UNIVERSIDADE FEDERAL FLUMINENSE
CENTRO TECNOLÓGICO
MESTRADO EM ENGENHARIA DE PRODUÇÃO

MÔNICA CARVALHO DA FONSECA

REGRESSÃO LOGÍSTICA APLICADA À MANUTENÇÃO DE CLIENTES DE CARTÕES DE CRÉDITO

NITERÓI
2008

MÔNICA CARVALHO DA FONSECA

REGRESSÃO LOGÍSTICA APLICADA À MANUTENÇÃO DE CLIENTES DE CARTÕES DE CRÉDITO

Dissertação apresentada ao Curso de Mestrado em Engenharia de Produção da Universidade Federal Fluminense como requisito parcial para obtenção do Grau de Mestre. Área de Concentração: Estratégia, Gestão e Finanças.

Orientador: Prof. Dr. HELDER GOMES COSTA

Niterói
2008

Ficha Catalográfica elaborada pela Biblioteca da Escola de Engenharia e Instituto de Computação da UFF

F676 Fonseca, Mônica Carvalho da.
Regressão logística aplicada à manutenção de clientes de cartões de crédito / Mônica Carvalho de Fonseca. – Niterói, RJ : [s.n.], 2008.
70 f.

Orientador: Helder Gomes Costa .
Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal Fluminense, 2008.

1. Regressão logística. 2. Cartão de crédito. 3. Processo de decisão. 4. Fidelização – Programa. 5. Engenharia de produção. I. Título.

CDD 658.5

MÔNICA CARVALHO DA FONSECA

REGRESSÃO LOGÍSTICA APLICADA À MANUTENÇÃO DE CLIENTES DE CARTÕES DE CRÉDITO

Dissertação apresentada ao Curso de Mestrado em Engenharia de Produção da Universidade Federal Fluminense como requisito parcial para obtenção do Grau de Mestre. Área de Concentração: Estratégia, Gestão e Finanças.

Aprovada em NOVEMBRO de 2008.

BANCA EXAMINADORA

Prof. Dr. Helder Gomes Costa – Orientador
Universidade Federal Fluminense

Prof. Dr. Annibal Parracho Sant'Anna
Universidade Federal Fluminense

Prof. Dr. Luiz Alberto Nascimento Campos Filho
Faculdades de Economia e Finanças - IBMEC RJ

Niterói
2008

AGRADECIMENTOS

Aos professores do Curso de Mestrado em Engenharia de Produção da UFF, pelos ensinamentos transmitidos.

Ao meu orientador Professor Helder Gomes Costa pela paciência, estímulo e dedicação ao mostrar o caminho a ser seguido na realização deste trabalho. Agradeço, principalmente, pelas dicas preciosas e inteligentes nos momentos mais difíceis da construção deste projeto.

A meus pais, os responsáveis diretos por toda essa história.

Aos amigos e colegas de trabalho que desenvolveram em mim o gosto pela modelagem e que colaboraram com sugestões, apoio e energia positiva, acreditando comigo no êxito deste estudo.

Aos amigos pelo incentivo. Um agradecimento especial a Ana Luzia, Paula Oliveira e Ana Maria, pela ajuda em pontos fundamentais.

A todos que de uma forma ou de outra contribuíram para a realização deste trabalho.

RESUMO

A concorrência entre as instituições financeiras na disputa por novos mercados, fidelização de clientes e a busca por novos clientes têm desencadeado uma permanente melhoria nos processos. Assim, o mercado busca utilizar cada vez mais diversas metodologias estatísticas para auxiliar suas decisões. Várias técnicas estatísticas já foram utilizadas, entre elas: análise discriminante, análise de sobrevivência, árvores de decisão, redes neurais e regressão múltipla. Neste âmbito, destacam-se as aplicações de modelos estatísticos a processos de: concessão de crédito; manutenção de clientes; cobrança de clientes em atraso; e, na retenção dos clientes. O presente estudo tem como objetivo desenvolver um modelo de regressão logística, (modelo *anti-attrition scoring*) para auxiliar na fidelização de clientes de um determinado cartão de crédito, identificando, antecipadamente, clientes que tenham alta probabilidade de não utilização no futuro. Os resultados obtidos mostram que o modelo é bastante eficiente em separar as contas que possivelmente deixarão de utilizar o cartão de crédito, dado que o modelo proposto apresentou um elevado grau de explicação.

Palavras-chave: Cartões de Crédito, Regressão Logística, Decisão.

ABSTRACT

The competition between financial institutions in the dispute for new markets, loyalty of customers and the search for new customers has triggered a permanent improvement in the processes of these institutions. According to this, the market increasingly seeks to use different statistics methodologies to assist their decisions. Several statistical techniques have been used, including: discriminant analysis, survival analysis, decision trees, Bayesian inference, neural networks and multiple regression. In this context, there are highlighted applications of statistical models in: credit concession, maintenance of customers; collection of customers in arrears, and the retention of customers. This study aims to develop a model of logistics regression to assist the retention of customers of a particular credit card, identifying, in advance, customers who have high potential for non-use of the product in the future. The results show that the model is very efficient on separating accounts that should no longer use the credit card, as the proposed model showed high performance.

Keywords: Credit cards, Logistic Regression, Decision.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

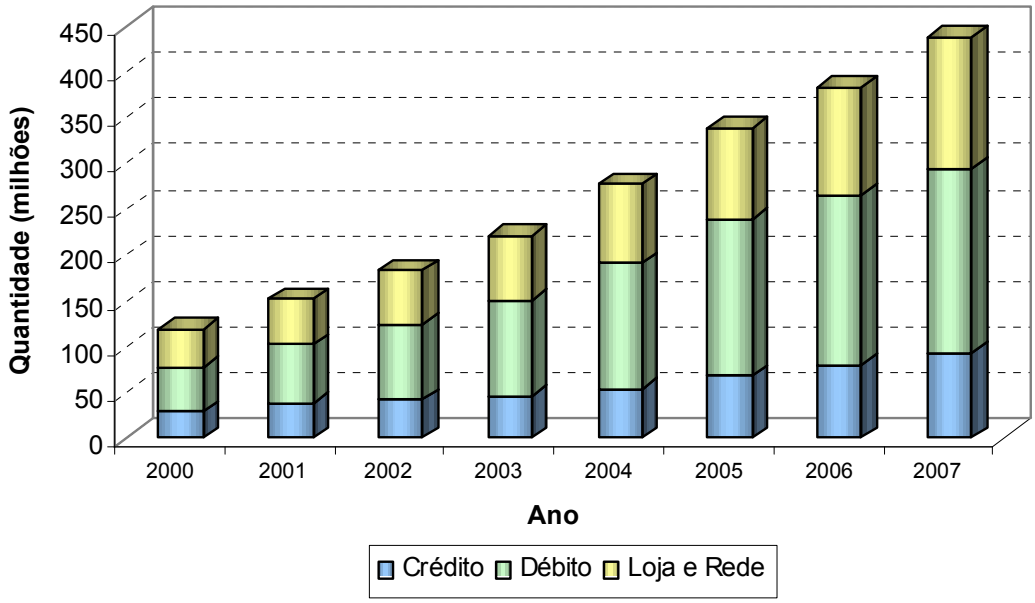
Os cartões de plástico estão substituindo cada vez mais as cédulas, moedas e cheques. Apesar de ser uma forma alternativa de pagamento, eles não têm poder de compra, eles apenas registram a intenção do pagamento com dinheiro real. Cedo ou tarde, a despesa deverá ser paga em espécie ou cheque. Pode-se definir então como uma forma imediata de autorização de pagamento, a débito, crédito ou transferência.

Os cartões de banco garantem o acesso a cheques, retirada de dinheiro em espécie e pagamentos de contas. Hoje existe no mercado uma grande variedade de tipos de cartões. São cartões de descontos, benefícios, afinidade, de apoio à campanhas sociais e ecológicas; cartões que atendem a diversos perfis socias e econômicos. Existem ainda os cartões de plástico que apenas concedem benefícios e descontos. Outros cartões oferecem ainda todas as opções num único plástico.

Instituições financeiras, bancos e grandes lojas de departamentos enviam para seus clientes cartões exclusivos para utilização em sua rede conveniada para compras de diversos bens e serviços. Além disto, eles podem ser utilizados em outras redes e até mesmo através da internet. O objetivo é fidelizar os clientes e facilitar a compra eliminando a burocracia na abertura do crédito.

Atualmente, o mercado de cartões está em grande expansão no Brasil, possuindo aproximadamente 438 milhões de cartões (ABECS, 2008). No ano de 2007 houve um aumento de 14% em relação a todo o ano de 2006. Comparando ainda o ano de 2007 com 2000 observa-se um incremento superior a 250% no número total de cartões de crédito, débito e de lojas emitidos no Brasil, conforme mostra o gráfico 1 (ABECS, 2008).

Gráfico 1 - Evolução da Quantidade de Cartões Emitidos no Brasil 2000-2007



Fonte: ABECS (2008)

1.1. MOTIVAÇÃO

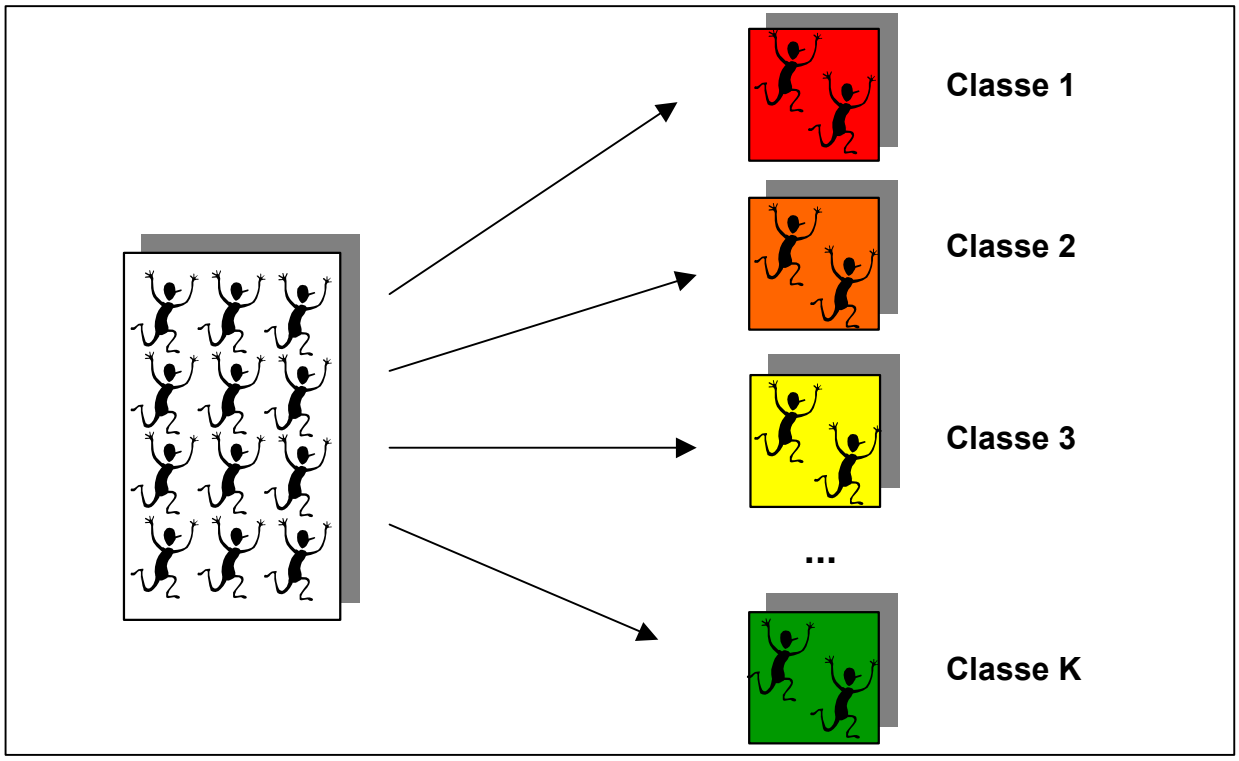
A concorrência entre as instituições financeiras na disputa por novos mercados e clientes e a fidelização dos antigos clientes tem desencadeado uma permanente melhoria nos processos destas instituições.

Assim, o mercado busca utilizar cada vez mais diversas metodologias estatísticas para auxiliar suas decisões. Várias técnicas estatísticas já foram utilizadas, entre elas: análise discriminante (West, 2000), análise de sobrevivência (Ferreira, 2007), árvores de decisão (Huang *et al*, 2006), redes neurais (Gonçalves, 2005) e regressão múltipla (Pereira, 2004). Neste âmbito, destacam-se as aplicações de modelos estatísticos a processos de: concessão de crédito; manutenção de clientes; cobrança de clientes em atraso; e, na retenção dos clientes.

Um fator decisivo para o crescimento e sobrevivência das empresas diz respeito ao comportamento dos clientes frente às mudanças do mercado. O desafio deixa de ser a obtenção de novos clientes e passa a ser a manutenção dos atuais. Este desafio também é colocado para as empresas financeiras, demandando o estudo e a construção de modelos estatísticos específicos para este setor e que permitam analisar o comportamento de seus clientes.

Um dos problemas centrais neste contexto é o de ranquear os membros de uma base de dados, no caso deste estudo, os clientes de uma instituição financeira. A classificação dos clientes em classes ordenadas, conforme ilustra a figura 2, é de grande importância para auxiliar a tomada de decisão das grandes instituições financeiras.

Figura 2 – Ranqueamento de Clientes



Fonte: Adaptada de Campello (2005)

1.2. OBJETIVO

Objetivando contribuir para a solução do problema destacado na seção anterior, o presente estudo tem como objetivo desenvolver um modelo de regressão logística, modelo *anti-attrition scoring* (Burez e Poel, 2008), para ranquear os clientes que possuem um determinado cartão de crédito, identificando, antecipadamente, clientes que tenham alta probabilidade de não utilização do produto no futuro.

Além de ranquear os clientes, pode-se identificar as principais características das contas do cartão de crédito que possuem alta probabilidade de inatividade.

1.3. ESTRUTURAÇÃO DO TRABALHO

Este trabalho é composto por seis capítulos e referências bibliográficas.

O capítulo 1 traz uma introdução sobre o assunto tratado, o objetivo e motivação deste trabalho, juntamente com sua estruturação.

O capítulo 2 discute a revisão bibliográfica através das diversas técnicas utilizadas para a identificação de possíveis rupturas no relacionamento de clientes com as empresas de cartões de crédito.

O capítulo 3 expõe os conceitos teóricos sobre regressão logística.

No capítulo 4 são apresentados detalhadamente os conceitos da base de dados, a análise exploratória dos dados, o desenvolvimento do modelo com o modelo final estimado e sua análise.

No último capítulo, apresentam-se as conclusões e também sugestões para trabalhos futuros.

2. REVISÃO BIBLIOGRÁFICA

Os altos níveis de inadimplência nas instituições financeiras, conforme apresentado na tabela 1, geram cada vez mais o interesse na utilização de metodologias estatísticas que auxiliem nas tomadas de decisão. A metodologia mais comumente utilizada é a regressão. Vários tipos de modelos são aplicados nas diversas fases do cliente na instituição. Os modelos de *application scoring* são mais conhecidos como modelos de concessão, utilizados para decidir se um novo crédito deve ou não ser dado a um determinado cliente (Thomas *et al*, 2002).

Tabela 1 – Indicadores de Inadimplência

Indicadores de Crédito e Inadimplência	Data	Último
Cheques sem Fundos (Usecheque)	Set/08	2,13M
Inadimplência P. Física (BC)	Ago/08	7,50%
Inadimplência P. Jurídica (BC)	Ago/08	1,70%
Inadimplência PF Novos Registros (SCPC)	Set/08	479,27K
Inadimplência PF Cancelamentos (SCPC)	Set/08	365,99K
Endividamento (% - Fecomércio)	Set/08	53
Inadimplência (% - Fecomércio)	Set/08	30
Comprometimento (% - Fecomércio)	Ago/08	34

Fonte: Infomoney (2008)

Após a entrada do novo cliente na instituição, modelos de acompanhamento são aplicados constantemente, os chamados *behaviour scoring*. Auxiliam na manutenção do risco de crédito, concedendo novos créditos, ajustando limites ou até mesmo ofertando novos produtos. São baseados principalmente em informações de comportamento do cliente dentro da instituição, gerando assim, uma maior discriminação quando comparados com os modelos de concessão (Hoper e Lewis, 1992).

Além dos modelos já citados, existe ainda o modelo de *anti-attrition scoring*, que tem como principal objetivo prever se o cliente possui alta probabilidade de não utilização ou cancelamento do produto nos próximos n meses. A grande importância deste modelo é que se pode tomar uma decisão antecipada que evite a ruptura no relacionamento.

Pereira (2004) compara três estratégias para a construção de modelos de risco de crédito de clientes de instituições financeiras. A primeira desenvolve o modelo em duas etapas, a segunda ajusta vários modelos simultâneos e finalmente a terceira, que ajusta vários modelos simultâneos com variáveis respostas diferentes. O melhor resultado obtido foi com a segunda estratégia, que poupa tempo e tem melhor desempenho.

Ferreira (2007) apresenta técnicas como análise de sobrevivência e regressão logística para prever o comportamento de clientes e assim focar estratégias de modo a alocar recursos para impedir a migração dos mesmos para outras instituiçoes financeiras. As variáveis utilizadas no modelo logístico são: média de faturamento anual da conta, número de meses em que o cliente fica no rotativo, grupo comportamental de consumo ao qual o cliente pertence, possui ou não débito automático e se possui seguro contra roubo/furto.

Régis (2007) analisa a aplicação do modelo multi-estado de Markov na área de risco associado ao uso de cartões de crédito, aproveitando as características de transições entre diversos estados de relacionamento entre os clientes e as instiuições ao longo do tempo. Modelos de regressão logística também foram estimados a fim de comparar os resultados com os obtidos pelo modelo anterior. Foram construídos modelos para estimar as probabilidades de o cliente ficar inativo e em atraso. O

comportamento dos clientes foi mensurado utilizando-se as seguintes variáveis: máxima quantidade de meses com utilização de crédito rotativo, máxima quantidade consecutiva de meses sem comprar, média da quantidade de compras, média do percentual de utilização do limite de crédito e maior percentual de utilização do limite de crédito. Nos modelos utilizados para prever os clientes em atraso, verificou-se um melhor desempenho com os modelos multi-estado de Markov. Já nos modelos de inatividade, obteve-se vantagem com os modelos ajustados a partir da regressão logística.

Abreu (2004) utiliza análise de sobrevivência e compara brevemente seus resultados com a regressão logística, mostrando uma similaridade grande entre os resultados das duas técnicas. Também é avaliado que a quantidade de clientes utilizados para o estudo é fundamental para uma análise acertiva.

Gonçalves (2005) desenvolve três modelos, aplicando três técnicas para a classificação de clientes: Regressão Logística, Redes Neurais e Algoritmos Genéticos. A regressão logística apresenta o melhor resultado apesar de muito próxima dos resultados de redes neurais. O modelo de algoritmos genéticos apresenta resultados bons, mas em um patamar inferior aos anteriores.

Vansconcellos (2002) aplica a regressão logística em uma base de clientes de pessoa física de uma instituição financeira para auxiliar a concessão de crédito. O estudo também sugere ferramentas para verificação da qualidade do modelo estimado, assim como, para acompanhamento da performance do modelo ao longo do tempo.

Onusic e Viana (2004) constroem modelos de previsão de insolvência utilizando Análise Envoltória de Dados e Regressão Logística com indicadores contábeis e financeiros derivados das demonstrações contábeis de empresas brasileiras.

Brito (2006) desenvolve um modelo de classificação de risco para avaliar o risco de crédito de empresas no mercado brasileiro. A técnica utilizada é a regressão logística. Os resultados do estudo indicam que o modelo de classificação de risco desenvolvido prevê eventos de *default* com bom nível de acurácia.

Kimura *et al* (2005) apresenta uma aplicação de redes neurais para a identificação de bons e maus pagadores em operações de crédito ao consumidor a partir de uma base de dados real.

West (2000) investiga a acurácia de modelos quantitativos que utilizam redes neurais para construir modelos de crédito. A pesquisa mostra que redes neurais é uma boa técnica para o desenvolvimento destes modelos e sugere ainda que a regressão logística é uma ótima alternativa.

Hayhoe *et al* (1999) utiliza a regressão logística e estuda a utilização do crédito por estudantes para prever a quantidade de cartões de crédito.

Huang *et al* (2006) constrói modelos de crédito utlizando algoritmos genéticos e compara com redes neurais, análise discriminante, regressão logística e árvore de decisão. Baseado no resultados empíricos, conclui que algoritmos genéticos têm uma melhor performance nos problemas de modelos de crédito.

Lim e Sohn (2007) propõem um *behaviour scoring* para auxiliar na identificação dos clientes com alto risco de inadimplência. O estudo possui uma limitação de tamanho de amostra.

Sohn e Shin (2006) aplicam um método de inferência dos negados em operações de crédito baseados em análise de sobrevivência.

Crook *et al* (2007) utiliza técnicas estatísticas como regressão logística e redes neurais no desenvolvimento de modelos de crédito.

Chuang (2008) compara técnicas estatísticas no desenvolvimento de modelos de crédito. Os melhores resultados são obtidos nas técnicas regressão logística e redes neurais.

Mendes (2008) desenvolve um modelo estatístico que relaciona variáveis transacionais, demográficas e sobre o histórico de eventos, com a probabilidade de cancelamentos dos clientes assinantes; e define o perfil dos clientes com maior risco de cancelamento em planos de saúde.

Ribeiro (2007) produz predições do valor de tempo de vida de clientes através de metodologias distintas, tendo em vista a comparação entre elas.

A seguir um quadro resumo dos trabalhos consultados contendo: autor, ano, metodologia aplicada, contexto da aplicação, tipo de modelo, variáveis utilizadas e o melhor resultado encontrado.

Autor	Ano	Metodologia	Contexto	Tipo
Pereira	2004	Regressão Logística utilizando várias simulações	Crédito	Behaviour scoring
Ferreira	2007	Análise de Sobrevivência Regressão Logística	Cartão de Crédito	Anti-attrition scoring
Régis	2007	Modelos Multi-Estado de Markov Regressão Logística	Cartão de Crédito	Anti-attrition scoring
Abreu	2004	Análise de Sobrevivência Regressão Logística	Crédito	Application scoring
Gonçalves	2005	Redes Neurais Regressão Logística Algoritmos Genéticos	Crédito	Application scoring
Vasconcellos	2002	Regressão Logística	Crédito	Application scoring
Onusic, Viana	2004	Análise Envoltória de Dados Regressão Logística	Empresas	Insolvência
Brito	2006	Regressão Logística	Empresas	Anti-attrition scoring, Insolvência
Kimura et al	2005	Redes Neurais	Crédito	Application scoring

Autor	Ano	Metodologia	Contexto	Tipo
West	2000	Redes Neurais Análise Discriminante Regressão Logística Árvore de Decisão Cluster	Crédito	Application scoring
Hayhoe et al	1999	Regressão Logística	Cartão de Crédito	Behaviour scoring
Huang et al	2006	Algoritmos Genéticos Redes Neurais Análise Discriminante Regressão Logística Árvore de Decisão	Crédito	Application scoring
Lim, Sohn	2007	Cluster	Crédito	Behaviour scoring
Sohn, Shin	2006	Análise de Sobrevivência	Crédito	Application scoring
Crook et al	2007	Regressão Logística Redes Neurais	Crédito	Behaviour scoring
Chuang	2008	Regressão Logística Redes Neurais Árvore de Decisão	Crédito	Application scoring
Mendes	2008	Regressão Logística	Plano de Saúde	Anti-attrition scoring
Ribeiro	2007	Modelos Estocásticos	Supermercado	Behaviour scoring

Autor	Ano	Variáveis	Melhor Resultado
Pereira	2004	Não divulgadas	Regressão Logística com vários modelos simultâneos
Ferreira	2007	Média de faturamento anual da conta Número de meses em que o cliente fica no rotativo Grupo comportamental de consumo ao qual o cliente pertence Possui ou não débito automático Possui seguro contra roubo/furto	Regressão Logística
Régis	2007	Máxima quantidade de meses com utilização de crédito rotativo Máxima quantidade consecutiva de meses sem comprar Média da quantidade de compras Média do percentual de utilização do limite de crédito Maior percentual de utilização do limite de crédito	Regressão Logística
Abreu	2004	Estado Civil Tipo de Cliente Sexo Tipo de Residência Possui Cartão Idade Comprometimento de Renda Valor do Crédito Concedido Região Profissão CEP	Análise de Sobrevivência
Gonçalves	2005	Sexo Estado Civil Telefone Tempo de Emprego Salário Qtd de Parcelas a serem quitadas Primeira Aquisição Tempo de Residência Valor da Parcela Valor Total do Empréstimo Idade CEP Profissão Salário	Regressão Logística

Autor	Ano	Variáveis	Melhor Resultado
Vasconcellos	2002	Nome Idade Sexo Estado Civil Escolaridade Profissão Tempo de Emprego Renda Patrimônio Informações Bancárias Data de Vencimento das Prestações Data de Pagamento das Prestações Valores Totais das Prestações Valores dos Juros das Prestações Valores das Amort	Regressão Logística
Onusic, Viana	2004	Endividamento Geral Endividamento de Longo Prazo Composição do Endividamento Crescimento de Vendas Retorno sobre o Ativo Giro do Ativo	Análise Envoltória de Dados
Brito	2006	Liquidez Retorno Giro Margem Lucro Patromônio Endividamento Capital Saldo Fluxo	Regressão Logística
Kimura et al	2005	Variáveis de Cadastro	Redes Neurais
West	2000	Idade Residência Trabalho Histórico de Crédito Empréstimo	Redes Neurais
Hayhoe et al	1999	Idade Sexo Formas de Utilização do Dinheiro	Regressão Logística
Huang et al	2006	Idade Sexo Estado Civil Trabalho Histórico de Crédito Proposta de Empréstimo	Algoritmo Genético
Lim, Sohn	2007	Application e Behavorial Variáveis	Cluster

Autor	Ano	Variáveis	Melhor Resultado
Sohn, Shin	2006	Idade Trabalho Residência Montante do Empréstimo Valor do Automóvel Número de Cartões de Crédito	Análise de Sobrevivência
Crook et al	2007	Sócio-Demográficas Histórico de Crédito	Regressão Logística
Chuang	2008	Variáveis Cadastrais Histórico de Crédito Emprego Residência	Regressão Logística Redes Neurais
Mendes	2008	Sexo Estado Civil Idade Bairro Agregados Dependentes Opcionais Tempo de Consulta Tempo de Exame Valor Pagamentos Internação Cirurgia	Regressão Logística
Ribeiro	2007	Receita Freqüência de compras Meses de compras Ticket Médio de compra Receita Média Mensal Tempo de Inatividade Recência Produtos distintos Lojas distintas Tempo médio entre compras	Modelos Estocásticos

A partir dos trabalhos citados, verifica-se que na maioria dos casos a regressão logística apresenta um melhor desempenho. A variação, entre ser a melhor técnica ou não, ocorre devido às diferentes bases de dados onde as técnicas são aplicadas e a forma de desenvolvimento dos modelos. Assim, a regressão logística desempenha um papel importante no auxílio de uma visão de risco comportamental, no caso deste estudo, a identificação da probabilidade de inatividade.

3. BASE CONCEITUAL: REGRESSÃO LOGÍSTICA

Uma das técnicas mais utilizadas para a construção de modelos *attrition scoring* tem sido, historicamente, a regressão linear com resposta binária. Este método combina os coeficientes com as respostas das variáveis explicativas de forma a gerar contribuições individuais que, somadas, resultam no escore final. Apesar de apresentar estimadores não viesados e consistentes este método apresenta problemas de heterocedasticidade (Cook e Weisberg, 1983) (já que a variância dos resíduos depende dos valores das variáveis explicativas, ou seja, não é constante), e a principal limitação é muito clara: os valores previstos para a variável resposta não pertencem necessariamente ao intervalo [0,1], podendo assumir valores negativos e até mesmo maiores que 1, ou seja, não condizentes com o problema estudado, em que o intervalo deve ser obedecido.

O método de regressão logística é, por definição, apropriado para estudos em que a variável resposta assume valores 0 ou 1, e formula uma equação não linear entre as variáveis explicativas e a variável resposta, devido à forma funcional do método, que apresenta funções exponenciais relacionadas às variáveis explicativas com a variável resposta. Assim, o método parece mais apropriado que a regressão linear. O método de regressão logística foi escolhido pelo fato do problema em questão ter uma variável dependente binária (a conta estar ativa ou inativa nos próximos seis meses), o que torna o método mais apropriado, além de ser computacionalmente simples.

A facilidade de revisar o modelo periodicamente é importante porque, visto que o modelo é baseado em utilizações das contas do cartão de crédito passadas, o modelo somente será robusto enquanto as características de utlização das novas contas forem semelhantes às características encontradas no passado, ou seja, se o perfil da população da carteira se mantiver estável.O objetivo de qualquer técnica de construção de modelos estatísticos é encontrar uma forma funcional adequada e parcimoniosa para descrever o relacionamento entre uma variável resposta (dependente) e um conjunto de variáveis independentes (explicativas).

O modelo de regressão logística pode ser extremamente útil para descrever a relação entre um desfecho binário de uma ou mais variáveis explicativas, pois permite estimar a magnitude e a direção dos efeitos dos preditores. O modelo ajustado também pode ser usado para predizer o desfecho para um particular indivíduo, com base nos correspondentes valores observados das variáveis explicativas. Se o modelo é parcimonioso e possui boa capacidade preditiva, então pode ser muito valioso na prática.

Levando-se em consideração que o principal objetivo deste trabalho é estudar a utilização da conta nos próximos seis meses, foram ajustados modelos de regressão logística para estimar a **probabilidade de uma conta estar inativa nos próximos seis meses.**

É conveniente descrever resumidamente os aspectos desta metodologia. Para tanto, considere Y_i uma variável aleatória binária que assume valores $Y_i = 0$ (fracasso) e $Y_i = 1$ (sucesso), enquanto as variáveis explicativas são denotadas pelo vetor $\underset{\sim i}{x}' = (x_{i1}, x_{i2}, ..., x_{ip})$.

Sendo Y_i uma variável binária, a probabilidade de sucesso da resposta binária será representada por $P(Y_i = 1) = \pi_i$, e naturalmente, a probabilidade de fracasso é $P(Y_i = 0) = 1 - \pi_i$, sendo $i = 1, ..., n$ (Vasconcellos, 2002). Logo, a probabilidade conjunta é dada por:

$$P(Y_1, Y_2, ..., Y_n) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \qquad (3.1)$$

O modelo de regressão logística relaciona a probabilidade de sucesso π_i com as variáveis explicativas através da seguinte forma funcional (Neter e Wasserman, 1974):

$$P(Y_i = 1) = \pi_i = \frac{1}{1 + e^{-\left(\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}\right)}} \qquad (3.2)$$

tal que $\beta_o, \beta_1, ..., \beta_p$ são os coeficientes de regressão logística a serem estimados pelo método da máxima verossimilhança. Apenas para facilitar a notação, a probabilidade de sucesso em 3.2 será definida como $\pi_i = F\left(x_i^t \beta\right)$ onde F é a função logística e $\underset{\sim}{\beta}^t = (\beta_o, \beta_1, ..., \beta_p)$ é o vetor de coeficientes logísticos.

Para n observações independentes, onde m observações $(m < n)$ apresentam $Y = 0$, enquanto que as $m - n$ observações restantes têm $Y = 1$, a função de verossimilhança é dada por:

$$L = P(Y_1 = 0) \cdot \ldots \cdot P(Y_m = 0) \cdot P(Y_{m+1} = 1) \cdot \ldots \cdot P(Y_n = 1) \qquad (3.3)$$

Usando a notação em 3.2 obtém-se:

$$L = [1 - \pi(x_1)] \cdot \ldots \cdot [1 - \pi(x_m)] \pi(x_{m+1}) \cdot \ldots \cdot \pi(x_n) = \prod_{j=1}^{m} [1 - \pi(x_j)] \cdot \prod_{j=m+1}^{n} \pi(x_j) \qquad (3.4)$$

Simplificando a equação 3.4, tem-se:

$$L = \prod_{j=1}^{n} \left\{ \pi(x_j)^{Y_j} \cdot [1 - \pi(x_j)]^{1-Y_j} \right\} = \prod_{j=1}^{n} \left\{ F(x_j^t \underset{\sim}{\beta})^{Y_i} \cdot \left[1 - F(x_j^t \underset{\sim}{\beta}) \right]^{1-Y_i} \right\} \qquad (3.5)$$

Aplicando uma transformação logarítmica em 3.5 obtém-se:

$$\ln \mathrm{L} = \sum_{\mathrm{j=1}}^{\mathrm{n}} \left\{ \mathrm{Y_i} \ln\left[F\left(x_j^t \underset{\sim}{\beta} \right) \right] + \left(1 - Y_j\right) \ln\left[1 - F\left(x_j^t \underset{\sim}{\beta} \right) \right] \right\} \tag{3.6}$$

O estimador de máxima verossimilhança ou $\hat{\beta}_{MV}$ é o vetor de coeficientes que maximiza a função de verossimilhança em 3.5.

Apesar da função de verossimilhança da regressão logística apresentar apenas um ponto de máximo, as equações de máxima verossimilhança são não lineares e não existe uma fórmula fechada para os estimadores, por isso as estimativas são obtidas numericamente através de algoritmos de otimização, por exemplo, pelo método de Escore de Fisher.

A estimativa assintótica da matriz de covariâncias dos coeficientes é dada por:

$$\hat{\Sigma}\left(\hat{\beta}\right) = \hat{I}^{-1}\left(\hat{\beta}\right) \tag{3.7}$$

sendo $\hat{I}^{-1}\left(\hat{\beta}\right)$ a inversa da matriz $\hat{I}\left(\hat{\beta}\right)$, calculada no ponto $\hat{\beta}$ e dada por :

$$\hat{I}\left(\hat{\beta}\right) = X^T V X \tag{3.8}$$

na qual

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & & x_{2p} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}_{nxp} \tag{3.9}$$

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}_{nxn} \quad (3.10)$$

A estimativa da variância de $\hat{\beta}_j$, denotada por $\hat{\sigma}^2(\hat{\beta}_j)$, é o elemento localizado na linha j e coluna j da matriz $\hat{\Sigma}(\hat{\beta})$ (Silva, 1992), e a covariância entre as estimativas dos coeficientes $\hat{\beta}_i$ e $\hat{\beta}_j$ é o elemento localizado na linha i e coluna j da matriz $\hat{\Sigma}(\hat{\beta})$.

Após estimar o modelo é necessário avaliar a significância dos coeficientes estimados e interpretar os respectivos valores. O estimador $\hat{\beta}_{MV}$ tem distribuição assintótica normal, o que permite a realização da inferência estatística da forma habitual.

3.1. RAZÃO DE VANTAGENS

Além da utilização da probabilidade de ocorrência (π) para expressar as chances de ocorrência um evento, é comum também expressar a vantagem em favor da ocorrência através da razão de probabilidades, definida como (Souza, 2006):

$$Vantagem(odds) = \pi/(1-\pi) \quad (3.11)$$

Por exemplo, em uma conta na qual a probabilidade de estar inativa nos próximos seis meses é 0,67, a vantagem em favor da ocorrência do evento é dada por 0,67 / (1-0,67) = 2, o que significa dizer que na mesma população é duas vezes mais provável encontrar uma conta inativa nos próximos seis meses do que uma conta ativa.

Se a não ocorrência de um evento for mais provável que sua ocorrência, a razão de probabilidades (*odds*) do evento é menor que um. Por outro lado, se a ocorrência de um evento for mais provável que a não ocorrência, a razão de probabilidades do evento

é maior que 1, por exemplo, se as chances de um evento ocorrer são de 3 para 1, a razão de probabilidades do evento é igual a 3.

A partir da probabilidade condicional em 3.2 é fácil definir a razão de probabilidades como uma função das variáveis explicativas dada por:

$$\frac{\pi(x_i)}{1-\pi(x_i)} = e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}} \tag{3.12}$$

Aplicando uma transformação logarítmica na equação 3.12, tem-se uma expressão onde o logaritmo da razão de probabilidades é expresso como uma função linear das variáveis explicativas.

$$\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} \tag{3.13}$$

3.2. MEDIDAS DE QUALIDADE DE AJUSTE

Para estruturar e selecionar o melhor modelo logístico utiliza-se os testes estatísticos que serão descritos a seguir.

a. Estatística de Wald

Após a estimação dos parâmetros deve-se proceder à investigação da significância estatística dos mesmos. O teste de Wald é utilizado para avaliar se o parâmetro é estatisticamente significativo, ou seja, testar (Grizzle *et al*, 1969):

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

A estatística teste utilizada é obtida através da razão do quadrado do coeficiente pela sua respectiva variância. Sob H_o, $W \sim \chi_1^2$, sendo seu valor comparado com valores tabulados. A estatística teste, para avaliar se o parâmetro β é igual a zero, é assim especificada:

$$W = \frac{\hat{\beta}^2}{Var(\hat{\beta})} \qquad (3.14)$$

O teste de Wald, todavia, freqüentemente, falha em rejeitar coeficientes que são estatisticamente significativos (Hauck e Donner, 1977). Sendo assim, aconselha-se que os coeficientes, identificados pelo teste de Wald como sendo estatisticamente não significativos, sejam testados pelo teste da razão de verossimilhanças.

b. Razão de Verossimilhanças

Este teste é utilizado para comparar o modelo de interesse[1] com um outro modelo, em geral, mais completo através das verossimilhanças, ou seja, testar as hipóteses:

1 O modelo de interesse é aquele no qual $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ com $p < N$, que se deseja testar/avaliar.

$$\begin{cases} H_0: \text{O modelo de interesse é tão adequado quanto o saturado.} \\ H_1: \text{O modelo de interesse não é tão adequado quanto o saturado.} \end{cases}$$

O Teste da Razão de Verossimilhanças é baseado na estatística conhecida por *Deviance* do modelo. A *deviance* (função desvio) de um modelo de regressão compara o logaritmo da verossimilhança deste modelo com o logaritmo da verossimilhança do modelo saturado[2]. Um modelo completo é um modelo que se ajusta completamente aos dados, isto é, para cada observação tem-se um parâmetro.

A *deviance*, para o modelo de regressão logístico, é dada por (Abreu, 2004):

$$D = 2[(\ln(\hat{\beta}_{\max}; y) - \ln(\hat{\beta}; y))] \qquad (3.15)$$

Onde, $\hat{\beta}_{\max}$ é o vetor de estimativas de máxima verossimilhança correspondente ao modelo saturado e $\hat{\beta}$ é o vetor de estimativas para o modelo de interesse.

Sob a hipótese nula, ou seja, o modelo de interesse é tão adequado quanto o saturado, a estatística $D \sim \chi^2_{N-p}$.

Se o modelo de interesse não for adequado, a estatística D será maior do que o esperado para uma distribuição χ_{N-p}. Se o modelo é adequado, é de se esperar que o valor D esteja próximo do "meio" da distribuição.

3.3. MÉTODOS DE SELEÇÃO DE VARIÁVEIS

Um dos maiores problemas na estimação de um modelo de regressão é a seleção das variáveis independentes que serão incluídas no modelo (Neter e Wasserman, 1974). Deve-se escolher as melhores variáveis independentes para predizer a variável dependente.

2 O modelo saturado é aquele no qual o número de parâmetros é igual ao número de observações.

Muitas vezes, existe um número grande de variáveis independentes para predizer a variável dependente. Certamente, muitas delas contribuam muito pouco ou até mesmo nada para estimar a variável dependente. A melhor predição será dada pela escolha correta das variáveis independentes, porém não se sabe quantas e quais selecionar.

Um grande número de variáveis independentes pode custar caro e ser de difícil avaliação. Além disso, muitas variáveis correlacionadas no modelo podem adicionar pouco poder de predição, enquanto diminui sua habilidade preditiva e aumenta os erros. Deve-se então reduzir as variáveis independentes de forma a obter a melhor seleção para um modelo parcimonioso.

Existem vários métodos de seleção de variáveis, mas nenhum deles é comprovadamente o melhor método. A escolha das variáveis possui julgamentos muitas vezes subjetivos, não existindo assim, um conjunto ótimo de variáveis. Os procedimentos mais conhecidos são todas as regressões possíveis, *backward*, *forward* e *stepwise*.

O método de todas as regressões possíveis consiste em ajustar todas as combinações possíveis entre as variáveis independentes, comparando todos os modelos ajustados. Para a comparação, utiliza-se o coeficiente de explicação ou quadrado médio dos resíduos. Este método seria inviável quando existe um grande número de variáveis independentes, pois a quantidade necessária de modelos a serem ajustados seria impraticável.

Stepwise é o método mais utilizado na seleção de variáveis. Ele computa uma sequência de equações de regressão, adicionando ou excluindo uma variável independente a cada passo, dependendo da contribuição da variável para o modelo. Uma limitação do método stepwise, é que ele algumas vezes surge com um conjunto de variáveis independentes razoavelmente fraco para predições, quando as variáveis independentes estão altamente correlacionadas (Draper e Smith, 1981).

Já o método *forward*, é uma simplificação do *stepwise*. A diferença é que no método *forward* as variáveis que já foram selecionadas para entrar no modelo não serão mais testadas para sair dele. Uma variável por vez é incorporada até que não haja mais inclusão, e as variáveis selecionadas definam o modelo.

O método de eliminação *backward* é o oposto do método *forward*. Ele começa com um modelo completo contendo todas as variáveis e testa se alguma variável deverá sair do modelo.

3.4. MULTICOLINEARIDADE

Em modelos de regressão com mais de uma variável explicativa, é comum que as variáveis explicativas apresentem algum tipo de interdependência. A multicolinearidade ocorre quando uma ou mais variáveis explicativas do modelo possuem relações exatas ou aproximadamente exatas, ou seja, um dos vetores é combinação linear do outro.

Apenas a correlação entre variáveis independentes causa problemas, pois a forte relação de cada variável explicativa com a variável dependente é desejável.

Quando as variáveis explicativas são perfeitamente correlacionadas entre si, a multicolinearidade é perfeita, impossibilitando o cálculo do estimador de mínimos quadrados, pois a matriz X'X tem o determinante igual a zero, portanto não possui inversa.

A multicolinearidade pode afetar as estimativas dos coeficientes de regressão e a aplicabilidade geral do modelo estimado.

Alguns indícios da presença de multicolinearidade:

- Altos valores do coeficiente de correlação;
- Alterações nas estimativas dos coeficientes de regressão;
- Obtenção de intervalos de confiança com elevadas amplitudes para os coeficientes de regressão, associados a variáveis independentes importantes;
- A rejeição da hipótese Ho: $\beta_1 = \beta_2 = ... = \beta_k = 0$ por meio da realização do teste F, mas nenhuma rejeição das hipóteses Ho: $\beta_i = 0$, i = 1, 2,..., k, por meio da realização dos testes t sobre os coeficientes individuais de regresssão;

- Obtenção de estimativas para os coeficientes de regressão com sinais algébricos contrários àqueles que seriam esperados a partir de conhecimentos teóricos disponíveis ou de experiências anteriores sobre o fenômeno estudado.

4. ESTUDO DE CASO

A metodologia adotada no presente trabalho está estruturada no desenvolvimento das seguintes etapas:

- Definição da base de dados;
- Análise exploratória da base de dados;
- Aplicação do modelo de regressão logística na base de dados, obtendo:
    - O ranqueamento dos clientes;
    - A classificação dos clientes em classes ordenadas.
- Análise dos resultados.

A seguir é apresentada a aplicação de cada uma destas etapas a uma situação real de classificação de clientes que possuem um determinado cartão de crédito de uma instituição financeira brasileira.

4.1. DEFINIÇÃO DA BASE DE DADOS

A base de dados utilizada neste estudo contempla clientes pessoa física que possuam cartão de crédito de uma grande instituição financeira brasileira. A utilização deste cartão de crédito é exclusiva em postos de combustíveis de uma determinada bandeira e em suas lojas de conveniência.

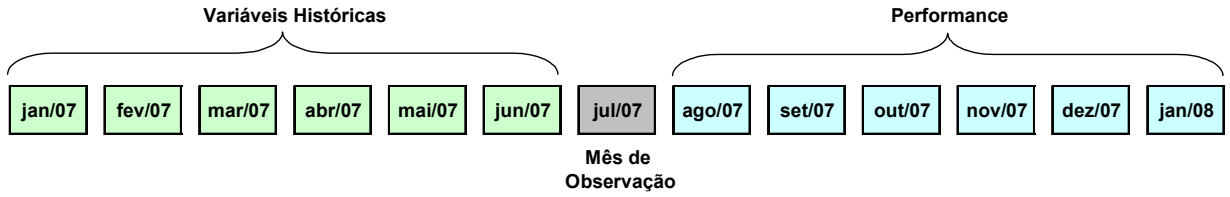
Para a estimação do modelo selecionam-se todas as contas ativas e sem atrasos na data de observação. A base de dados possui cerca de 110 mil contas com data de observação em Julho de 2007. Somente consideram-se neste estudo as contas que possuem as seguintes características:

- Possuir cartão de crédito ativo em Julho de 2007
- Ter pelo menos 6 meses de ativação em Julho de 2007
- Sem atraso de pagamento em Julho de 2007

4.1.1. Janela de Observação

Analisa-se a utilização das contas do cartão de crédito nos seis meses anteriores à data de observação, de Janeiro de 2007 até Junho de 2007, para a construção das variáveis históricas de utilização do produto que são utilizadas como variáveis explicativas no modelo. A variável resposta que contém o status da conta de cartão de crédito nos próximos seis meses é criada com base no mês de Janeiro de 2008, conforme apresentado na figura 3.

Figura 3 - Janela de observação da Base de Dados



Fonte: o Autor

Foi realizado um estudo preliminar englobando toda a base de contas deste cartão de crédito em diversos períodos e não foi observada sazonalidade nos dados de utilização do cartão. Sendo assim, para não utilizar uma base de dados muito antiga, utiliza-se neste estudo uma base de dados que contém os seis meses anteriores ao

mês de observação para estimar a probabilidade de inatividade nos seis meses posteriores.

4.1.2. Variáveis Selecionadas

O banco de dados possui 17 variáveis comportamentais. A tabela 2 apresentada a seguir ilustra as variáveis disponíveis no banco de dados: as variáveis Número da conta, Número do CPF, Datas de Abertura e Data de Ativação não foram utilizadas no estudo.

Tabela 2 - Descrição das Variáveis Utilizadas no Estudo

Variável	Código	Descrição
Número da conta	NO_CONTA	Número da conta a qual o cartão está associado. Uma conta poderá ter vários cartões associados a ela.
Número do CPF	NO_CPF	Número do CPF do cliente
Dígitos do CPF	CD_CONTROLE_CNPJ	Número de controle do CPF
Data de Abertura	DT_ABERTURA_CONTA	Data de abertura da conta
Data de Ativação	DT_ATIVACAO_CONTA	Data mais antiga entre as datas de ativação dos cartões (titular e adicional)
Total de Transações	TOT_TRANSACOES	Total de transações realizadas no período
Número de Meses de Utilização	NO_MESES_UTILIZA	Número de meses em que pelo menos um cartão da conta foi utilizado
Total de Saques	TOT_SAQUE	Total de saques realizados no período
Total de Compras à Vista	TOT_COMPRA_VISTA	Total de compras à vista realizadas no período
Total de Compras Parceladas	TOT_COMPRA_PARC	Total de compras parceladas realizadas no período
Valor Total das Faturas	VL_TOT_FATURA	Valor total das faturas no período
Valor Médio das Faturas	VL_MED_FATURA	Valor médio das faturas no período
Percentual Médio de Utilização	PE_UTILIZACAO_MEDIO	Percentual médio de utilização do limite no período
Tipo de Transações	TIPO_TRANSACAO	Indica os tipos de transações efetuadas no período
Tempo de Abertura	TMP_ABERTURA_CONTA	Tempo em dias desde a abertura da conta
Tempo de Ativação	TMP_ATIVACAO_CONTA	Tempo em dias desde a ativacao da conta
Utilização	UTILIZA	Variável resposta (Indica se a conta foi utilizada nos seis meses posteriores à data de observação)

Fonte: o Autor

4.2. ANÁLISE EXPLORATÓRIA DA BASE DE DADOS

Uma análise exploratória dos dados foi realizada para obter um maior conhecimento acerca do comportamento das contas selecionadas para o estudo.

A tabela 3 a seguir mostra algumas estatísticas descritivas das variáveis números da base de dados descrita.

Tabela 3 - Estatísticas Descritivas das Contas

Variável	Mínimo	1º Quartil	Média	Mediana	3º Quartil	Máximo	Desvio Padrão	Coef. de Variação
Total de Transações	1,00	6,00	14,88	11,00	20,00	262,00	14,09	0,95
Total de Saques	0,00	0,00	0,07	0,00	0,00	11,00	0,31	4,62
Total de Compras Parceladas	0,00	0,00	0,02	0,00	0,00	34,00	0,33	18,14
Total de Compras à Vista	0,00	5,00	14,80	11,00	20,00	262,00	14,10	0,95
Número de Meses de Utilização	1,00	3,00	4,54	5,00	6,00	6,00	1,65	0,36
Valor Médio das Faturas	0,00	81,99	192,17	143,87	248,71	5.394,96	169,65	0,88
Valor Total das Faturas	0,00	341,99	1.033,22	730,93	1.388,71	32.369,76	1.035,12	1,00
Percentual Médio de Utilização	0,00	0,15	0,33	0,28	0,47	1,56	0,22	0,68
Tempo de Abertura	185,00	294,00	409,48	378,00	538,00	725,00	142,00	0,35
Tempo de Ativação	1,00	203,00	323,24	298,00	451,00	716,00	169,55	0,52

Fonte: o Autor

Em média, as contas estudadas possuem pouco mais de 1 ano e no mínimo 6 meses desde sua abertura. O tempo desde a ativação, ou seja, quando o cartão foi utilizado pela primeira vez, é em média de 11 meses e no mínimo de 1 dia de acordo as premissas utilizadas para a construção da base de dados. O percentual de utilização médio do limite de crédito disponível é de 33%. Em alguns casos o limite de crédito é excedido em até 56% conforme mostrado na tabela 3.

Nos seis meses de utilização que foram observados na base, os clientes em média utilizam o cartão em aproximandamente 5 destes meses. Os clientes realizam em média 15 transações em 6 meses, alguns chegam a realizar mais de 200 transações em apenas 6 meses. Se paga em média R$192,17 por fatura. Percebe-se

que a utilização deste cartão é maior quando comparado aos valores de outros cartões do mercado, conforme ilustrado na tabela 4.

Tabela 4 – Indicadores do Mercado de Cartões de Crédito

Indicadores	jan/08	fev/08	mar/08	abr/08	mai/08	jun/08	jul/08	ago/08	set/08
Transações por Cartão	1,0	1,0	1,1	1,0	1,1	1,1	1,1	1,1	1,1
Gasto Médio por Cartão R$	64	60	65	63	69	66	68	70	67

Fonte: ABECS (2008)

A maioria das transações efetuadas é de compra à vista, sendo que cerca de 94% das contas utilizam somente este tipo de operação. As contas que realizam saques e compras à vista totalizam 5%, como pode-se observar na tabela 5 abaixo. As contas que só efetuam saque ou realizam apenas compras parceladas possuem um maior percentual de contas inativas nos seis meses posteriores ao ponto de corte da observação.

Tabela 5 - Contas por Tipo de Transação e Utiliza

Tipo de Transação	Conta Inativa	Conta Ativa	Total
01 - Somente Saque	90%	10%	1%
02 - Somente Compra à vista	28%	72%	94%
03 - Somente Compra Parcelada	68%	32%	0%
04 - Saque e Compra à vista	32%	68%	5%
06 - À vista e Parcelado	19%	81%	1%
07 - Saque, Compra à vista, Parcelado	29%	71%	0%
Total	29%	71%	100%

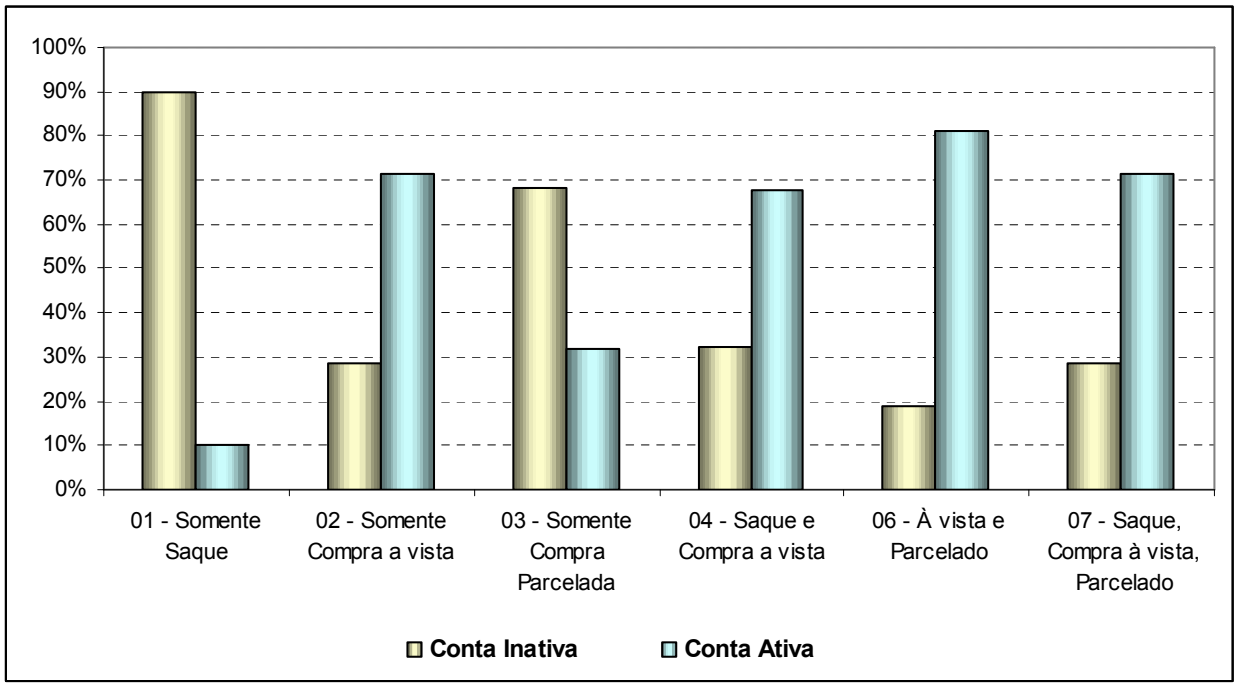
Fonte: o Autor

Pode-se observar ainda na tabela 5 que o percentual de contas que se tornam inativas nos seis meses posteriores ao ponto de corte da observação é de 29%, um valor razoavelmente alto, justificando mais uma vez o objetivo deste estudo.

O gráfico 4 a seguir mostra a distribuição do tipo de transação pela variável resposta utiliza, que é utilizada como variável dependente no ajuste do modelo. Cerca de 90% das contas que realizam somente saque se tornam inativas nos seis meses seguintes ao ponto de corte da observação, e das contas que realizam compras à vista

e parcelas 81% continuam ativas nos próximos seis meses posteriores ao ponto de corte da bservação.

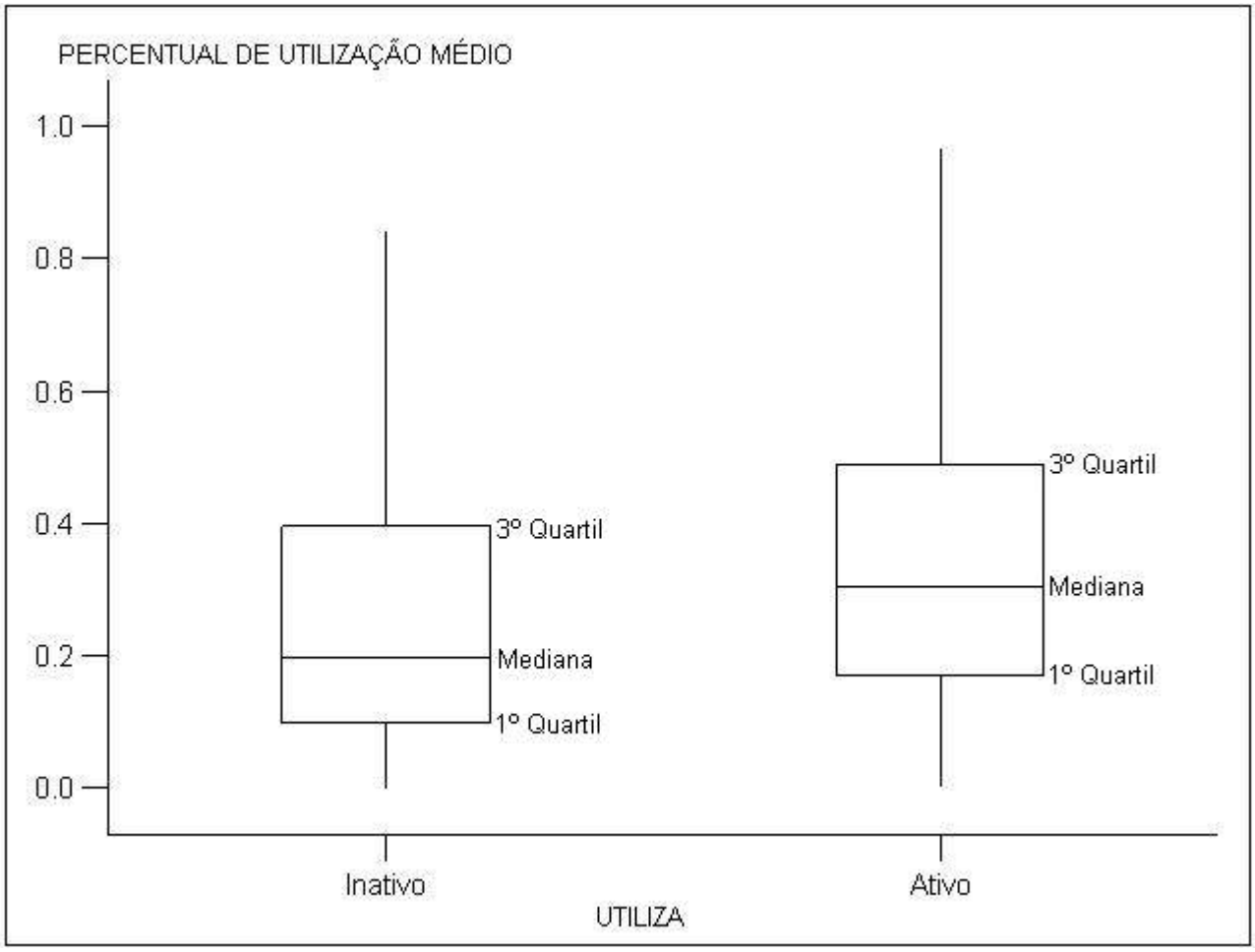
Gráfico 4 - Contas por Tipo de Transação e Utiliza



Fonte: o Autor

O gráfico 5 apresenta os quartis (inclusive a mediana) do percentual de utilização médio dos cartões nos seis meses posteriores ao ponto de corte da observação. Observa-se, neste gráfico, que as contas ativas possuem a mediana do percentual médio de utilização do limite, 31%, um pouco maior do que e mediana, 20%, das contas que ficam inativas. Além disto, ocorre maior concentração de contas com percentual de utilização acima de 30% dentre as contas ativas do que nas contas inativas.

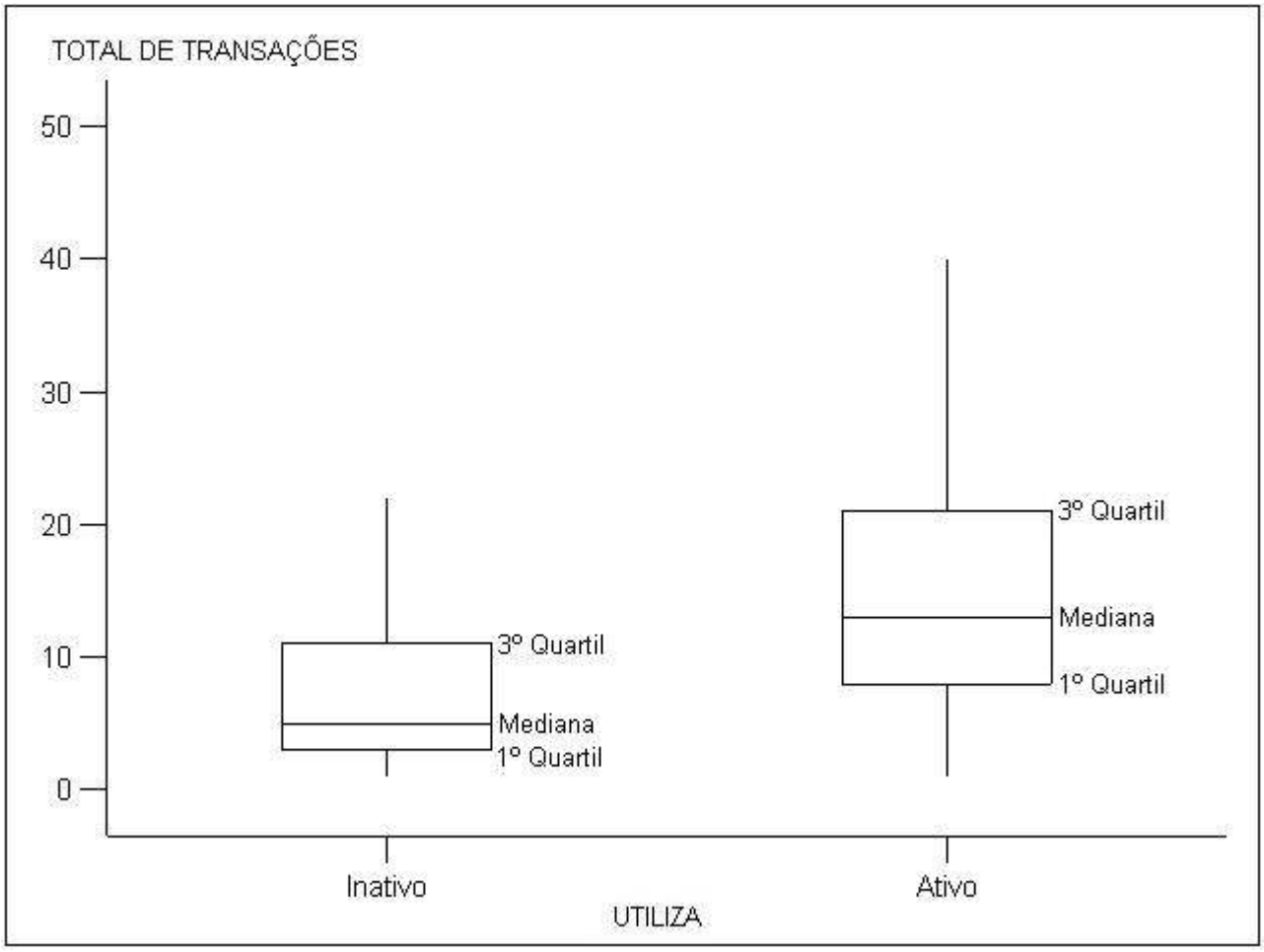
Gráfico 5 - Contas por Percentual Médio de Utilização e Utiliza



Fonte: o Autor

De acordo com o gráfico 6 apresentado a seguir, a mediana do total de transações é maior nas contas ativas comparadas às contas inativas. A mediana do total de transações para as contas ativas é de 5 transações em seis meses e, para as contas inativas é de 14 transações em seis meses.

Gráfico 6 - Contas por Total de Transações e Utiliza



Fonte: o Autor

4.3. APLICAÇÃO E AJUSTAMENTO DO MODELO

Nesta parte do trabalho, define-se o modelo estatístico que será utilizado para descrever a relação entre a probabilidade da conta deixa de utilizar o cartão de crédito e suas características.

Conforme descrito no capítulo 2, no contexto da modelagem por regressão logística, é efetuado um ajustamento de modelos - o qual consite na construção e análise de vários modelos até que se encontre aquele que melhor se ajuste ao problema estudado.

Nesta seção é efetuada passo a passo a construção e análise dos modelos intermediários construídos. No total, foram construídos 22 modelos, enumerados de Ai até I. Para os testes de cada um dos modelos utiliza-se o pacote estatístico SAS sobre a base de dados descrita na seção 4.1.

- **Modelos Intermediários Ai e B:**

Para avaliar quais variáveis são significativas no modelo foi primeiramente ajustado um modelo de regressão logística considerando-se cada variável explicativa separadamente, o modelo Ai referente à variável i, e um modelo B com todas as variáveis simultaneamente. Nesta primeira etapa, retiram-se do modelo as variáveis que não são significativas sozinhas e nem em conjunto. Será utilizada a estatística da razão de verossimilhanças, nas respectivas decomposições denominadas do tipo 1 e tipo 3, para comparar os diversos modelos de interesse.

A decomposição do tipo 1[3] refere-se à comparação de modelos tal que a diferença entre o modelo simplificado da hipótese nula e o considerado na hipótese alternativa é de apenas uma variável, sendo que os modelos são ajustados seqüencialmente a partir daquele com apenas o intercepto até o modelo com todas as variáveis explicativas em estudo (procedimento "*forward*").

Por outro lado, a decomposição de tipo 3[4] considera na hipótese alternativa sempre o modelo completo, com todas as variáveis explicativas em estudo, e na hipótese nula um modelo com todas as variáveis exceto aquela que se deseja testar. Neste caso, retira-se uma variável de cada vez do modelo completo.

Por exemplo, em um estudo com as variáveis X1, X2, X3 e X4, sendo todas as variáveis contínuas, no qual se deseja estudar a significância da variável X2 usando a estatística da razão de verossimilhanças pela decomposição de tipo 1, testa-se as hipóteses:

3 LR TYPE1 - Consiste em ajustar modelos seqüencialmente começando com somente o intercepto e continuando colocando variáveis até o modelo especificado.

4 LR TYPE3 - Estatísticas do tipo 3 são computadas para cada efeito especificado no modelo.

$$\begin{cases} H_0 : \ln\left(\dfrac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_1 \\ H_1 : \ln\left(\dfrac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 \end{cases}$$

Enquanto que, utilizando-se a estatística da razão de verossimilhanças pela decomposição de tipo 3, testa-se as hipóteses:

$$\begin{cases} H_0 : \ln\left(\dfrac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 \\ H_1 : \ln\left(\dfrac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \end{cases}$$

Antes de iniciar o ajustamento dos modelos, será calculada a correlação entre as variáveis explicativas para saber se existe alguma variável altamente correlacionada com outra, o que causaria problema de multicolinearidade.

Tabela 6 - Correlação das Variáveis Explicativas

Variáveis	Total de Transações	Núm de Meses de Utilização	Total de Saques	Total de Compras à Vista	Total de Compras Parceladas	Tempo de Abertura	Valor Total das Faturas	Valor Médio das Faturas	Perc. de Utilização Médio	Tempo de Ativação
Total de Transações	100%	57%	-3%	100%	2%	10%	54%	48%	37%	22%
Núm de Meses de Utilização		100%	-8%	57%	2%	9%	47%	34%	33%	36%
Total de Saques			100%	-6%	0%	2%	0%	0%	-6%	-2%
Total de Compras à Vista				100%	0%	10%	54%	48%	37%	22%
Total de Compras Parceladas					100%	1%	3%	3%	1%	0%
Tempo de Abertura						100%	13%	11%	-9%	79%
Valor Total das Faturas							100%	97%	56%	24%
Valor Médio da Faturas								100%	57%	17%
Percentual de Utilização Médio									100%	5%
Tempo de Ativação										100%

Fonte: o Autor

A variável Total de Compras à Vista possui correlação igual a 100% com a variável Total de Transações, o que impossibilita o cálculo dos estimadores de mínimos quadrados, pois apresenta um problema de multicolinearidade perfeita. Assim, a variável Total de Compras à Vista não será considerada no modelo. Outras variáveis possuem correlação elevada, como é o caso de Tempo de Ativação com Tempo de Abertura e Valor Total das Faturas e Valor Médio das Faturas. Estes resultados já eram esperados, uma vez que a maioria das contas faz compras à vista e a data de abertura da conta é sempre muito próxima, quando não igual, à data de ativação da conta.

Na tabela 7 a seguir, apresenta-se a correlação entre as variáveis explicativas e a variável resposta para a escolha da variável correlacionada que sairá do modelo além da ordem de entrada das variáveis explicativas no método *forward*.

Tabela 7 - Correlação das Variáveis Explicativas com a Variável Resposta

Variáveis Explicativas	**Correlação**
Núm de Meses de Utilização	43,5%
Total de Compras à Vista	29,5%
Total de Transações	29,4%
Valor Total das Faturas	22,4%
Valor Médio das Faturas	18,3%
Perc. de Utilização Médio	15,3%
Tempo de Ativação	14,4%
Tempo de Abertura	6,3%
Total de Saques	-5,8%
Total de Compras Parceladas	1,3%

As variáveis correlacionadas que não entrarão no modelo são: Total de Transações, Valor Médio das faturas e Tempo de Abertura da Conta.

Neste primeiro momento, foram gerados nove modelos, um contendo todas as variáveis explicativas (Modelo B) e mais oito modelos (Modelos Ai) contendo cada um uma única variável explicativa.

A seguir a tabela 8 que contém os resultados dos modelos Ai e B.

Tabela 8 - Decomposição do Tipo 1 e 3 dos Modelos Ai e B

Variável	Modelo B Tipo 1	Modelo B Tipo 3	Modelo Ai Tipo 1	*Deviance* Tipo 3
Número de Meses de Utilização	<,0001	<,0001	<,0001	4563,84
Total de Compras à Vista	<,0001	<,0001	<,0001	946,58
Valor Total da Fatura	0,0107	0,286	<,0001	1,14
Percentual de Utilização Médio	<,0001	<,0001	<,0001	35,47
Tempo de Ativação da Conta	0,1334	0,1645	<,0001	1,93
Total de Saques	<,0001	0,0182	<,0001	5,58
Total de Compras Parceladas	0,0458	0,3332	<,0001	0,94
Tipo de Transação	<,0001	<,0001	<,0001	237,78

Fonte: o Autor

Verifica-se que a variável Total de Compras Parceladas não é significativa no modelo B, conforme a tabela 8, o que é explicado pela baixa frequência de utilização de compras parceladas na base de dados. Assim, esta variável não participará da próxima etapa da modelagem, onde será ajustado um modelo C contendo todas as variáveis explicativas exceto esta variável.

- **Modelos Intermediários Ai e C:**

Um modelo C foi ajustado considerando todas as variáveis menos a variável não significativa observada anteriormente. Na tabela 9 a seguir, verifica-se a decomposição do tipo 1 e do tipo 3 do modelo C ajustado e novamente dos modelos Ai.

Tabela 9 - Decomposição do Tipo 1 e 3 do Modelo C

Variável	Modelo C Tipo 1	Modelo C Tipo 3	Modelo Ai Tipo 1	*Deviance* Tipo 3
Número de Meses de Utilização	<,0001	<,0001	<,0001	4569,91
Total de Compras à Vista	<,0001	<,0001	<,0001	945,64
Valor Total da Fatura	0,0107	0,2744	<,0001	1,19
Percentual de Utilização Médio	<,0001	<,0001	<,0001	35,6
Tempo de Ativação da Conta	0,1334	0,163	<,0001	1,95
Total de Saques	<,0001	0,0181	<,0001	5,59
Tipo de Transação	<,0001	<,0001	<,0001	240,83

Fonte: o Autor

Percebe-se que após a retirada da variável Total de Compras Parceladas, ainda existe uma variável não significativa no modelo C. Logo, na próxima etapa da modelagem será ajustado um modelo D onde se deve retirar a variável Tempo de Ativação da Conta que não está significativa.

- **Modelos Intermediários Ai e D:**

Assim, a tabela 10 apresenta o modelo D que foi ajustado contendo todas as variáveis do modelo C menos aquela considerada não significativa conforme descrito na tabela 8, Tempo de Ativação da Conta.

Tabela 10 - Decomposição do Tipo 1 e 3 do Modelo D

Variável	Modelo D Tipo 1	Modelo D Tipo 3	Modelo Ai Tipo 1	*Deviance* Tipo 3
Número de Meses de Utilização	<,0001	<,0001	<,0001	5836,16
Total de Compras à Vista	<,0001	<,0001	<,0001	913,73
Valor Total da Fatura	0,2297	0,0544	<,0001	3,7
Percentual de Utilização Médio	<,0001	<,0001	<,0001	32,75
Total de Saques	<,0001	0,0238	<,0001	5,11
Tipo de Transação	<,0001	<,0001	<,0001	207,47

Fonte: o Autor

Após a retirada da variável Tempo de Ativação da Conta, ainda existe uma variável não significativa no modelo D. Logo, no próximo modelo não será considerada a variável Valor Total da Fatura que não está significativa.

- **Modelos Intermediários Ai e E:**

A tabela 11 apresenta o modelo E que foi ajustado contendo todas as variáveis do modelo D menos aquela considerada não significativa conforme descrito na tabela 10.

Tabela 11 - Decomposição do Tipo 1 e 3 do Modelo E

Variável	Modelo E Tipo 1	Modelo E Tipo 3	Modelo Ai Tipo 1	*Deviance* Tipo 3
Número de Meses de Utilização	<,0001	<,0001	<,0001	6030,01
Total de Compras à Vista	<,0001	<,0001	<,0001	1050,62
Percentual de Utilização Médio	<,0001	<,0001	<,0001	29,72
Total de Saques	<,0001	0,0204	<,0001	5,38
Tipo de Transação	<,0001	<,0001	<,0001	206,94

Fonte: o Autor

Pode-se observar que todas as variáveis são significativas no modelo E e nos modelos Ai. Agora, deve-se analisar os parâmetros estimados para observar se há a necessidade de realizar agregações nos níveis dos fatores.

Há necessidade de agregações nos níveis dos fatores toda vez que houver categorias com parâmetros não significativos em fatores com mais de duas categorias, ou quando em um fator houver categorias cujos valores dos parâmetros estimados para as diferentes categorias sejam muito próximos.

Para agregar as categorias não significativas com o fator utilizado como categoria de referência, utiliza-se o teste de Wald para testar $H_0 : \beta_j = 0$ contra $H_1 : \beta_j \neq 0$. Analisa-se o p-valor do teste para decidir sobre a aceitação da hipótese nula e assim, agregar a categoria que está sendo avaliada com a categoria de referência.

Para agregar fatores com parâmetros estimados muito parecidos, utiliza-se o teste do contraste. Este teste consiste em testar a igualdade entre dois parâmetros, ou seja, testar as hipóteses $H_0 : \beta_i - \beta_j = 0$ contra $H_1 : \beta_i - \beta_j \neq 0$ para $i \neq j$. Analisa-se o p-valor do teste do contraste para decidir sobre a aceitação da hipótese nula, podendo assim, agregar os fatores testados.

A tabela 12 apresenta os parâmetros estimados e o p-valor do teste de Wald para o modelo E.

Tabela 12 - Estatísticas do Modelo E

Variável	g.l.	Parâmetro Estimado	DP	Qui-Quadrado	Pr > Qui-Quadrado
Intercepto	1	1,91	0,37	26,42	<,0001
Número de Meses de Utilização	1	-0,46	0,01	5.873,66	<,0001
Total de Compras à Vista	1	-0,03	0,00	844,90	<,0001
Percentual de Utilização Médio	1	0,20	0,04	29,77	<,0001
Total de Saques	1	-0,13	0,06	5,22	0,0223
Tipo de Transação - 01 - Somente Saque	1	1,05	0,38	7,62	0,0058
Tipo de Transação - 02 - Somente Compra a Vista	1	-0,53	0,37	2,06	0,1515
Tipo de Transação - 03 - Somente Compra Parcelada	1	-0,38	0,52	0,54	0,4614
Tipo de Transação - 04 - Saque e Compra a vista	1	-0,28	0,37	0,57	0,4487
Tipo de Transação - 06 - À vista e Parcelado	1	-0,77	0,38	4,08	0,0433
Tipo de Transação - 07 - Saque, Compra à vista, Parc.	0	0	0	,	,

Fonte: o Autor

Há necessidade de agregação das seguintes categorias não significativas com as linhas de referência: no fator Tipo de Transação, deve-se agregar 02 (Somente Compra á Vista), 03 (Somente Compra Parcelada), 04 (Saque e Compra à Vista) e 06 (À Vista e Parcelado) com 07 (Saque, Compra à Vista e Parcelada). A agregação será feita com a linha de referência, sendo assim, somente analisando o p-valor do teste de Wald já se pode chegar a esta conclusão.

Ajusta-se então o modelo F agregando-se os níveis dos fatores.

- **Modelo Intermediário F:**

Chega-se ao modelo F contendo todas as variáveis simples significativas e com os níveis dos fatores agregados. Na tabela 13 a seguir, apresenta-se o modelo ajustado F com a agregação nos níveis dos fatores.

Tabela 13 - Estatísticas do Modelo F

Variável	g.l.	Parâmetro Estimado	DP	Qui-Quadrado	Pr > Qui-Quadrado
Intercepto	1	1,3845	0,0214	4193,27	<,0001
Número de Meses de Utilização	1	-0,4588	0,006	5905,62	<,0001
Total de Compras à Vista	1	-0,0313	0,0011	844,16	<,0001
Percentual de Utilização Médio	1	0,1990	0,0364	29,86	<,0001
Total de Saques	1	0,0408	0,0249	2,68	0,1016
Tipo de Transação - 01 - Somente Saque	1	1,3293	0,1107	144,25	<,0001
Tipo de Transação - 02,03,04,05,06 e 07	0	0,0000	0	,	,

Fonte: o Autor

Pode-se observar que com a agregação dos níveis dos fatores, a variável Total de Saques deixou de ser significativa. Assim, será ajustado um modelo G sem esta variável.

- **Modelo Intermediário G:**

O modelo G contém todas as variáveis do modelo F exceto a variável Total de Saques. Na tabela 14 a seguir, apresenta-se o modelo ajustado G.

Tabela 14 - Estatísticas do Modelo G

Variável	g.l.	Parâmetro Estimado	DP	Qui-Quadrado	Pr > Qui-Quadrado
Intercepto	1	1,3870	0,0213	4229,78	<,0001
Número de Meses de Utilização	1	-0,4587	0,006	5903,59	<,0001
Total de Compras à Vista	1	-0,0313	0,0011	845,53	<,0001
Percentual de Utilização Médio	1	0,1978	0,0364	29,51	<,0001
Tipo de Transação - 01 - Somente Saque	1	1,3870	0,1051	174,14	<,0001
Tipo de Transação - 02,03,04,05,06 e 07	0	0,0000	0	,	,

Fonte: o Autor

Com todas as variáveis simples significativas, deve-se testar quais interações duplas são significativas. No modelo G, existem quatro variáveis simples significativas. Deve-se então, gerar todas as combinações possíveis entre estas quatro variáveis simples, logo, combinam-se as quatro variáveis simples formando-se 6 interações duplas.

Assim, ajustam-se 6 modelos, cada um contendo as quatro variáveis simples correspondentes aos efeitos mais uma interação dupla a ser testada. No tabela 14 apresenta-se a significância de todas as 6 interações estudadas.

Tabela 15 - Resultado do Ajustamento das Interações Duplas no Modelo G

Interações Duplas	Tipo 3
Número de Meses de Utilização*Total de Compras à Vista	0,0007
Número de Meses de Utilização*Percentual de Utilização Médio	<,0001
Número de Meses de Utilização*Tipo de Transação	0,0519
Total de Compras à Vista*Percentual de Utilização Médio	<,0001
Total de Compras à Vista*Tipo de Transação	0,0961
Percentual de Utilização Médio*Tipo de Transação	0,1419

Fonte: o Autor

Algumas interações duplas, apesar de significativas quando inseridas no modelo G fazem com que algumas variáveis simples deixem de ser significativas, assim estas interações duplas não serão consideradas na próxima modelagem.

Existem três interações duplas significativas e que não tornam nenhuma variável simples não significativa. Ajusta-se um modelo H que contenha todas as variáveis simples significativas e mais as três variáveis duplas significativas.

- **Modelo Intermediário H:**

Apresenta-se na tabela 16 a seguir a decomposição do tipo 1 e do tipo 3 do modelo H ajustado.

Tabela 16 - Decomposição Tipo 1 e Tipo 3 do Modelo H

Variável	Tipo 1	Tipo 3	*Deviance* Tipo 3
Número de Meses de Utilização	<,0001	<,0001	1429,34
Total de Compras à Vista	<,0001	<,0001	40,07
Percentual de Utilização Médio	<,0001	0,0023	9,32
Tipo de Transação	<,0001	<,0001	228,75
Número de Meses de Utilização*Total de Compras à Vista	0,0007	<,0001	43,13
Número de Meses de Utilização*Percentual de Utilização Médio	<,0001	0,0893	2,89
Total de Compras à Vista*Percentual de Utilização Médio	<,0001	<,0001	198,27

Fonte: o Autor

Será retirada a interação dupla Número de Meses de Utilização e Percentual de Utilização Médio, pois não é significativa no modelo com todas as interações duplas além de deixar algumas variáveis simples não significativas.

- **Modelo Intermediário I:**

O modelo I ajustado a seguir não considera a interação dupla Número de Meses de Utilização e Percentual de Utilização.

Tabela 17 - Decomposição Tipo 1 e Tipo 3 do Modelo I

Variável	Tipo 1	Tipo 3	Deviance Tipo 3
Número de Meses de Utilização	<,0001	<,0001	2859,34
Total de Compras à Vista	<,0001	<,0001	38,43
Percentual de Utilização Médio	<,0001	<,0001	74,9
Tipo de Transação	<,0001	<,0001	226,52
Número de Meses de Utilização*Total de Compras à Vista	0,0007	<,0001	52,11
Total de Compras à Vista*Percentual de Utilização Médio	<,0001	<,0001	249,9

Fonte: o Autor

Todas as variáveis simples e duplas são significativas no modelo I. Assim, na tabela 18 apresentam-se os parâmetros estimados e o p-valor do teste de Wald do modelo I.

Tabela 18 - Resultado do Modelo I

Variável	g.l.	Parâmetro Estimado	DP	Qui-Quadrado	Pr > Qui-Quadrado
Intercepto	1	1,3811	0,03	2499,95	<,0001
Número de Meses de Utilização	1	-0,4014	0,01	2889,27	<,0001
Total de Compras à Vista	1	-0,0276	0,00	37,93	<,0001
Percentual de Utilização Médio	1	-0,4722	0,05	76,19	<,0001
Tipo de Transação - 01 - Somente Saque	1	1,3633	0,11	165,86	<,0001
Tipo de Transação - 02,03,04,05,06 e 07	0	0,0000	0,00	,	,
Número de Meses de Utilização*Tot. de Compras à Vista	1	-0,0058	0,00	52,69	<,0001
Total de Compras à Vista*Percentual de Utilização Médio	1	0,0600	0,00	262,18	<,0001

Fonte: o Autor

O modelo I possui todas as variáveis simples e duplas significativas. Não será necessário testar as interações triplas. Para uma interação tripla ser inserida no modelo é necessário que o modelo possua as três variáveis simples significativas e mais as três interações duplas resultantes das combinações destas variáveis simples. Assim, como o modelo I tem somente duas interações duplas significativas, não existe nenhuma interação dupla a ser testada.

O modelo I possui todas as variáveis simples e duplas significativas, sendo assim, este é o modelo final ajustado, o que melhor se adequa aos dados estudados.

O modelo final pode ser escrito da seguinte forma:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \mu + \beta^{no\_meses\_utiliza} X_i^{no\_meses\_utiliza} + \beta^{tot\_compra\_vista} X_i^{tot\_compra\_vista} +$$

$$\beta^{pe\_utilizacao\_medio} X_i^{pe\_utilizacao\_medio} + \beta_i^{tipo\_transacao} +$$

$$\beta^{no\_meses\_utiliza} * \beta^{tot\_compra\_vista} * X_i^{no\_meses\_utiliza} * X_i^{tot\_compra\_vista} +$$

$$\beta^{tot\_compra\_vista} * \beta^{pe\_utilizacao\_medio} * X_i^{tot\_compra\_vista} * X_i^{tpe\_utilizacao\_medio}$$

4.4. ANÁLISE DO MODELO FINAL

Nesta etapa, o Modelo I (modelo final) foi aplicado à base de dados, com o objetivo de identificar os riscos associados às contas do cartão de crédito estudado neste trabalho. Apresentam-se a seguir os resultados encontrados nesse estudo.

Tabela 19 - Estimativas dos Parâmetros do Modelo Final

Variável	Parâmetro Estimado	P-Valor
Intercepto	1,3811	<,0001
Número de Meses de Utilização	-0,4014	<,0001
Total de Compras à Vista	-0,0276	<,0001
Percentual de Utilização Médio	-0,4722	<,0001
Tipo de Transação		
01 - Somente Saque	1,3633	<,0001
02,03,04,06 e 07 - Demais		
Número de Meses de Utilização*Total de Compras à Vista	-0,0058	<,0001
Total de Compras à Vista*Percentual de Utilização Médio	0,0600	<,0001

Fonte: o Autor

Os resultados obtidos com o modelo final indicam que a probabilidade de a conta deixar de utilizar o cartão é maior para quem:

1. Utiliza o cartão em um menor número de meses;

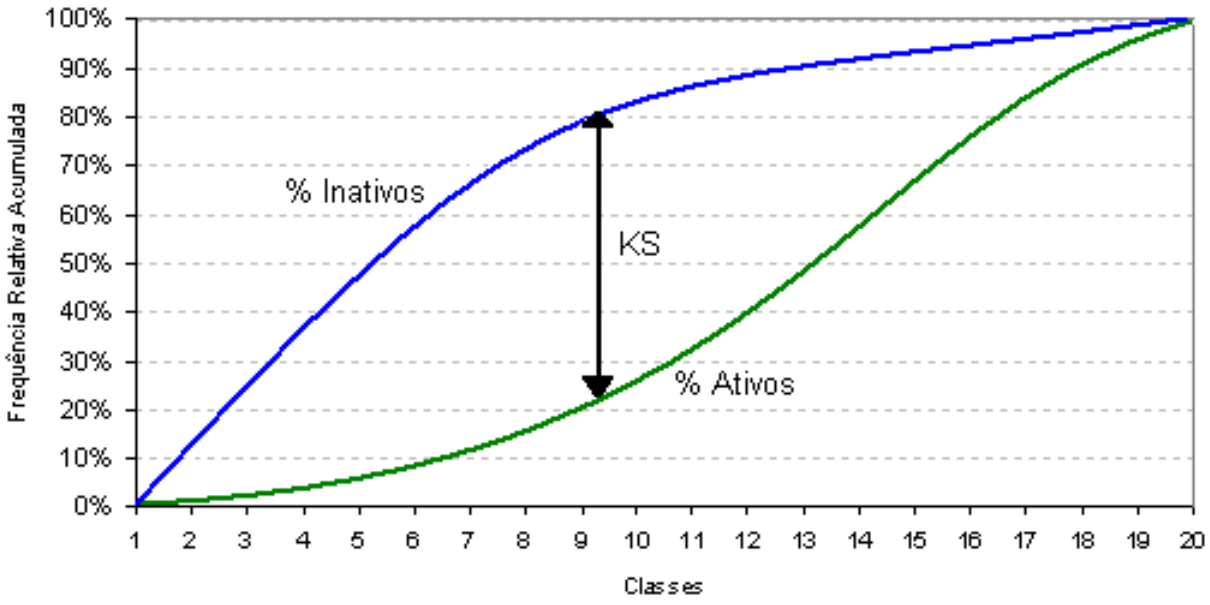
2. Realiza um total de compras à vista menor;

3. Possui um menor percentual de utilização médio do limite;

4. Realiza somente saque.

Utilizou-se a estatística de Kolmogorv–Smirnov (KS) para avaliar o desempenho do modelo apresentado. Esta estatística mede a máxima distância entre a distribuição de frequência relativa acumulada entre os valores observados da variável resposta e os valores estimados, podendo variar de 0 a 1. Quanto mais próximo de 1, melhor é a performance do modelo. Segue ilustração na figura 7 de um exemplo de gráfico de KS.

Figura 7 – Exemplo de Gráfico de KS



Fonte: o Autor

No caso do modelo final estimado o valor do KS encontrado foi 42,58%, indicando uma boa discriminação.

4.4.1. Razão de Vantagens no Modelo Estimado

Nesta seção serão estimadas as razões de vantagens para o modelo final. Esta estatística indica que categoria tem maior ou menor vantagem em favor da ocorrência do usuário deixar de utilizar o cartão de crédito em relação à categoria comparada.

Conforme a tabela 20 a seguir, a categoria Somente Saque do fator Tipo de Transação, apresenta maior vantagem em favor de a conta estar inativa nos seis meses posteriores à data de corte de observação.

Tabela 20 - Razão de Vantagens em Favor de a Conta estar Inativa por Tipo de Transação

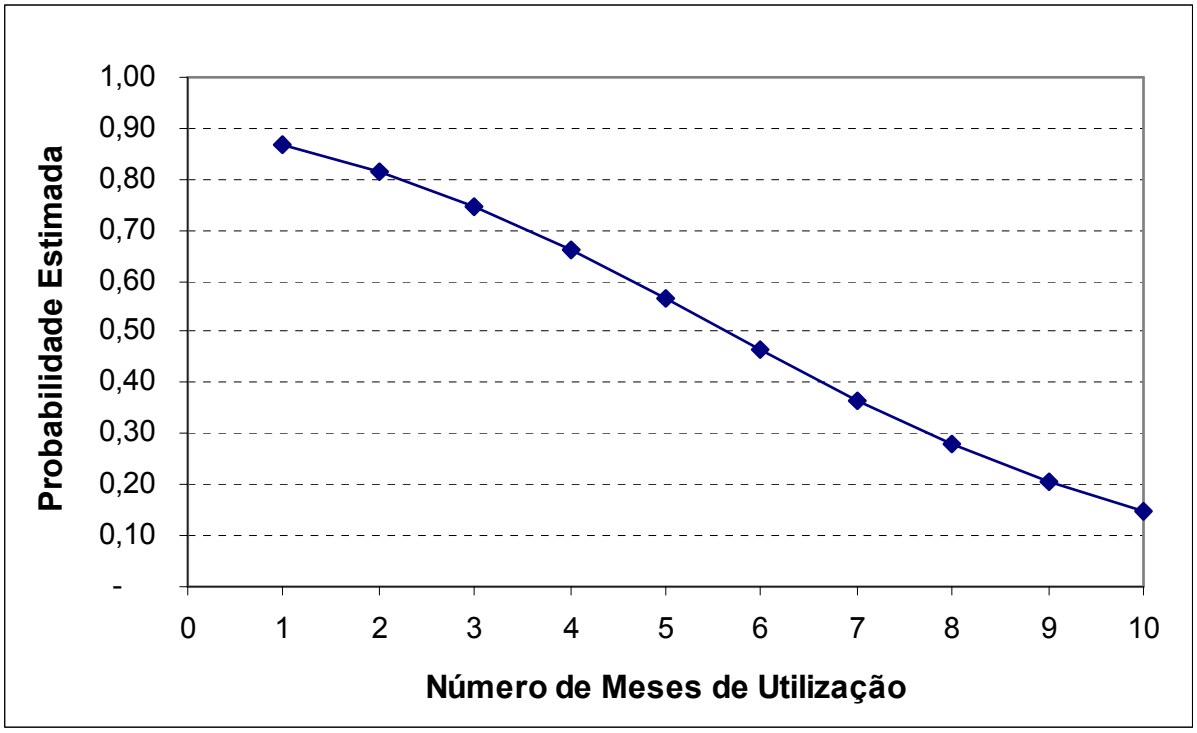
Variáveis	Razão de Vantagens
Tipo de Transação	-
01 - Somente Saque	3,9091
02,03,04,06 e 07 - Demais Transações	1,0000

Fonte: o Autor

4.4.2. Probabilidades Estimadas

Para enriquecer a análise, são apresentadas as probabilidades estimadas de a conta deixar de utilizar o cartão de crédito, condicionada a variações nas variáveis explicativas utilizadas no modelo final.

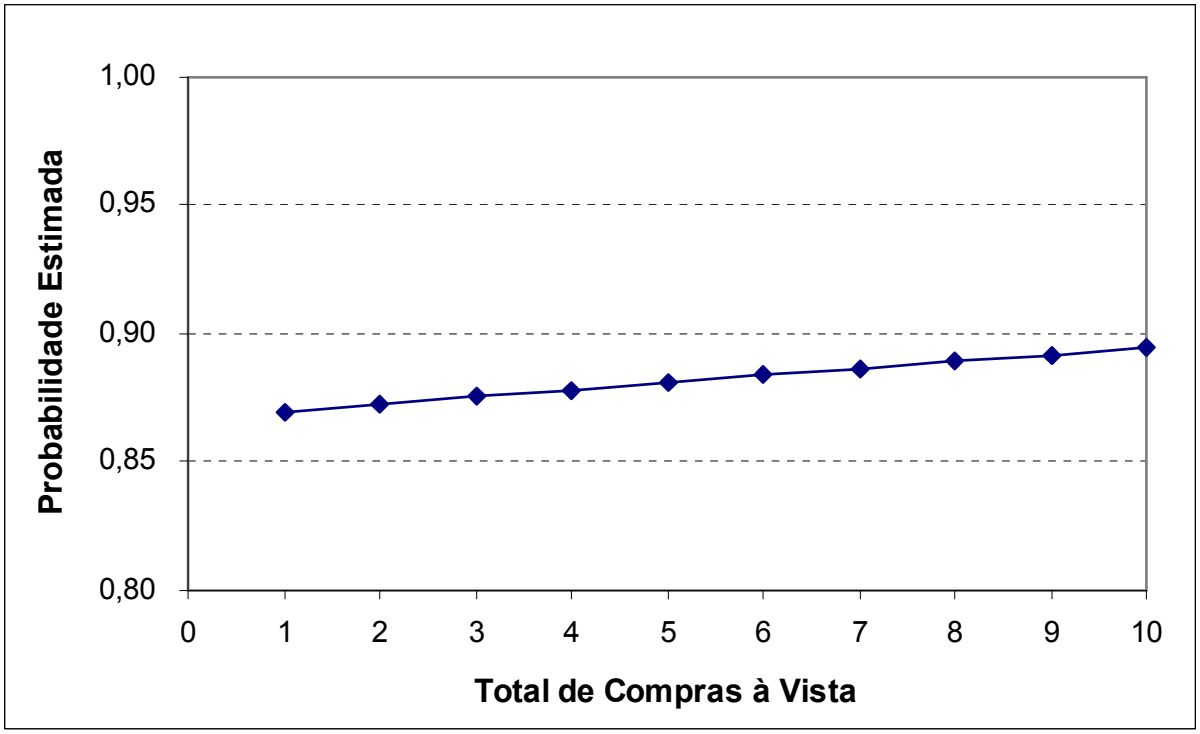
Gráfico 8 - Probabilidades Estimadas de a Conta estar Inativa por Número de Meses de Utilização



Fonte: o Autor

Observa-se que quanto maior Número de Meses de Utilização da conta, menor a probabilidade estimada da conta ficar inativa. Quanto maior o número de meses em que o cliente utiliza o cartão de crédito menos chances ele tem de deixar de utilizá-lo no futuro, o que condiz com o esperado.

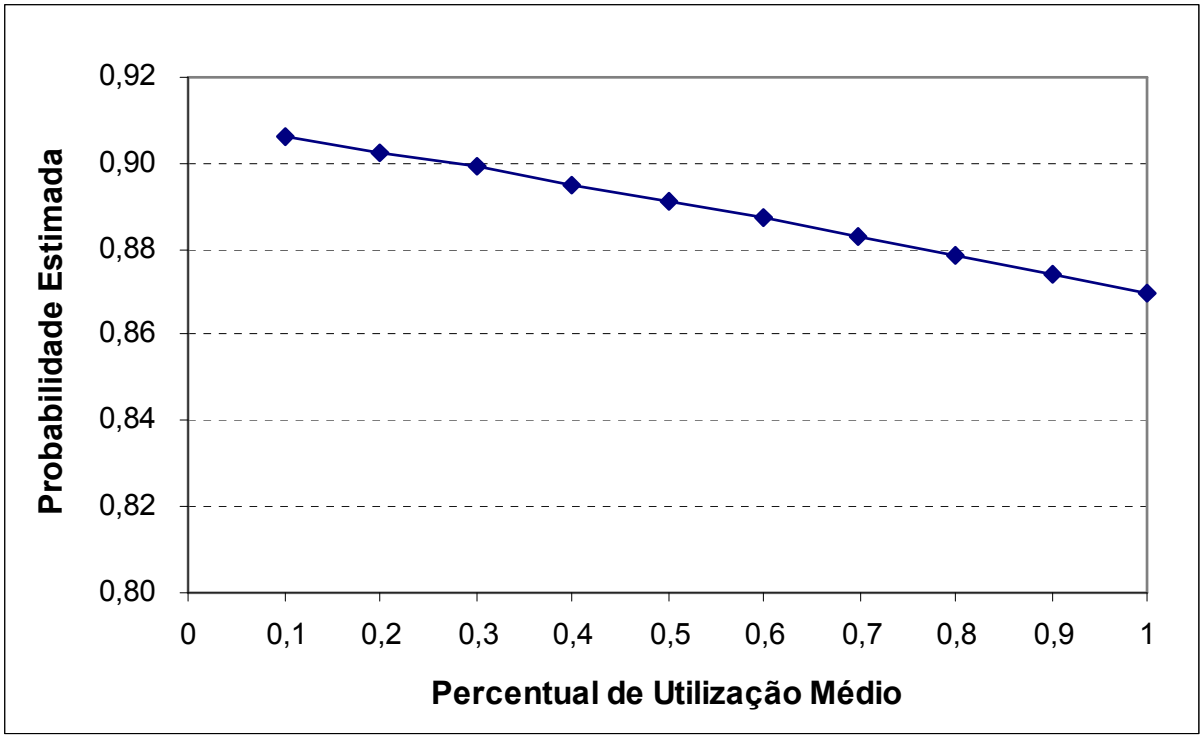
Gráfico 9 - Probabilidades Estimadas de a Conta estar Inativa por Total de Compras à Vista



Fonte: o Autor

Observa-se que quanto maior o total de compras à vista, maior a probabilidade estimada da conta ficar inativa.

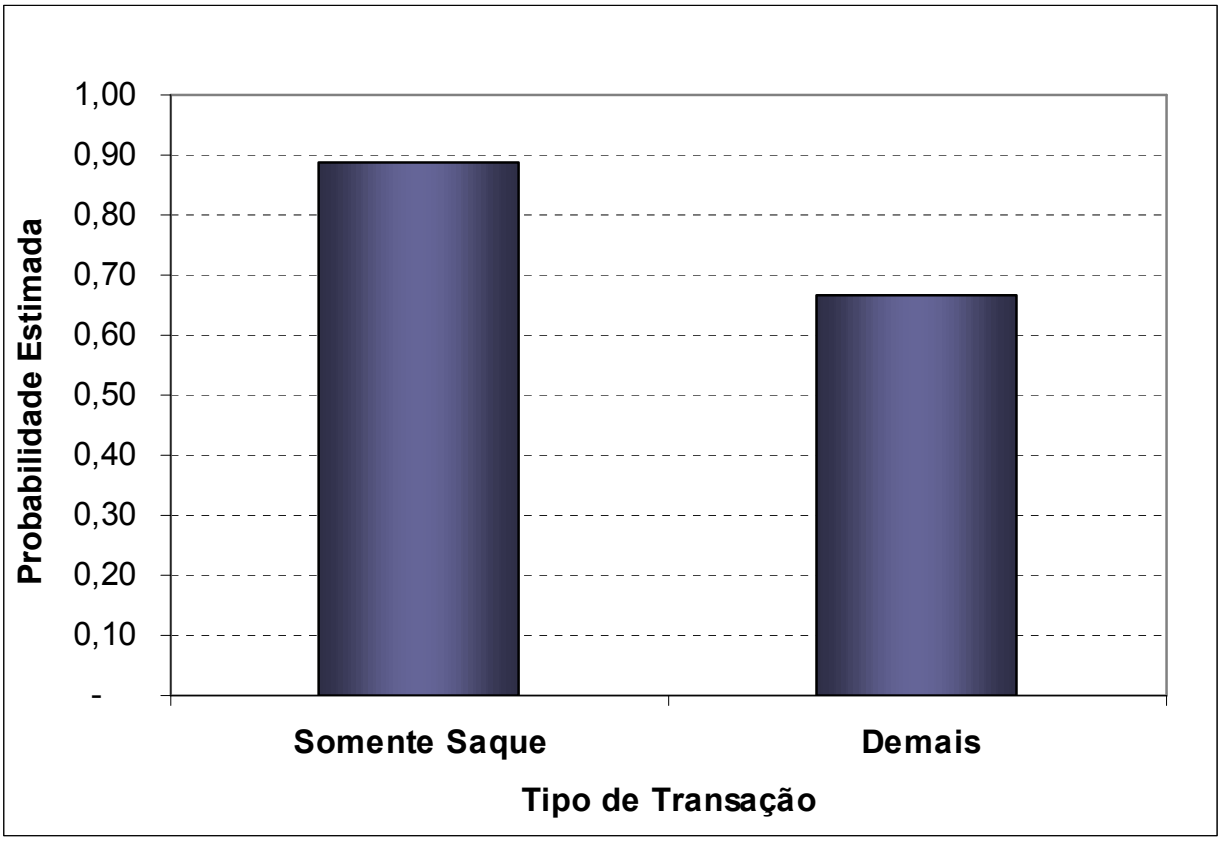
Gráfico 10 - Probabilidades Estimadas de a Conta estar Inativa por Percentual de Utilização Médio



Fonte: o Autor

Observa-se que quanto maior o percentual de utilização médio da conta, menor a probabilidade estimada da conta ficar inativa.

Gráfico 12 - Probabilidades Estimadas de a Conta estar Inativa por Tipo de Transação



Fonte: o Autor

A categoria somente saque do fator tipo de transação apresenta maior probabilidade de ocorrência de ficar inativo nos seis meses posteriores à data de corte da observação, e a categoria demais transações, a menor probabilidade de ficar inativo nos seis meses posteriores à data de corte da observação.

4.4.3. Classificação de Clientes

É comum a segmentação da carteira de clientes em diferentes grupos, a partir do perfil e do comportamento dos clientes ao longo do tempo, atráves da criação de classes de *score*. Desta forma, podem ser desenvolvidas diversas políticas específicas

para cada classe em função do perfil do cliente e de seu comportamento em determinado produto.

A seguir, apresenta-se na tabela 21 a comparação por classes de *score* entre o percentual de inativos real e o percentual de inativos estimados para os próximos seis meses pelo modelo ajustado.

Tabela 21 – Distribuição por Classes de Score

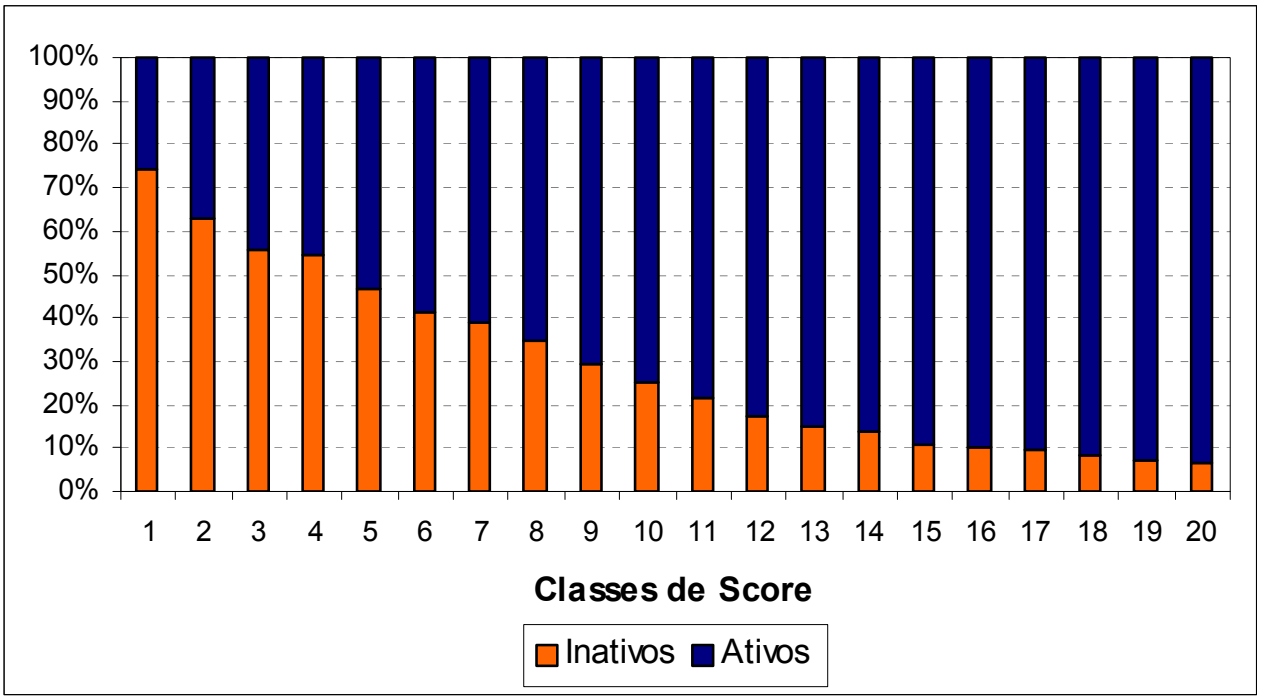
Classes	***% Contas***	***% Inativos Real***	***% Inativos Estimado***
1	5,0%	74,16%	74,69%
2	5,0%	62,72%	65,35%
3	5,0%	55,73%	59,42%
4	5,0%	54,67%	51,03%
5	5,0%	46,52%	46,74%
6	5,0%	41,12%	38,79%
7	5,0%	39,06%	35,29%
8	5,0%	34,69%	31,14%
9	5,0%	29,26%	26,03%
10	5,0%	25,03%	23,99%
11	5,0%	21,82%	21,26%
12	5,0%	17,39%	17,92%
13	5,0%	14,83%	16,24%
14	5,0%	13,71%	15,00%
15	5,0%	10,60%	13,85%
16	5,0%	10,11%	12,68%
17	5,0%	9,32%	11,41%
18	5,0%	8,18%	9,94%
19	5,0%	7,46%	7,99%
20	5,0%	6,72%	4,35%

Fonte: o Autor

Primeiramente, pode-se observar que a probabilidade real está diminuindo conforme a classe aumenta além de estar linear. O percentual de inativos estimado apresenta-se muito próximo do percentual real em cada classe de *score*.

A seguir, apresenta-se no gráfico 13 a distribuição das contas ativas e inativas nos seis meses posteriores à data de corte de observação por classes de *score*.

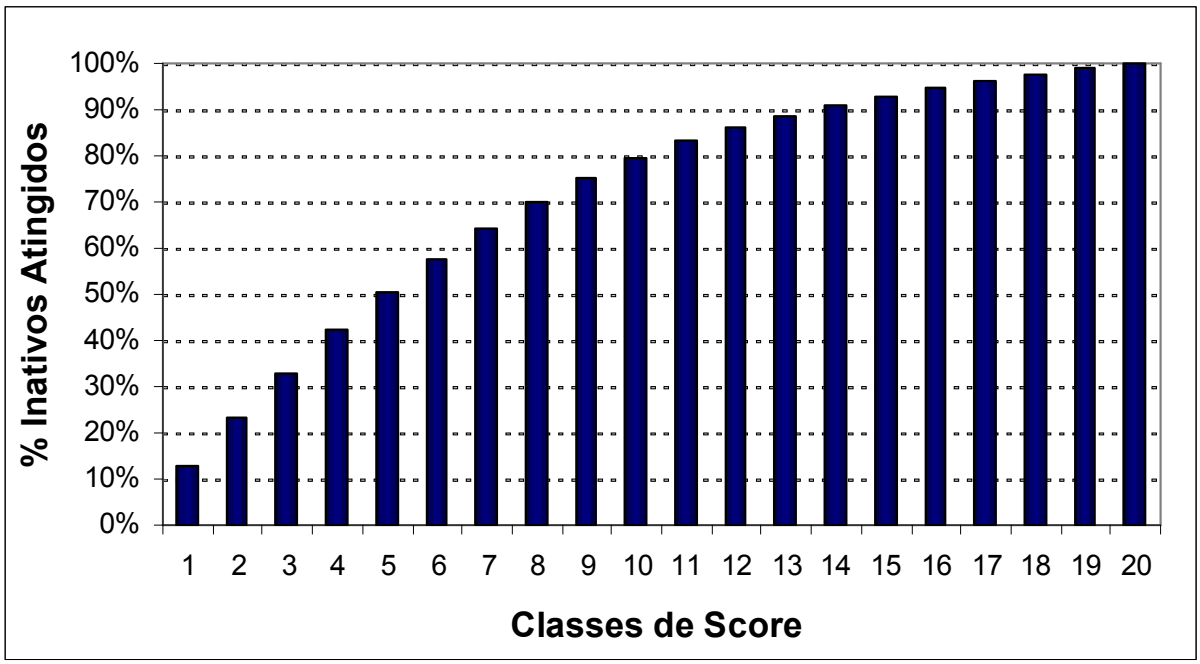
Gráfico 13 – Distribuição das Contas Ativas e Inativas por Classes de Score



Fonte: o Autor

Os clientes das menores classes de fato apresentam proporcionalmente maior probabilidade de se tornarem inativos, assim como as maiores classes apresentam maior probabilidade de se tornarem ativos.

Gráfico 14 – Distribuição das Contas Inativas por Classes de Score



Fonte: o Autor

No gráfico 14, apresenta-se o percentual de inativos atingidos no caso de uma ação de marketing realizada selecionando as classes de score da menor para a maior. Pode-se observar que se for selecionado somente metade dos clientes (classe 10), será atingido 80% dos inativos seis meses posteriores à data de corte da observação.

5. CONCLUSÕES E TRABALHOS FUTUROS

O objetivo deste estudo foi construir um modelo de regressão logística como ferramenta no auxílio à fidelização de clientes de um deteriminado cartão de crédito de uso exclusivo em postos de combustíveis e em suas lojas de conveniência. Na revisão de literatura, foram destacados os modelos empregados para tal estudo e verificou-se a presença de modelos multi-estado de Markov, análise de sobrevivência e regressão logística, como principais ferramentas estatísticas.

Adotou-se como técnica estatística a regressão logística utilizando a base de dados de clientes de uma financeira para desenvolver um modelo que determine o risco da não utilização das contas do cartão de crédito, para podermos traçar o perfil dos clientes com maior probabilidade de não utilização e assim, propor ações de retenção para aumentar o lucro da carteira de cartões de crédito da empresa.

A utilização da regressão logística como método de desenvolvimento do modelo *anti-attrition* foi correto, pois os resultados do modelo demonstraram robustez. Os resultados obtidos mostram que o modelo é bastante eficiente em separar as contas que possivelmente deixarão de utilizar o cartão de crédito, dado que o modelo proposto apresentou um elevado grau de explicação.

Este trabalho atingiu o objetivo proposto: construir um modelo de regressão logística para ser usado como ferramenta no auxílio à fidelização de clientes de um determinado cartão de crédito. Os resultados obtidos indicam a aplicabilidade da regressão logística para identificar de forma segmentada potenciais clientes com grandes chances de cancelamento.

Pode-se sugerir como trabalhos futuros a utilização das informações fornecidas pelo modelo para implementar uma ação focada de marketing com o objetivo de fidelizar os clientes. O modelo indicará o grupo com maior risco de cancelamento com mais eficiência do que sem sua utilização. Ao longo do tempo, deve-se verificar se o ocorrido com as contas do cartão foi o que o modelo apresentado já previa.

A carteira de cartões de crédito utilizada neste estudo é restrita, comtemplando somente clientes de um cartão de crédito que pode ser utilizado somente em determinados postos de combustíveis. A aplicação desta técnica em outras bases de dados de cartões de crédito seria interesante.

Outro ponto importante é a utilização de variáveis comportamentais do cliente no mercado, o que poderia gerar modelos de melhor desempenho dado um conjunto maior de variáveis explicativas. Estas variáveis podem ser obtidas nos bureaus de crédito.

Recomenda-se também a segmentação dos clientes da carteira para a construção de diversos modelos, pois na atual competitividade do mercado, é fundamental saber distinguir os diferentes comportamentos e capacidades de gerar lucros dos diversos nichos de clientes. Sem uma segmentação adequada, os tomadores de decisão não conseguirão obter todas as informações necessárias para seguir a melhor estratégia, não podendo assim, aumentar os lucros e os resultados da companhia.

Poderão ser construídos também modelos estatísticos para a adequação de limites de crédito das contas de cartão de crédito. Este modelo seria de grande valia para ser utilizado juntamente com o modelo de cancelamento. Uma das causas da não utilização do cartão de crédito pelo cliente pode estar relacionada ao baixo limite que lhe foi concedido.

Uma das contribuições deste trabalho é a identificação dos perfis de interesse das contas de cartão de crédito. Esta identificação de perfil poderá ser reutilizada por outras instituições que concordem com o modelo proposto. Desta forma, a metodologia aqui apresentada é mais uma contribuição à necessidade de diferenciação dos potenciais clientes à ruptura, partindo da premissa de que a instituição financeira deve possuir informações consistentes para escolher as ações mais eficientes em sua tomada de decisão.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ABECS. **Associação Brasileira das Empresas de Cartões de Crédito e Serviços**. Disponível em: http://www.abecs.org.br/mercado_cartoes.asp. 2008.

Abreu, H. J. **Aplicação da Análise de Sobrevivência em um Problema de Credit Scoring e Comparação com a Regressão Logísitica**. Dissertação (Mestrado em Estatística), Universidade Federal de São Carlos, São Carlos. 2004.

Agresti, A. **Categorical Data Analysis**. New York: John Wiley & Sons. 1990.

Brito, G. A. S.; Neto, A. A. Modelo de Classificação de Risco de Crédito de Empresas. **Revista Contabilidade & Financas**. V.19, n.46, p.18-29. 2008.

Burez, J.; Poel, D. V. Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department **. Expert Systems with Applications.** V.35, n.1-2, p.497–514. 2008.

Bussab, W. O. e Morettin, P. A. **Estatística Básica**. São Paulo: Editora Saraiva, 2005.

Campello, R. C. **Análise Multicritério Aplicada à Classificação da Solvência de Operadoras de Assistência à Saúde**. Dissertação (Mestrado em Engenharia de Produção), Universidade Federal Fluminense, Rio de Janeiro. 2005.

Carpenter, E. M. L. **Um Modelo de Análise de Risco de Crédito de Clientes em Relações B2B. Dissertação** (Mestrado em Administração), PUC, Rio de Janeiro. 2006.

Crook, J. N.; Edelman, D. B.; Thomas, L. C. Recent developments in consumer credit risk assessment. **European Journal of Operational Research.** V.183, n.3, p.1447-1465. 2007.

Cook, R. Dennis e Weisberg, Sanford. Diagnostics for Heteroscedasticity in Regression. **Biometrika.** V.70, n.1, p.1-10. 1983.

Draper, N.R.; Smith, H. **Applied Regression Analysis**. New York: Wiley. 1981.

Dobson, A. **An Introduction to Generalized Linear Models***,* London: Chapman and Hall, 1990.

Ferreira, J. M. **Análise de Sobrevivência: Uma visão de Risco Comportamental na Utilização de Cartão de Crédito***.* Dissertação (Mestrado em Biometria), Universidade Federal Rural de Pernambuco, Pernambuco. 2007.

Gonçalves, E. B. **Análise de Risco de Crédito com o Uso de Modelos de Regressão Logística, Redes Neurais e Algoritmos Genéticos***.* Dissertação (Mestrado em Administração), Universidade de São Paulo, São Paulo. 2005.

Grizzle, J. E., Starmer, C. F.; Koch, G. G. Analysis of Categorical data by Linear Models*.* **Biometrics**. V.25, n.3, p.489-504. 1969.

Hauck, W. W., Donner, A. Wald's test as applied to hypothesis in logit analysis*.* **Journal of the American Statistical Association**. V.72, n.360, p.851-853. 1977.

Hair, J. and Joseph; F. **Multivariate Data Analysis***,* New Jersey: Prentice Hall, 1998.

Hayhoe, C. R.; Leach , L.; Turner P. R. Discriminating the number of credit cards held by college students using credit and money attitudes. **Journal of economic psychology.** V.20, n.6, p.643-656. 1999.

Hoper, M. A.; Lewis, E. M**. Behaviour Scoring and Adaptive Control Systems. In Credit Scoring and Credit Control**, ed L. C. Thomas, J. N. Crook, D. B. Edelman, Claredons Press: Oxford. 1992.

Huang, J. J.; Tzeng, G.; Ong, C. Two-stage genetic programming (2SGP) for the credit scoring model. **Applied Mathematics and Computation.** V.174, n.2, p.1039–1053. 2006

Infomoney. Disponível em: http://web.infomoney.com.br/investimentos/cotacoes/indicadores/. 2008.

Jonhson, R. A. and Wichern, D. W. **Applied Multivariate Statistical Analysis**. New York: Prentice-Hall, 2002.

Karan, K. A*.* **Regressão Logística: Um Modelo de Risco de Cancelamento de Clientes**. Dissertação (Mestrado em Administração), PUC, Rio de Janeiro. 2006.

Kimura, H.; Lima, F. G.; Perera, L. C. J.; Donzelli, C. R.; Filho, A. C. Aplicação de Redes Neurais na Análise e na Concessão de Crédito ao Consumidor. **ENANPAD**, Brasília. 2005.

Lazaridis, A. A note regarding the problem of perfect multicollinearity**. Quality and Quantity.** v.20, n.2-3, p.297–306. 1986.

Lim, M. K.; Sohn, S. Y. Cluster-based dynamic scoring model**. Expert Systems with Applications.** V.32, n.2, p.427–431. 2007.

Mendes, V. P. **Modelos de Churn de Clientes em Plano de Saúde.** Dissertação (Mestrado em Engenharia de Produção), Universidade Federal Fluminense, Rio de Janeiro. 2008.

Neter, J.; Wasserman, W.; **Applied Linear Statistical Models: regression, analysis of variance and experimental designs**. Homewood, Ill: R. D. Irwin. 1974.

Onusic, L. M.; Viana, A. B. N. **Comparação dos Resultados de utilização de Análise por Envoltória de Dados e Regressão Logística em Modelos de Previsão de Insolvência: Um Estudo Aplicado a Empresas Brasileiras***.* FACEF pesquisa, Franca. V.7, n.1, p.19-33, 2004.

Ohtoshi, C. **Uma Comparação de Regressão Logística, Árvores de Classificação e Redes Neurais: Analisando Dados de Crédito**. Dissertação (Mestrado em Estatística), Universidade de São Paulo, São Paulo. 2003.

Pereira, G. H. A**. Modelos de Risco de Crédito de Clientes: Uma Aplicação a Dados Reais***.* Dissertação (Mestrado em Estatística), Universidade de São Paulo, São Paulo. 2004.

Régis, D. E. **Aplicação de Modelo Multi-Estado de Markov em Cartões de Crédito***.* Dissertação (Mestrado em Economia), IBMEC, São Paulo. 2007.

Ribeiro, R. O. A. **Análise Comparativa de Metodologias de Cálculo para o Valor do Tempo de Vida do Cliente. Estudo de caso Aplicado a uma Grande Rede de Supermercados.** Dissertação (Mestrado em Engenharia de Produção), Universidade Federal Fluminense, Rio de Janeiro. 2007.

Rosa, P. T. M. **Modelos de Credit Scoring: Regressão Logística, Chaid e Real**. Dissertação (Mestrado em Estatística), Universidade de São Paulo, São Paulo. 2000.

Silva, G. L. **Modelos Logísticos para Dados Binários. Dissertação** (Mestrado em Estatística), Universidade de São Paulo, São Paulo. 1992.

Sohn, S. Y.; Shin, H. W. Reject Inference in credit operations based on survival analysis. **Expert Systems with Applications.** V.31, n.1, p.26–29. 2006.

Souza, E. C. **Análise de Influência Local no Modelo de Regressão Logística**. Dissertação (Mestrado em Agronomia), Universidade de São Paulo, São Paulo. 2006.

Thomas, L. C.; Edelman, D. B.; Crook, J. N. **Credit Scoring and Its Applications***.* Siam: Philadelphia. 2002

Vasconcellos, M. S. **Proposta de Método para Análise de Concessões de Crédito a Pessoas Físicas**. Dissertação (Mestrado em Economia), Universidade de São Paulo, São Paulo. 2002.

West, D. Neural Network Credit Scoring Models. **Computers & Operations Research**. V. 27, n.11-12, p.1131-1152. 2000.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo