495 Limite Total/ Renda Comprovada: <=0,533
17	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: <=0,965 Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: >0,655 e <=0,655
21	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses:<=0,495 Limite Total/ Renda Comprovada: >0,533 Quantidade de compras últimos 3 meses/Quantidade de Compras últimos 6 meses: <=0,746 Quantidade de compras último mês/Quantidade de Compras últimos 3 meses:<=0,335
22	Quantidade de compras último mês/Quantidade de Compras últimos 3 meses:<=0,495 Limite Total/ Renda Comprovada: >0,533 Quantidade de compras últimos 3 meses/Quantidade de Compras últimos 6 meses:<=0,746 Quantidade de compras último mês/Quantidade de Compras últimos 3 meses: >0,335

Fonte: Elaborada pela autora (2014)

O Gráfico 1 mostra a ordem de importância das variáveis selecionadas para a construção do modelo.

Gráfico 1 – Importância das variáveis independentes no modelo



Fonte: Elaborada pela autora (2014)

A variável de maior importância para o CART, de acordo com o Gráfico 1, é a variável "Quantidade de compras último mês/Quantidade e compras últimos 3 meses", logo seguida por "Quantidade de compras último mês/Quantidade e compras últimos 6 meses", em último grau de importância para o modelo está a variável "Praça", estas classificações de importância demonstram que a informação se os clientes que utilizaram mais ou menos o seu cartão no último mês é mais valiosa para o modelo comparada à variável que informa a praça do cliente.

Uma das formas utilizadas na avaliação da classificação do modelo é a tabela de classificação, que está apresentada na Tabela 10.

Tabela 10– Tabela de classificação – CART

Posto				
Amostra		Previsto		
		Bom	Mau	Porcentagem Correta
Treinamento	Bom	83.783	43.654	65,7%
	Mau	52.440	74.735	58,8%
	Porcentagem global	53,5%	46,5%	62,3%
Teste	Bom	35.694	18.433	65,9%
	Mau	22.265	32.168	59,1%
	Porcentagem global	53,4%	46,6%	62,5%

Fonte: Elaborada pela autora (2014)

Com os resultados da tabela de classificação é possível encontrar as medidas de análise de ajuste acurácia, erro, *recall*, especificidade e ajuste, para os modelos de treinamento e validação, conforme mostra Tabela 11.

Tabela 11 - Medidas de ajustamento do modelo

Medidas	Modelo	Validação
Acurácia	62,3%	62,5%
Erro	37,7%	37,5%
Recall	65,7%	65,9%
Especificidade	58,8%	59,1%
Precisão	61,5%	61,6%

Fonte: Elaborada pela autora (2014)

Como mostra a Tabela 9, para a amostra de treinamento, dentre os 136.223 (83.783 + 52.440) registros que foram classificados como bons, 83.783 foram classificados corretamente (53,5%), e os 118.389 (74.735+ 43.654) registros eu foram classificados como maus, 74.735 (46,5%) foram classificados corretamente. Simultaneamente, dos 137.547 (83.783 + 43.654) registros que atenderem o mínimo para a classificação como bom pelo critério, somente 83.783 foram classificados corretamente (65,7%), e dentre os 127.175(52.440 + 74.735) registros que não atenderam ao critério, e foram classificados como maus somente 74.735 foram classificados corretamente (58,8%).

O modelo classificatório da base de teste levou em consideração o número de corte de 0,5. Pode se avaliar o resultado de outros pontos de corte calculando novas especificidades e recall para cada corte, e encontrando novas eficiências. Conforme mostra Tabela 12 o ponto de corte de maior equilíbrio é o 0,5.

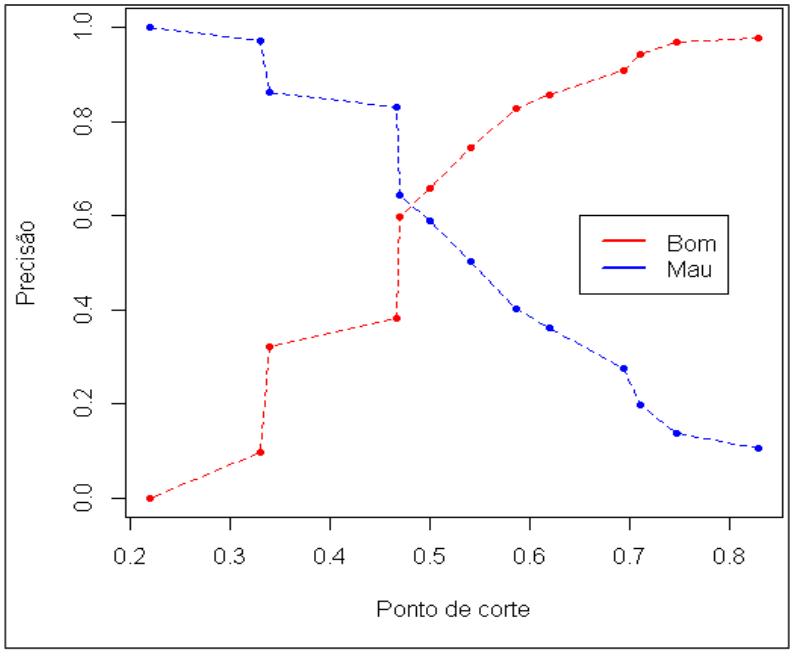
A curva ROC confirmou a escolha do ponto de corte, apresentando também o ponto de corte com equilíbrio maior em 0,5, como mostra o Gráfico 2.

Tabela 12 – Eficiência dos cortes – árvore de classificação

Corte	Especificidade	Sensitividade	Eficiência
0	50,0%	0,0%	25,0%
0,1	50,0%	0,0%	25,0%
0,2	50,0%	0,0%	25,0%
0,3	51,8%	77,8%	65,0%
0,4	57,4%	69,4%	63,0%
0,5	66,2%	59,9%	63,0%
0,6	71,6%	57,0%	64,0%
0,7	77,8%	54,0%	66,0%
0,8	82,9%	52,2%	68,0%
0,9	0,0%	50,0%	25,0%
1	0,0%	50,0%	25,0%

Fonte: Elaborada pela autora (2014)

Gráfico 2 – Ponto de corte na curva ROC para árvore de classificação



Fonte: Elaborada pela autora (2014)

Como as probabilidades são limitadas à quantidade de nós e folhas, o gráfico não é contínuo. Então, ao invés de ser escolhido o ponto onde as curvas se encontram, escolhe-se o ponto de corte onde a precisão dos dois grupos mais se aproxima.

4.4 Regressão Logística

Em regressão logística não existe método único, com o objetivo de encontrar o modelo que mais se ajuste, foi construída uma tabela onde vários pontos de corte foram testados, variando entre 0.1 e 0.9, para encontrar o ponto de maior eficiência do modelo.

Uma maneira de escolha do ponto de corte é calcular a eficiência, ou seja, a média aritmética do recall (sensitividade) e especificidade. O melhor ponto teria maior eficiência. Este ponto seria o que melhor separaria os maus clientes dos bons clientes.

Porém, antes de definir pelo ponto de corte, há a situação de ponderar qual o erro é o mais tolerável, se os falsos bons clientes ou os falsos maus clientes. No caso da empresa X, onde o objetivo é a identificação de maus clientes para uma ação de marketing influenciando o uso do cartão, há uma propensão para a classificação de bons clientes como maus à classificação de maus clientes como bons.

Contudo, o critério deste trabalho foi o ponto de corte de melhor equilíbrio, e este demonstrou ser o ponto 0,5, conforme tabela 13.

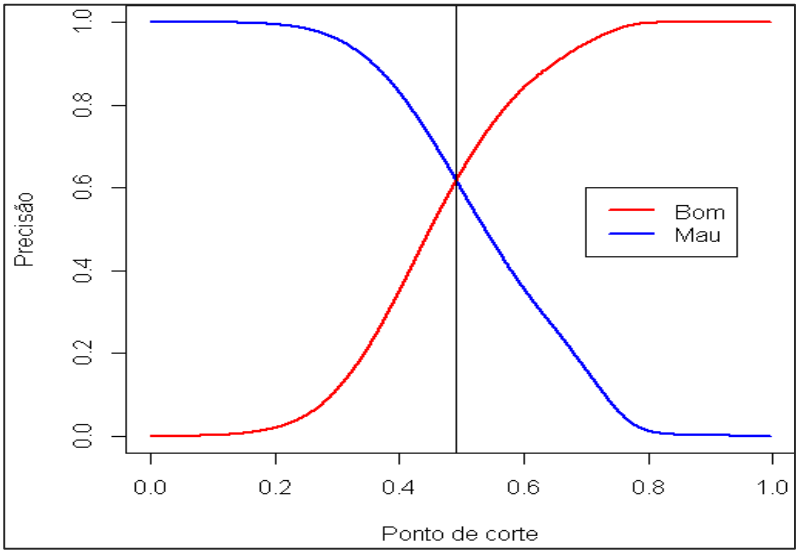
A curva ROC também apresentou um ponto de corte com equilíbrio maior em 0,5, como mostra o Gráfico 3.

Tabela 13 - Eficiência dos cortes – regressão logística

Corte	Especificidade	Sensitividade	Eficiência
0,1	100,0%	0,1%	50,1%
0,2	99,5%	1,9%	50,7%
0,3	96,0%	11,4%	53,7%
0,4	83,1%	35,1%	59,1%
0,5	59,7%	64,0%	61,9%
0,6	36,0%	84,0%	60,0%
0,7	35,0%	85,0%	59,6%
0,8	1,0%	99,9%	50,6%
0,9	0,0%	100,0%	50,0%

Fonte: Elaborada pela autora (2014)

Gráfico 3 – Ponto de corte na curva ROC para regressão logística



Fonte: Elaborada pela autora (2014)

Para construção do modelo de regressão logística utilizou-se o método *forward stepwise* com 0,05 de significância para entrada e saída de variáveis. Esse método permite a correção da multicolinearidade, pois desconsidera variáveis que apresentam esse problema, optando por deixar no modelo as de maior significância.

A variável dependente pode ser influenciada pela presença de qualquer uma das variáveis quantitativa e qualitativas, as variáveis quantitativas podem ser transformadas em outra escala, porém as variáveis qualitativas necessitam de um método para tratamento de seus atributos. A solução é a criação de variáveis artificiais, que assumem valores 0 e 1, conhecidas como variáveis *dummy*.

A tabela 14 traz as variáveis *dummy* criadas para o modelo.

Tabela 14 – Variáveis categóricas e suas respectivas variáveis *dummy*

Variáveis		Frequência	Codificação de parâmetro				
			(1)	(2)	(3)	(4)	(5)
Praça	Campina	18	1	0	0	0	0
	Fortaleza	252.325	0	1	0	0	0
	Juazeiro	22.492	0	0	1	0	0
	Mossoró	31.214	0	0	0	1	0
	Natal	35.037	0	0	0	0	1
	Sobral	22.086	0	0	0	0	0
Grau de Instrução	Até 1º Grau completo	83.164	1	0	0		
	Até 2º Grau completo	216.905	0	1	0		
	Outros	27.060	0	0	1		
	Superior (completo + incompleto)	36.043	0	0	0		
Sexo	Feminino	262.003	1				
	Masculino	101.169	0				
Natureza ocupação	Autônomos	161.524	1				
	Empregados	201.648	0				
Se tem dependentes	Sim	259.804	1				
	Não	103.368	0				
Possui outro cartão	Não	107.846	1				
	Sim	255.326	0				
Estado civil	Casado	179.883	1				
	Solteiro	183.289	0				
Possui adicional	Não	327.523	1				
	Sim	35.649	0				
Segmento	Não Supermercado	237.259	1				
	Supermercado	125.913	0				

Fonte: Elaborada pela autora (2014)

O método *forward stepwise* conduziu um modelo com 18 variáveis:

Segmento, Praça, Ocupação, Grau de Instrução, Possui outro cartão, Possui adicional, Estado Civil, Se tem dependentes, Tempo de cadastro, Tempo de emprego, Limite total, Renda comprovada, Extratos/tempo, Comprometimento do saldo, Quantidade de compras ultimo mês/Quantidade de compras últimos 3 meses, Quantidade de compras ultimo mês/Quantidade de compras últimos 6 meses, Quantidade de compras últimos 3 meses/Quantidade de compras últimos 6 meses, Tempo para ativação e Limite/Renda.

Somente as variáveis: idade, sexo e meses de cadastro não entraram no modelo.

O resultado do teste de Qui Quadrado de Quald foi de 31.185,77 com 25 graus de liberdade. *Cox & Snell* R^2 apresentou valor de 0,082. *Nagelkerke* R^2 para este modelo foi de 0,11.

A tabela de classificação, conforme visto nesta mesma seção, apresentou resultado geral de 61,9%. A Tabela 15 mostra o *recall* e especificidade de ambas as classificações possíveis.

Tabela 15 – Tabela de classificação - regressão logística

Observado	Classe Preditiva		
	Bom	Mau	Porcentagem correta
Bom	116.230	65.334	64,0%
Mau	73.266	108.342	59,7%
	Porcentagem Global		61,8%

Fonte: Elaborada pela autora (2014)

A Tabela 16 traz as informações de ajustamento do modelo de regressão logística.

Tabela 16 – Medidas de ajustamento

Medidas	%
Acurácia	61,8%
Erro	38,2%
Recall	64,0%
Especificidade	59,7%
Precisão	61,3%

Fonte: Elaborada pela autora (2014)

Como mostra a Tabela 14, dentre os 189. 496 (116.230 + 73.266) registros que foram classificados como bons, 116.230 foram classificados corretamente (61,3%), e os 173.676 (65.334+ 108.342) registros eu foram classificados como maus, 108.342 (62,4%) foram classificados corretamente. Simultaneamente, dos 181.564 (116.230 + 65.334) registros que atenderem o mínimo para a classificação como bom pelo critério, 116.230 foram classificados corretamente (64,0%), e dentre os 181.608 (73.266 + 108.342) registros que não atenderam ao critério, e foram classificados como maus, 108.342 foram classificados corretamente (59,7%).

A Tabela 17 traz as informações da equação final.

Tabela 17 – Variáveis e coeficientes da equação

							95% C.I. para EXP(B)	
Variáveis	B	S.E.	Wald	df	Sig.	Exp(B)	Inferior	Superior
Segmento (1)	0,261	0,008	1003,286	1,000	0,000	1,298	1,277	1,319
Praça			522,337	5,000	0,000			
Praça (1)	-0,808	0,508	2,528	1,000	0,112	0,446	0,164	1,207
Praça (2)	0,132	0,015	74,109	1,000	0,000	1,141	1,107	1,176
Praça (3)	0,340	0,020	288,169	1,000	0,000	1,404	1,350	1,461
Praça (4)	0,274	0,019	206,057	1,000	0,000	1,315	1,267	1,365
Praça (5)	0,032	0,019	2,733	1,000	0,098	1,032	0,994	1,072
Natureza Ocupação (1)	-0,032	0,007	19,878	1,000	0,000	0,969	0,955	0,982
Grau de Instrução			39,602	3,000	0,000			
Grau de Instrução (1)	0,019	0,014	2,002	1,000	0,157	1,020	0,993	1,047
Grau de Instrução (2)	-0,016	0,012	1,630	1,000	0,202	0,984	0,961	1,008
Grau de Instrução (3)	0,061	0,017	12,275	1,000	0,000	1,062	1,027	1,099
Possui Outro Cartão(1)	-0,042	0,008	27,273	1,000	0,000	0,959	0,944	0,974
Possui adicional (1)	0,273	0,012	522,020	1,000	0,000	1,314	1,284	1,345
Estado Civil (1)	-0,016	0,007	5,280	1,000	0,022	0,984	0,970	0,998
Se tem dependentes (1)	0,024	0,008	8,432	1,000	0,004	1,024	1,008	1,040
Tempo de emprego	0,002	0,001	21,920	1,000	0,000	1,002	1,001	1,003
Limite Total	0,000	0,000	344,762	1,000	0,000	1,000	1,000	1,000
Renda Comprovada	0,000	0,000	485,577	1,000	0,000	1,000	1,000	1,000
Total de extratos/tempo	-0,496	0,021	569,674	1,000	0,000	0,609	0,584	0,634
Comprometimento do saldo	-0,144	0,011	174,604	1,000	0,000	0,866	0,847	0,884
Quantidade de compras ultimo mês/Quantidade de compras últimos 3 meses	2,695	0,037	5236,451	1,000	0,000	14,808	13,766	15,930
Quantidade de compras ultimo mês/Quantidade de compras últimos 6 meses	-1,990	0,052	1485,997	1,000	0,000	0,137	0,124	0,151
Quantidade de compras últimos 3 meses/Quantidade de compras últimos 6 meses	1,446	0,035	1667,627	1,000	0,000	4,248	3,963	4,553
Tempo para ativação	0,007	0,001	43,715	1,000	0,000	1,007	1,005	1,009
Limite/renda	-0,468	0,017	775,420	1,000	0,000	0,626	0,606	0,647
Tempo de cadastro	0,092	0,004	559,034	1,000	0,000	1,097	1,088	1,105
Constante	-1,749	0,039	1984,635	1,000	0,000	0,174		

Fonte: Elaborada pela autora (2014)

Os coeficientes associados a cada variável e as correspondentes razões de chance e intervalo de confiança (95%) permitem um melhor conhecimento das características dos cientes que os tornam mais ou menos propensos a pararem de utilizarem o cartão.

A interpretação dos coeficientes do modelo foi feita através de análise das razões de chance.

Ao analisar as variáveis quantitativas com coeficientes negativos, conforme Tabela 16, verifica-se que para aumento de uma unidade nas variáveis Quantidade de extratos/Tempo, comprometimento do saldo, quantidade de compras último mês/ quantidade de compras últimos 6 meses e limite/renda, a possibilidade de um cliente ser mau diminui 3,9%, 13,4%, 86,3% e 37,4%, respectivamente.

Quanto as variáveis quantitativas com contribuição positiva, observa-se que como o aumento de uma unidade das variáveis meses para ativação e tempo de cadastro (da proposta), a chance de um cliente ser mau aumenta 0,7% e 9,7% respectivamente.

Ao observar as razões de chance associadas às variáveis quantidade de compras último mês/quantidade de compras últimos 6 meses e quantidade de compras últimos 3 meses/quantidade de compras últimos 6 meses, o aumento de uma unidade nessas variáveis, a chance de ser mau fica 15 e 4 vezes maior, respectivamente.

A razão de chance igual a 1 indica que a possibilidade de ser bom ou mau resultado do aumento das variáveis de Tempo de Emprego, Limite Total e Renda comprovada é igual.

Em relação às variáveis categóricas, as razões de chance indicam quantas vezes é mais ou menos provável que um cliente desta categoria seja mau comparado a um cliente que esteja na categoria de referência. Para cada variável, a categoria referencial é a primeira que surge na ordem de codificação.

Ao observar a razão de chance associada à variável Segmento, detecta-se que um cliente que possui o cartão da categoria não supermercado, tem 1,29 mais chances de ser mau cliente em relação a um cliente que possua cartão de supermercado.

Em relação à praça, um cliente de Campinas tem 55% a menos de chances de ser mau cliente, que um cliente da Praça de Sobral. Por outro lado, os clientes de Juazeiro apresentam 1,4 mais chances de se tornarem maus clientes.

No que diz respeito a variável Natureza ocupação, um cliente com ocupação de Empregado tem 3,9% de chances a menos de se tornar mau cliente do que um cliente com ocupação autônomo.

Quanto ao grau de instrução, um cliente cuja escolaridade seja Outros apresenta 6,2% de possibilidade de ser mau cliente ao ser comparado com um cliente com escolaridade Superior Completo.

Nota-se que os clientes que possuem outro cartão apresentam 4,1% de possibilidade de serem maus clientes ao serem comparados com aqueles clientes que não

possuem outro cartão. De forma oposta, os clientes que não possuem adicionais apresentam 31,4% de chance de deixarem de usar o cartão do que aqueles que possuem adicional.

Ao analisar a variável do estado civil, notou-se que aqueles clientes cuja o estado civil é Casado a chance de se tornar mau cliente é 1,6% menor que aquele cliente com estado civil solteiro. A variável que analisa se o cliente possui dependentes apresentou resultado de razão de chance igual a 1, e não esclarece qual a possibilidade de ser bom ou mau cliente quando se aumenta ou diminui esta variável no modelo.

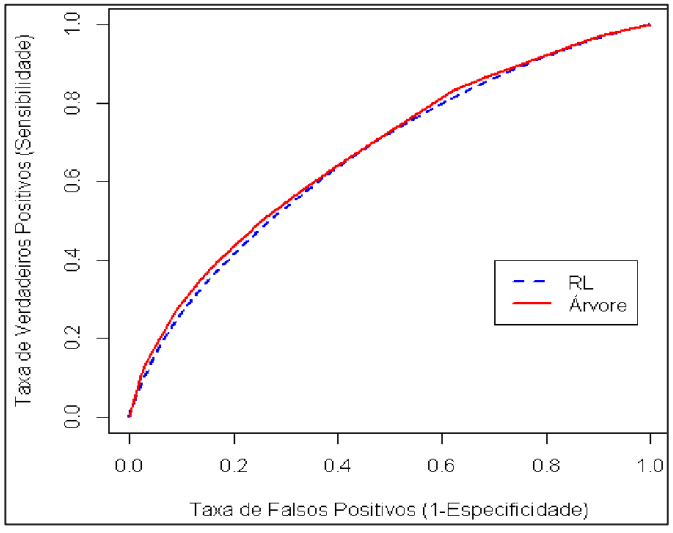
4.5 Análise comparativa do desempenho dos modelos

O modelo de Árvore de classificação resultou com um KS de 21,7%, e o modelo de Regressão Logística com 23,5%. Ambos classificados na categoria de baixa discriminação.

Ao se analisar a área abaixo da curva ROC dos dois modelos o de árvore de classificação apresentou uma área levemente maior que regressão logística, 0,676 e 0,664, respectivamente, indicando maior discriminação. Quanto maior a capacidade do teste em discriminar segundo estes dois grupos, mais a curva se aproximaria do canto superior esquerdo do gráfico e a área sob a curva seria próxima de 1.

O Gráfico 4 demonstra um maior detalhamento da curva dos modelos de regressão logística e árvore de classificação.

Gráfico 4- Curva ROC árvore gini(modelação) x Regressão Logística



Fonte: Elaborada pela autora (2014)

Ao comparar as medidas de desempenho mostradas nas Tabelas 9 e 15, para os modelos de árvore de classificação e regressão logística, respectivamente, o modelo de Regressão logística apresentou um menor numero de acurácia, consequentemente um número maior de erros, o modelo de CART apresentou um número pouco maior na taxa de *recall* ou sensibilidade, 3%, maior comparado ao da regressão logística, podendo-se afirmar que este

modelo classifica melhor os bons, por outro lado, o modelo de regressão logística apresentou uma taxa de especificidade maior, classificando 1,5% melhor os clientes maus, objetivo deste estudo. O resultado da precisão, como apresentado na seção 2.4.3, apresenta o resultado quando se aceita os bons clientes classificados como maus, o que não é o objetivo deste estudo, contudo, o modelo de regressão logística apresentou uma taxa 19% maior que a expressa pelo modelo de árvore de classificação.

5 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi um estudo de caso comparando aplicações das técnicas de *Data mining* mais difundidas, regressão logística e árvore de classificação, além de caracterizar os clientes da empresa selecionada, avaliar a influência de cada variável sócio demográfica, caracterização e histórico na empresa com o comportamento do cliente, aplicar modelos de classificação que permitam prever o comportamento do cliente e comparar o desempenho destes modelos.

Com aplicação na base de clientes de uma empresa de cartão de crédito, pôde- se mostrar que esses modelos são capazes de oferecer rendimento monetário para a instituição, visto que a empresa encontra quem são seus clientes mais propícios a deixarem de utilizar o cartão. Com esse conhecimento pode-se criar campanha de marketing lembrando o seu produto. A resposta esperada do uso de modelagem é acertar o público de clientes que irão receber a ação, obtendo o maior retorno possível.

Desta forma, foram analisados 363.172 registros, de forma que estavam igualmente divididos entre as duas classificações do critério. Construiu-se ainda, 21 variáveis que retratavam características sócio demográficas, a relação com a empresa e histórico na empresa. Onde esses indicadores foram submetidos à aplicação de modelos classificatórios.

As árvores de classificação foi o primeiro método a se testar. A técnica que usou divisões binárias com o objetivo de purificação de resultado chegou a uma classificação geral correta de 62,3%. O modelo acertou 65,7% os clientes classificados como bons e 58,8% os clientes classificados como maus.

A segunda técnica utilizada foi a de regressão logística. Onde se adotou o procedimento de *foward stepwise,* e inclusão das mesmas variáveis utilizadas na árvore de classificação. O modelo apresentou um acerto geral de 61,8%, 64% no grupo dos bons clientes e 59,7% do grupo dos maus clientes.

Como forma de melhor avaliação de desempenho utilizou-se a análise das curvas ROC resultantes dos métodos. A análise foi efetuada considerando os valores abaixo da área da curva. O modelo com resultado mais elevado foi o de árvores de classificação, com uma área abaixo da curva ROC de 0,676, considerada como sendo de discriminação aceitável.

Comparou-se ainda, o indicador KS desses modelos. Técnica muito utilizada no setor empresarial. A técnica de regressão logística apresentou um KS maior, porém, ainda considerado valor de pouca discriminação.

Na execução deste trabalho os primeiros desafios surgiram na escolha das técnicas a serem aplicadas. Outro fator foi a escolha do banco de dados. É necessário escolher e

preparar um grande volume de dados, sendo o preenchimento das variáveis com informações verídicas, com poucos dados *missing* e *outliers.* Sendo assim, é importante ressaltar o tratamento minucioso das informações para o processo de correto de modelagem.

Cabe ressaltar que, com as divisões dos dados em parcelas de treinamento com 70% e teste 30%. A presar de toda a aleatoriedade, algum vicio e distorção pode ter surgido. Talvez seja mais robusto como forma de avaliar a eficiência dos modelos testá-los com indivíduos em outro período de tempo ou realizar validação cruzada.

A respeito das variáveis, sugere-se a inclusão de outros tipos de variáveis independentes, a reavaliação do critério de classificação dos clientes em bons e maus, e a realização de estudos comparativos com aplicação de outros métodos de *Data mining,* tais como redes neurais, máquinas de suporte vetorial e *random forest.*

REFERÊNCIAS

AMO, Sandra. Mineração de dados. In: Congresso da Sociedade Brasileira de Computação, n. 24, v. 2, Salvador, 2004. **Anais.** Urbelândia: Universidade Federal de Urbelândia, 2004.

ANGELONI, Maria; REIS, Eduardo S. Business Intelligence como tecnologia de Suporte a definição de estratégias para melhoria da qualidade de ensino. In: Encontro da ANPAD, 2006, Salvador. **XXX Encontro Nacional de Pós Graduação em Administração**, v. 1, p. 16, 2006.

ANSOFF, H.I. **Estratégia empresarial**. São Paulo: McGraw Hill, 1977.

ANTONELLI, Ricardo. **Conhecendo o Business Intelligence (BI):** Uma ferramenta de auxilio a tomada de decisão. Rev. TECAP, Rio de Janeiro, v. 3, n. 3. 2009

BASGALUPP, Márcio Porto. **LEGAL-Tree**: Um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão. Tese (Doutorado em Ciências da computação e matemática computacional) – Universidade de São Paulo, São Paulo, 2010.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados**: Conceitos, Tarefas, Métodos e Ferramentas. Relatório Técnico – Universidade De Goiás, Goiás, 2009.

CARNUT, Leonardo *et al*. **Validação inicial** ... Recife: ed. Universidade de Pernambuco, 2008.

CASA NOVA, Silvia Pereira de Castro *et al.* Modelos de previsão de insolvência utilizando a análise por envoltória de dados: aplicação a empresas brasileiras. Rev. adm. contemp. Curitiba , v. 11, n.2, 2007. Disponível em:< http://www.scielo.br/scielo.php?pid=S1415-65552007000600005&script=sci_arttext>. Acesso em: 10 de maio de 2014

CONFEDERAÇÃO NACIONAL DE DIREITOS LOJISTAS; SISTEMA DE PROTEÇÃO AO CREDITO BRASIL. **Cartões de credito**: quem paga a conta afinal? [20--].

CONSUMIDOR. Cartões de crédito. Disponível em: < http://www.soleis.com.br/cartao_credito.htm>. Acesso em: 22 de mar de 2014.

CÔRTES, Sergio; PORCARO, Rosa Maria; LIFSCHITZ, Sergio. **Mineração De Dados: Funcionalidades, Técnicas E Abordagens**. Rio De Janeiro: Ed. PUC, 2002.

ELMASRI, Ramez; NAVATH, Shamkant B. **Sistema de Banco de dados**. 4 ed. Pearson: São Paulo, 2006. 643 p. Tradução: Fundamentais of Database systems.

FAYYAD, Usama. M., *et al***. Advances in Knowledge Discovery and Data mining**. 1 ed. California: AAAI Press/The MIT Press, p. 1-34, 1996.

FIGUEIRA, Cleonis Viater. **Modelos de Regressão Logística**. Porto Alegre: ed. Universidade Federal do Rio grande do Sul. 2006.

FREITAS, Paulo Springer de. Mercado de cartões de credito no Brasil: problemas de regulamentação e oportunidade de aperfeiçoamento de legislação. Serie textos para discussão, **Consultoria Legislativa do Senado Federal**. Brasília, n. 37, dez. 2007
FROTA, Diego. **Estudos em Regressão Logística**. São Paulo: Unicamp, 2011. Disponível em: < http://vigo.ime.unicamp.br/Projeto/2011-2/ms777/ms777_Diego.pdf>. Acesso em: 20 de abr de 2014.

GARCIA, Simone Carboni. **O uso de árvores de decisão na descoberta de conhecimento na área da saúde.** Rio Grande do Sul: ed. Universidade Federal do Rio Grande do Sul. 2003

GONÇALVES, Eric Barcconi; GOUVÊA, Maria Aparecida; MANTOVANI, Daielly Melina Nassif. **Análise do Risco de Credito com a regressão logística**. Revista Contemporânea de Contabilidade, Florianópolis, v.10, n.20, p.139-160, mai/ago. 2013.

HAIR, Joseph, *et al*. **Análise Multivariada de dados**. 6 ed. Bookman: Rio Grande do Sul, 2009. 679 p. Tradução: Multivariate data analysis.

HASTIE, Trevor. et al.. 2009. **The elementes of statistical learning**: Data mining, inference, and predction. Second edition. Springer series in statistics. Springer. 2nd ed 2009.

KANSO, Solange. Utilização da regressão logística para a classificação de famílias quanto à condição de pobreza nas RMs do Rio de Janeiro e Recife nos anos de 1970,1980 e 1991. In: Encontro nacional de Estudos Populacionais, 14, 2004, Minas Gerais, **ABEP.** Minas Gerais: IPEA, 2003.

LEMOS, Eliane P. **Análise de credito bancário com o uso de Data mining**: redes neurais e árvores de decisão. Dissertação (Mestrado em Ciências) – Universidade do Paraná, Curitiba, 2003.

LEITE, Cesar Eduardo; VITORIO JUNIOR, Daudt. A utilização de ferramenta estatística na tomada de decisão gerencial. In: Congresso Nacional de Excelência em Gestão, 4, 2008, Niterói. **Anais**. Niterói**:** Congresso Nacional, 2008.

LIMA, Josimara Alves De. **Liderança E Tomada De Decisão Na Organização**. Dissertação (MBA Em Administração Estratégica e Financeira) – Universidade do Oeste de Santa Catarina, Santa Catarina, 2012.

MARTINHAGO, Dariana Zanella *et al.* **Data mining**: definição, importância, aplicação na gestão organizacional. Lavras: Universidade federal de lavras, [2004?]

MATUSSE, Euclides Alfredo *et al.* **Uma estratégia de alocação**... Paraná: ed. Universidade Estadual de Maringá, [2012?].

MENDES. Carlos André Bulhões; VEGA, Fausto Alfredo Canales. **Técnicas de regressão logística aplicada à analise ambiental**. Revista Geográfica, Rio Grande do Sul, v. 20, n.1, p. 5-30, jan/abr. 2011.

MENEZES, Igor Gomes; Bittencourt, Antonio Virgílio. Construção, desenvolvimento e validação da escala de intenções comportamentais de comprometimento organizacional. **Instituto Brasileiro de avaliação psicológica**, v.9, n. 1, p. 119-127, abr. 2010.

MORAES, Luciane de Godói. **Uma abordagem alternativa de behavioral scoring usando modelagem híbrida de dois estágios com regressão logística e redes neurais**. Porto Alegre: ed. Universidade Federal do Rio Grande do Sul. 2012.

MUSZINSKI, Andre Amaral; BERTAGNOLLI, Silvia De Castro. **Business Intelligence**: Um Sistema De Apoio a decisões gerenciais. Porto Alegre: Centro Universitário Ritter Dos reis. Porto Alegre, [20--]

NAVEGA, Sergio. ***Principios Essenciais do Data mining***. Anais do Infoimagem 2002, Cenadem, Novembro/2002. Disponível em: < http://www.intelliwise.com/snavega>.

PAULA, Glberto A. **Modelos de regressão com apoio computacional.** São Paulo: Universidade de São Paulo, 2010.

PALMUTI, Claudio Silva; PICCHIAI, Djair. **Mensuração do risco de crédito por meio de análise estatística multivariada**. Revista Economia Ensaios, v. 26, n. 2, p. 7–22. 2012.

PEDRO, Silvia Ferreira. **Aplicação de métodos estatísticos na avaliação da satisfação dos utentes com internamento hospitalar**. Dissertação (mestrado em gestão empresarial)- Universidade do Algarve, Faro, 2007.

PEDRO, Silvia Maria Dias. **Exploração de dados aplicada a analise de risco de credito**. Inescp Id: Instituto Superior Técnico, 2001.

PINTO, Ana Valéria Monteiro**. Percepção dos alunos da feaac em relação ao curso de ciências atuariais**. Fortaleza: Ed. Universidade Federal do Ceará, 2011.

PREARO, Leandro Campi; GOUVÊA, Maria Aparecida; MONARI, Carolina. **Avaliação do emprego da técnica de análise de regressão logística em teses e dissertações de algumas instituições de ensino superior.** Londrina: ed. Semina: Ciências Sociais e Humanas, v. 30, n. 2, p. 123-140. 2009

PRIMAK, Fábio V. **Decisões com o B.I. (Business Intelligence)**. Rio de janeiro: Ciência Moderna, 2008.

REBOUÇAS, Silvia Pedro. **Árvores de classificação e regressão**. Fortaleza: Universidade Federal do Ceará, 2013. 16p.

ROBBINS, S. P.; DECENZO, D. A. **Fundamentos de Administração**: conceitos e aplicações, São Paulo: Prentice Hall 2006.
Sistema de Proteção ao Crédito. SPC BRASIL. Relatório de Pesquisa. Uso do Crédito. Junho de 2013.

RODRIGUES, Willian. **Metodologia Científica**. Paracambi: FAETEC/IST, 2007. Disponível em: < http://pesquisaemeducacaoufrgs.pbworks.com/w/file/fetch/64878127/Willian%2520Costa%2520Rodrigues_metodologia_cientifica.pdf>. Acesso em: 20 de abr de 2014.

SANTOS, Lucas Maia dos; FERREIRA, Marco Aurélio Marques; FARIA, Evandro Rodrigues de. **Utilização de modelos de regressão logística para a previsão de risco de liquidez em micro e pequenas empresas**. Ed: ABCustos Associação Brasileira de Custos, Rio Grande de Sul, v. 4, n 3, set/dez. 2009.

SANTOS, Paulo Jorge dias dos. **Testes não paramétricos para validação de modelos extremais**. Dissertação (mestrado em estatística) – Universidade De Lisboa, Lisboa, 2011.

SANTOS, Roberto de Souza. **Aplicação de um modelo um modelo preditivo de mineração de dados para apoio à decisão de credito.** Dissertação (mestrado em ciências da informação) – Universidade federal de Minas Gerais. 2006

SILVA, Ana Bela Costa da. **Análise estatística de inquéritos online**. Dissertação (Mestrado em Estatística de Sistema) – Universidade do Minho, Portugal, Outubro de 2011.

SOARES, G. O. G.; COUTINHO, E. S.; CAMARGOS, M. A. Determinantes do rating de crédito de companhias brasileiras. **Contabilidade Vista & Revista**, v. 23, n. 3, p. 109-143, 2012.

SOARES, Rômulo Alves. **Modelos de classificação aplicados à previsão de insolvência de empresas brasileiras de capital aberto**. Fortaleza: Universidade Federal do Ceará, 2013.

SOUZA, Elton Samaroque De. NEIVERTH, Jaques. **Estudo de caso sobre o Business Intelligence.** Ponta Grossa: Universidade Estadual De Ponta Grossa, 2007.

STEINER, Maria Teresinha Arns *et al.* Data mining como suporte à tomada de decisões -uma aplicação no diagnóstico médico. In: Simpósio Brasileiro de Pesquisa Operacional, 26, 2004, Minas Gerais. **SBPO**. Minas Gerais, [s.n], 2004.

WALTER, Silvana Anita *et al.* **Lealdade de estudantes: um modelo de regressão logística**. Revista de administração FACES Journal, v. 10, n. 4, p. 139-151, set/dez. 2010.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin**. Introdution to Data mining**. 1. ed. Boston: Addison-Wesley, 2005.

TARAPANOFF, Kira *et al.* Inteligência obtida pela aplicação de Data mining em base de teses francesas sobre o Brasil**. Ci. Infor,** Brasília, v 30, n 2, p. 20-28, maio/ago, 2001.

TSUKAHARA, Fábio Yasuhiro. **Adequação das técnicas de validação dos modelos de probabilidade de *defeault* em carteiras simuladas.** São Paulo: Universidade Presbiteriana Mackenzie, 2013.

WANDERLEY, Ana Valéria Medeiros. **Sistemas de Inteligência Competitiva.** Disponível em: < http://www.profcordella.com.br/unisanta/textos/sin24_conceito_inteligencia_competitiva.htm >. Acesso em: 10 de maio de 2014.