UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA
CURSO DE ESTATÍSTICA

BRUNO PINHEIRO DE ANDRADE

CONSTRUÇÃO E VALIDAÇÃO DE UM MODELO DE CLASSIFICAÇÃO DE RISCO DE CRÉDITO

FORTALEZA
2013

BRUNO PINHEIRO DE ANDRADE

CONSTRUÇÃO E VALIDAÇÃO DE UM MODELO DE CLASSIFICAÇÃO DE RISCO DE CRÉDITO

Monografia apresentada ao curso de Estatística do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará, como requisito parcial para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Rafael Bráz Azevedo Farias

FORTALEZA

2013

BRUNO PINHEIRO DE ANDRADE

CONSTRUÇÃO E VALIDAÇÃO DE UM MODELO DE CLASSIFICAÇÃO DE RISCO DE CRÉDITO

Monografia apresentada ao curso de Estatística do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará, como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovada em 17/12/2013.

BANCA EXAMINADORA

Prof. Dr. Rafael Bráz Azevedo Farias (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Juvêncio Santos Nobre
Universidade Federal do Ceará (UFC)

Prof. Dr. José Ailton Alencar Andrade
Universidade Federal do Ceará (UFC)

À Deus acima de tudo
À minha mãe Iseuda, por ser uma mãe muito dedicada
À toda minha família por me apoiar em todos os momentos.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por ter me possibilitado a sorte de estudar Estatística, sou muito feliz por essa oportunidade. Agradeço por iluminar meu caminho, por me dar forças para lutar, continuar sonhando sem perder o foco e por guiar pessoas importantes para minha vida, que sempre me apoiaram nas escolhas que tomei e estiveram presentes nos momentos bons e ruins. Agradeço por ter me escolhido no meio de tantos!

Às minhas mães Iseuda e Fátima, por terem sempre me dado muito amor e ter me possibilitado a oportunidade de sonhar, isso que potencializa minha vontade de vencer e continuar vencendo.

À minha irmã Amanda, por sempre esta do meu lado, me fortalecendo e me encorajando em todos os meus sonhos.

Ao meu pai Isaias, por ter sempre me dado o suporte necessário para me manter na universidade durante esse período de graduação.

À minha familia, por estarem presentes em todos os momentos de minha vida, por sempre terem dado os melhores concelhos e por sempre acreditarem em mim.

À minha namorada Renata que tanto amo, por estar ao meu lado, me dar carinho e amor.

Ao professor Rafael Farias pela orientação tanto na monografia quanto na vida acadêmica.

Aos meus professores do DEMA que contribuíram bastante para minha formação, e em especial ao professor Juvêncio que por muitas vezes me motivou.

Agradeço ao meu grupo de estudos composto por Yuri, Janaína e Kelly, vocês contribuíram muito para que eu chegasse ao final do curso, obrigado pela força e apoio.

Agradeço a minha segunda família à GAUSS, pois participar da GAUSS foi uma grande oportunidade que eu tive para aprender muito mais as práticas da estatística, me desenvolver como pessoa e amadurecer, OBRIGADO FAMÌLIA GAUSSIANA.

“A gente muda o mundo na mudança da mente, e quando a mente muda a gente anda pra frente.”

Gabriel O Pensador

“Dê-me uma alavanca e um ponto de apoio, e moverei o mundo.”

Arquimedes

“Nenhum vento sopra a favor de quem não sabe para onde ir.”

Oscar Wilde R. Sêneca

RESUMO

A análise de crédito é uma técnica bastante comum na área financeira. Quando uma instituição financeira vende um crédito a um cliente ela estará comprando um risco, em que esse risco é medido pela probabilidade do cliente não cumprir com suas obrigações. A fim de prever o risco em uma operação de crédito, foi desenvolvido o modelo de risco de crédito por meio de análise de regressão logística. A partir desses modelos de crédito as instituições financeiras vendem crédito a um risco menor maximizando os lucros com segurança. Esse trabalho visa apresentar as principais técnicas estatísticas para uma análise de risco de crédito e uma aplicação como exemplo didático. Os resultados encontrados na aplicação, mostram que o modelo ajustado conseguiu classificar bem os indivíduos.

Palavras-Chave: Análise de crédito. Análise de regressão logística.

ABSTRACT

The credit analysis is a common technique in finance. When a financial institution sells a credit to a customer, it will be buying a risk, which is the likelihood that the customer does not meet its obligations. In order to predict risk in a credit transaction, a the credit risk model was proposed using logistic regression analysis. From these models credit financial institutions selling loans to a lower risk maximizing profits safely. This work presents the main statistical techniques for analysis of credit risk and an application as a illustrative example. The results in the application show that the model was able to rank well adjusted individuals.

Keywords: Credit analysis. Logistic regression analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A gestão de risco de crédito representa um dos principais problemas enfrentados pelas instituições financeiras desde o inicio da sua atividade em meados dos anos 90, quando o mercado mundial tornou-se mais competitivo. Isso ocorre porque os bancos e as instituições financeiras em geral, tem como principal função a intermediação financeira. Portanto uma gestão de crédito otimizada, eficiente e segura impactaria de tal forma, que a evolução nesse quesito possibilitaria um maior desenvolvimento por parte de algumas instituições. Consequentemente as instituições melhores preparadas teriam o domínio desse mercado.

Em relação às operações de crédito, o banco concede crédito, sob a promessa de um recebimento futuro do capital com juros de acordo com o plano de reembolso contratado. Existe, contudo, na carteira de crédito da instituição, contratantes que podem não vir a cumprir as obrigações contratadas, implicando em prejuízos. Estes contratantes podem ser pessoas físicas ou jurídicas e, ao não cumprimento das responsabilidades, são chamados de inadimplentes. Nos últimos anos, devido sobretudo a pressões regulamentares, as instituições financeiras têm procurado criar metodologias mais eficientes para inferir a probabilidade de descumprimento esperado em cada operação de crédito.

O processo decisório de concessão de crédito era antigamente feito de forma essencialmente intuitivo, estruturando-se no "feeling" e na experiência dos analistas de crédito, onde esse procedimento era feito de tal forma que não havia nenhum padrão para classificação dos clientes. Os atuais analistas de crédito, munidos das metodologias estatísticas, conseguem analisar variáveis que explicam quais fatores podem levar os clientes a se tornarem inadimplentes. A partir da estatística os analistas conseguem mensurar o risco de cada cliente, desta forma consegue-se observar um padrão lógico, onde é possível classificar de forma mais coerente os processos de decisão.

O aumento da concorrência entre as instituições financeiras e a crescente pressão para a maximização das receitas impulsionam as instituições financeiras a procurarem meca-

nismos mais eficientes de atrair novos clientes com baixo perfil de risco e, ao mesmo tempo, controlar e minimizar as perdas. O desenvolvimento de novas tecnologias, o aumento da procura por crédito, bem como por uma questão de qualidade de serviço, a necessidade de responder o mais rápido possível às solicitações levou ao desenvolvimento e aplicação de sofisticados modelos estatísticos na gestão de risco de crédito, designados por *Credit Scoring*.

Lewis (1992) define os modelos de risco de crédito como sistemas que atribuem escores às variáveis de decisão de crédito de um requerente, mediante a aplicação de técnicas estatísticas. Esses modelos visam sumariar todas as características que permitem distinguir entre bons e maus pagadores.

A partir de uma equação estimada com base nas características dos solicitantes de crédito, é gerado um escore que representa o risco de perda em cada operação. O escore que resulta da equação, é interpretado como a probabilidade de descumprimento, que comparado com o ponto de corte previamente estabelecido associado a um conjunto de regras e filtros, permite ajuizar quanto à concessão ou não de crédito. Assim, a ideia básica dos modelos de risco de crédito é identificar certos fatores chaves que influenciam a probabilidade de descumprimento dos clientes, permitindo a classificação dos mesmos em grupos distintos e, como consequência, a decisão sobre a aceitação ou não da proposta em análise.

Os métodos usados no risco de crédito incluem várias técnicas estatísticas e de investigação operacional, sendo as mais utilizadas a regressão logística segundo Thomas (2000), a análise discriminante, árvores de decisão (Joos *et al*, 1998) e análise de regressão com procedimentos bayesianos (Campos, 2007).

1.1 Motivação

A concessão de crédito desempenha um papel fundamental no desenvolvimento de uma economia, em decorrência da dinâmica que introduz no processo econômico, seja como uma oportunidade para as empresas (especialmente as pequenas e médias empresas) aumentarem os seus níveis de produção ou como estímulo ao consumo dos indivíduos.

Segundo Baptista (2006), o reconhecimento de que os mercados financeiros, através do negócio de crédito privado, contribuem para o desenvolvimento econômico, é bem marcante na literatura financeira. A título de exemplo, o mercado de crédito ao consumo nos Estados Unidos da América tem demonstrado que há estabilidade econômica baseada

em políticas sólidas de crédito é sinônimo de prosperidade econômica, baixas taxas de desemprego e baixas taxas de juros. Ao longo das últimas décadas, o crédito ao consumo nos EUA tem crescido em ritmo fenomenal, tendo atingido em 2007 a marca de $13 trilhões de dólares, superando em 40% o crédito concedido ao setor industrial e, em 24% ao crédito às empresas, (Thomas, 2009). A par de outros fatores, o risco de crédito, dado o automatismo que assegura, foi o fator que mais permitiu a abertura do mercado de crédito a todos os consumidores, mantendo o risco em um nível controlável.

A Tabela 1 apresenta as principais vantagens e desvantagens citadas por Caouette *et al* (1998) para mensurar o risco de forma objetiva:

Tabela 1 - Vantagens e desvantagens

Vantagens	Desvantagens e limitações
Revisão de crédito mais consistente	Custo de Desenvolvimento
Informações organizadas	Modelos com "excesso de confiança"
Eficiência no trato de dados fornecidos por terceiros	Informações errôneas
Diminuição da metodologia subjetiva	Interpretação equivocada dos escores
Maior Eficiência do processo	Limitações geográficas e temporais

Fonte: Caouette *et al* (1998).

Sicsú (2010) cita abaixo algumas vantagens das metodologias de risco de crédito utilizando as metodologias estatísticas.

- **Consistência nas informações:** Ao submeter uma mesma solicitação de crédito a diferentes analistas, pode-se obter diferentes avaliações subjetivas, pois a experiência e o envolvimento com o cliente diferem entre eles. Um mesmo analista pode ter diferentes avaliações para uma mesma proposta se submetida em momentos diferentes. No entanto, isso não ocorrerá se aplicado um modelo quantitativo de risco de crédito. Mantidas inalteradas as características da solicitação, o escore será o mesmo, independentemente do analista, da agência ou da filial do credor.

- **Decisões rápidas:** Os recursos computacionais hoje disponíveis permitem que os escores sejam obtidos quase que instantaneamente, logo após cadastro dos dados da solicitação. Centenas ou milhares de decisões são tomados por dia, de forma segura. A pronta resposta a um cliente potencialmente é uma vantagem competitiva do credor.

- **Decisões adequadas:** Em função do risco quantificado, o credor poderá adotar diferentes regras de concessão.

  - O conhecimento das probabilidades de perda permite calcular perdas e ganhos esperados com as operações. Isso permite precificar as operações de forma adequada.
  - Os clientes podem ser divididos em classes de risco conforme seu escore. Para cada classe, o credor pode adotar diferentes regras de concessão de crédito, diferenciando, por exemplo, as taxas a aplicar. Ao reduzir essas taxas para clientes de baixo risco, terá como efeito a conquista de maior número de clientes, ou seja, de ampliação de mercado.

- **Decisões à distância:** Atualmente, com os recursos de transmissão de dados disponíveis, o credor não precisa alocar um analista de crédito em cada loja ou filial. O vendedor insere os dados no ponto de venda e, logo após submeter essas informações, receberá a decisão de crédito em sua tela.

- **Monitorar e administrar o risco de um portfólio de crédito:** Sem a quantificação do risco individual esta tarefa é inviável. Para a avaliação do risco do portfólio são necessárias, além dos escores, outras medidas que não serão discutidas neste texto, veja em Sicsú (2010) outras medidas.

  A utilização de medidas objetivas para o risco de crédito permite também:

  - verificar o grau com que se atendemos aos requisitos de órgãos regulares;
  - estabelecer uma linguagem comum entre as decisores de crédito e
  - definir níveis de alçada para concessão de crédito.

1.2 A importância do modelo de risco de crédito

O modelo de risco de crédito representa um papel fundamental nas práticas de gestão de risco da maioria dos bancos. São usados para quantificar o risco de crédito da contraparte ou da transação durante as diferentes fases do ciclo de crédito (por exemplo, modelos de solicitação, comportamentais ou de cobrança). O escore de crédito permite que os usuários tomem decisões rápidas, ou mesmo automatizem decisões. Isso é desejável quando os bancos lidam com grande número de clientes e margens de lucro relativamente pequenas nas transações individuais (ou seja, crédito ao consumidor) como também incrementando negócios com as pequenas empresas.

1.3 Definição do problema

No cenário mundial dos bancos, financeiras, seguradoras e outras instituições, existe a necessidade de otimizar seus processos a fim de competir ganhando mais clientes que a concorrência, por isso a importância de mensurar o risco de crédito. O risco de uma solicitação de crédito pode ser avaliado de forma subjetiva ou mensurado de forma objetiva utilizando metodologias quantitativas. A avaliação subjetiva, apesar de incorporar a experiência do analista, mas não utilizar as ferramentas estatísticas, não consegue quantificar o risco de crédito. Dizer que uma operação é de alto risco não é suficiente para estimar de maneira precisa as perdas ou ganhos esperados e, consequentemente, tomar a decisão mais adequada.

1.4 Objetivo

1.4.1 *Objetivo Geral*

Apresentar a metodologia de desenvolvimento, validação e monitoramento do modelo de risco de crédito, a fim de popularizar a técnica e incentivar novos alunos no aprendizado e desenvolvimento.

1.4.2 *Objetivos Específicos*

- Apresentar as técnicas estatísticas utilizadas para desenvolver um modelo de risco de crédito.
- Desenvolver um modelo de risco de crédito em um banco de dados de forma didática.
- Validar o modelo desenvolvido.
- Monitorar o modelo.

1.5 Justificativa

Esta monografia se justifica pelo fato desta metodologia estatística ter grande aplicação no segmento financeiro e estar em grande expansão. Ultimamente o mercado financeiro está absorvendo muitos profissionais que possuem esse conhecimento. Então, este tra-

balho apresenta as técnicas estatísticas básicas para a modelagem do risco de crédito a fim de divulgar esta metodologia entre alunos que visam este tipo de mercado.

1.6 Estrutura do trabalho

O trabalho foi desenvolvido em cinco partes, são eles: Introdução, Metodologia, Qualidade do ajuste, Aplicação e Conclusão. A Introdução aborda alguns aspectos históricos, a motivação do estudo e objetivo do trabalho. Na Metodologia são abordadas as técnicas estatísticas para o desenvolvimento do estudo, através da amostragem, da análise exploratória, discretização e categorização de variáveis e regressão logística. Na análise da Qualidade do ajuste serão apresentadas as medidas utilizadas para validar o modelo proposto, tais como: estatística de Kolmogorov-Smirnov, Curva ROC, coeficiente de GINI e algumas medidas de eficiência do modelo. A Aplicação será composta pela análise exploratória do banco de dados, desenvolvimento do modelo e validação, e por último, a Conclusão com os resultados finais de toda a análise e a conclusão final a respeito do modelo proposto.

2 METODOLOGIA

Esta monografia tem caráter empírico-descritivo e se propõe a análise de dados já conhecidos da literatura e, com base nessa análise, propor um modelo de risco de crédito baseado na regressão logística. Esse modelo de aprovação de crédito tem como objetivo principal servir de auxílio na avaliação de decisão do analista sobre a concessão de crédito.

2.1 Modelos de risco de crédito

Os modelos de risco de crédito são usualmente divididos em quatro categorias: modelos de aprovação, prospecção, escoragem comportamental e recuperação.

Modelos de aprovação, são modelos que tem por finalidade mensurar o risco de crédito em clientes que vão até há instituição financeira em busca de um produto de crédito.

Já os modelos de prospecção tem por finalidade captar novos clientes, isto ocorre da seguinte forma: as instituções financeiras costumam comprar de empresas como a "SERASA", bancos de dados com informação dos clientes, em que os bancos de dados são analisados a fim de identificar quais clientes são de baixo risco, essa identificação é feita através de modelos de prospecção de crédito. Após identificar os bons clientes, as instituições começam a oforecer produtos de crédito aos clientes com a esperança de fechar contratos.

Modelos de escoragem comportamental auxiliam na administração dos créditos já existentes, ou seja, aqueles clientes que já possuem uma relação creditícia com a instituição, então esse modelo tem a finalidade de prever em que momento seus clientes irão descumprir, gerindo de melhor forma a cobrança preventiva.

A partir da pontuação comportamental, os modelos de recuperação de crédito, estuda o desempenho de cobrança de certos grupos, na tentativa de reduzir custos e aumentar as recuperações. Os clientes com pouca probabilidade de efetuar o pagamento, por sua vez, são classificados como de alto risco, e passam a ser acompanhados de maneira mais

próxima. O fato dos clientes não cumprirem com suas obrigações, causa um grande prejuízo, pois a instituição terá que mover uma ação na justiça contra o inadimplente e o fato de recuperar o dinheiro e o cliente com os modelos de recuperação minimizam as perdas.

Neste trabalho abordaremos apenas o modelo de aprovação de crédito. No entanto, a principal diferença entre eles são as variáveis a serem anlisadas, pois o procedimento utilizado na análise é similar.

2.2 Amostragem

Em Bolfarine e Bussab (2005), quando se tem o interesse de estudar um fenômeno, ocorre que a população desse fenômeno é geralmente grande e esse estudo acaba tendo o obstáculo de fazer o levantamento de toda a população, tornando-se inviável o estudo. Para a solução desse problema, pesquisadores desenvolveram a técnica estatística de amostragem, que tem por objetivo retirar uma amostra de tal forma que pudesse representar a população de estudo. Essa técnica fez com que muitos estudos e pesquisas tivessem mais representatividade e confiabilidade, e o ganho dessa técnica foi menor custo e mais rapidez no levantamento dos dados.

2.2.1 *Amostragem em modelos com resposta binária*

Esta técnica de amostragem é utilizada quando a modelagem estatística apresenta variável resposta binária, por exemplo: bom ou mau cliente, tem ou não tem câncer (Bolfarine e Bussab, 2005). Em modelos de risco de crédito, adota-se, como variável resposta Y_i a característica de ser bom ou mau o $i-$ésimo cliente do conjunto de dados da população. A classificação dos clientes é usualmente dada por:

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

Uma característica interessante apresentadas nos bancos de dados das instituições financeiras, é a baixa frequência de maus clientes em relação aos bons clientes. Para a modelagem de risco de crédito, o ideal é que exista uma proporcionalidade entre bons e maus clientes, pois assim o modelo não favorece nenhuma das partes. Thomas *et al* (2002) sugerem que as amostras em um modelo de risco de crédito estejam um para um,

de bons e maus clientes, ou algo em torno desse valor.

A seguir são listadas algumas considerações básicas sobre o tamanho da amostra em modelos de escoragem sugeridas por Sicsú (2010).

- O tamanho da amostra deve ser muito maior que o número de variáveis utilizadas para o modelo. Quanto maior a amostra, mais confiáveis são os resultados;
- O número de maus clientes deve ser suficientemente grande para que o modelo seja confiável. Alguns autores recomendam que este número não seja inferior a 1.000 (em geral 500 maus e 500 bons) clientes;
- Sempre que possível, a amostra deve ser suficientemente grande para que possa dividi-la em duas partes, uma para desenvolvimento e outra para validação do modelo.

2.3 Análise exploratória de dados

Tal análise é bastante utilizada quando o analista tem por necessidade conhecer os dados que está trabalhando, pois esse primeiro contato tem por objetivo verificar a consistência dos dados e assim poder validá-los para dar progressão à análise. Em análise de risco de crédito, o objetivo principal é gerar um modelo de risco, mas para tal feito é necessário uma boa análise exploratória, pois é importante conhecer o comportamento de cada variável e ter uma clara noção do perfil dos clientes que compõem a amostra. Caso ocorra algum problema com os dados e esse problema não seja solucionado, independentemente do modelo ajustado, este modelo deverá fornecer péssimas predições, podendo comprometer a saúde financeira da instituição. A análise exploratória consiste basicamente em uma análise univariada, bivariada, medida de associação e medidas de discriminação.

2.3.1 *Análise univariada*

A análise univariada é o estudo das variáveis de forma individual. Esse método tem por objetivo:

- entender a forma da distribuição de frequências da variável;

- identificar valores inconsistentes, ou seja, valores cuja existência não faz sentido dentro do contexto do estudo ou da definição operacional da variável;
- verificar a existência de valores "em branco" ou *missing value* e decidir quanto à maneira de tratá-los;
- identificar valores discrepantes (*outliers*), investigar sua origem e decidir como tratá-los de modo que não comprometam a validade e adequabilidade dos modelos estatísticos;
- verificar o excesso de valores registrados como "outros" (conhecidos como síndrome de outros), pois essa categoria de resposta podem levar a muitos outros tipos de respostas, nessecitando assim de um estudo mais apurado.

2.3.2 *Análise bivariada*

A análise bivariada é o estudo da a relação entre duas variáveis. O interesse maior é na relação das covariáveis relacionadas com a variável resposta. Essa análise tem por objetivo:

- analisar o potencial discriminador das variáveis preditoras;
- analisar e refinar a categorização das variáveis;
- identificar comportamentos estranhos ou inesperados de uma variável;
- identificar correlação entre as variáveis preditoras.

2.3.3 *Risco Relativo (RR)*

O risco relativo é utilizado para medir o potencial discriminatório de cada nível da variável categórica. Essa medida é importante na identificação de níveis discriminatórios, caso não tenham, essa variável terá que passar por um processo de categorização de variáveis para poder ter em todos os níveis uma boa discriminação. O cálculo do risco relativo dar-se por:

$$\text{RR} = \frac{\mathbb{P}(\text{pertencer à uma categoria j}|Bom)}{\mathbb{P}(\text{pertencer à uma categoria j}|Mau)}.$$

Portanto pode-se dizer que uma variável está bem categorizada, quando o cálculo para todas os níveis estejam bem afastados entre os cálculos dos outros níveis da mesma categoria.

2.3.4 *Peso de Evidência*

O Peso de Evidência (*Weight of Evidence*, WOE) é um critério amplamente utilizado. Ele tem a mesma finalidade do RR, ele mensura o risco de uma determinada classe. O WOE é obtido a partir do logaritmo do risco relativo. Ele tem a vantagem de ter o valor zero como ponto de referência. A interpretação é semelhante ao do RR. Entretanto, o WOE toma valores em toda a reta real e o valor é uma função das probabilidades condicionais dentro de todas as classes da categoria de status (Bom/Mau). O peso de evidência é uma forma de mensurar a distância entre as categorias das duas distribuições (distribuição de bons e distribuição de maus clientes). A fórmula do cálculo do WOE é dada por:

$$\text{WOE} = \log\{\text{RR}\}.$$

2.3.5 *Teste Qui-Quadrado de Pearson (χ^2)*

Ao se realizar a análise bivariada, uma possibilidade é fazer o cruzamento das variáveis preditoras com a variável resposta, Nesse cruzamento pode-se utilizar o teste de χ^2 de Pearson a fim de verificar se existe relação entre as duas variáveis. Caso exista relação entre as variáveis, essa variável pode ser uma possível candidata a compor o modelo, caso contrário ela não terá associação, indicando que não explica a variável resposta e provavelmente não fará parte do modelo, veja Campos (1983). As hipóteses do teste são:

$$\begin{cases} H_0: & \text{Não existe associação entre as variáveis} \\ H_1: & \text{Existe associação entre as variáveis} \end{cases}$$

A estatística de teste é dada por:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

em que O_i é a frequência observada e E_i é a frequência esperada.

A regra de decisão no teste χ^2, quando o valor-p é maior que $\alpha\%$, indica que não rejeita-se H_0, logo a variável em questão não explica a inadimplência, caso contrário o valor-p seja menor que $\alpha\%$, indica que rejeita-se H_0, logo a variável em questão explica a inadimplência e esta variável poderá ser uma caditada a compor o modelo.

2.4 Discretização de variáveis

Discretizar ou categorizar é um processo que divide em classes variáveis quantitativas ou qualitativas. Em Sicsú (2010), pode-se encontrar algumas vantagens e desvantagens a respeito desse procedimento. Algumas dessas vantagens são citadas logo mais abaixo:

- a discretização de uma variável quantitativa permite entender com mais facilidade a relação dessa variável com os *status*(bom/mau) do cliente. Ou seja, quais categorias da variável apresentam maior risco de crédito, quais são neutras e quais apresentam menor risco;
- para que um modelo seja implantado é necessário "vende-lo" à área de crédito. A experiência mostra claramente que os analistas desta área entendem melhor o comportamento de uma variável quando ela é apresentada discretizada (na forma de tabela). Recursos estatísticos como o *boxplot*, por exemplo, que permitem comparar o comportamento de uma variável contínua entre bons e maus clientes, não são compreendidos por leigos em Estatística;
- no caso de uma variável ordenada (quantitativa ou qualitativa), a discretização permite analisar se a variação do risco também segue essa ordenação;
- quando a relação entre uma variável quantitativa e a medida de risco não é monotônica, a discretização é extremamente vantajosa, pois cada categoria será tratada de forma independente ao calcular os pesos. Isso decorre do uso de variáveis binárias geradas a partir dessas categorias.

A discretização em contra partida apresentam algumas desvantagens, essas são citadas a seguir:

- perde-se informação ao agrupar indivíduos com valores distintos e trata-los de forma semelhante;

- favorece ou prejudicam indivíduos que estão próximos às fronteiras de cada categoria;
- aumenta a dimensionalidade do problema.

Neste trabalho será aplicada a técnica de discretização, por entender que essa técnica apresenta mais vantagens do que desvantagens, sem comprometer significativamente a eficiência do modelo.

2.4.1 *Discretização de variáveis quantitativas*

A discretização pode ser obtida a partir da expertise técnica do analista de crédito. No caso em que o analista não tem experiência em discretização, o autor sugere alguns passos para esse processo:

- *1° passo:* Deve-se dividir a variável contínua em dez níveis de frequência aproximadamente iguais a 10%;
- *2° passo:* Em seguida deve ser feito o cruzamento desta variável com a variável resposta;
- *3° passo:* O analista deve notar se existe caselas com zero, caso exista, deve-se unir esta casela a uma casela com nível mais próximo que apresenta maior frequência;
- *4° passo:* Em seguida deve-se calcular o risco relativo e verificar quais valores estão mais próximos e em seguida uni-los.
- *5° passo:* Esse processo de união será repetido pelo analista até que tenha poucas categorias e com as medidas dos riscos relativos distantes.
- *6° passo:* Por fim deve ser feito o teste de Qui-Quadrado de Pearson e em seguida verificar se houve relação entre as variáveis. Deve-se lembrar de que cada nível não deve ter uma frequência baixa e que o número final de categorias não deve ser muito superior a cinco, exceto em casos excepcionais.

2.4.2 *Fusão de variáveis qualitativas ou quantitativas*

Quando se tem variáveis categóricas do tipo que não apresenta sequência lógica os passos para a fusão são os seguintes:

- *1° passo:* Devem-se cruzar estas variáveis com a variável resposta;
- *2° passo:* Em seguida calcula-se seus respectivos riscos relativos.
- *3° passo:* No caso de não existir ordem lógica, tem-se a possibilidade de a ordem ser construída a partir do risco relativo, definido isso verifica-se se existe riscos relativos próximos, caso exista, deve-se unir esses níveis até que não tenham riscos relativos próximos.
- *4° passo:* Após definir os níveis é feito o teste de χ^2 a fim de verificar a associação das variáveis. Quando a variável categórica tiver ordem lógica o analista deve calcular o risco relativo e manter a ordem lógica da variável, caso tenha as medidas do risco relativo próximas deve-se analisar a possibilidade de fazer a fusão com a variável que faça mas sentido ou a categorização deve permanecer assim, pois essa forma é a maneira natural de como os dados se comportam.

2.5 Regressão logística

A regressão logística é comumente utilizada para análise de dados com resposta binária ou dicotômica e consiste em relacionar, por meio de uma função, a variável resposta (variável dependente binária) com fatores que influenciam ou não a probabilidade de ocorrência de determinado evento (variáveis independentes).

Supondo um evento dependente em que a variável Y é uma variável binária que assume valores 0 ou 1, e variáveis independentes $\mathbf{X} = (x_1, x_2, \ldots, x_k)^t$. Neste caso, o modelo logístico pode ser adequado para modelar uma relação entre a probabilidade de um cliente ser mau ou bom pagador com um conjunto de fatores ou atributos que o caracterizam, variáveis explicativas. Esta relação é definida pela função logito e dada pela expressão (McCcullagh & Nelder, 1989):

$$\log\left\{\frac{\pi(\mathbf{x_i})}{1-\pi(\mathbf{x_i})}\right\} = \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_k$$

em que $\pi(x_i)$ pode ser escrito como:

$$\pi(\mathbf{x_i}) = \mathbb{P}(Y = 1|x_i) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_k)} \tag{2.1}$$

e pode ser interpretado como a probabilidade de um proponente ao crédito ser um mau

ou bom pagador dado as características que possui, representadas por $\mathbf{x_i}$.

Além disso, essa passagem não se dá de forma linear, mas sim de acordo com a forma não linear da função logística.

Analisando o significado da função de regressão logística no contexto de risco de crédito, tem-se que a variável dependente consiste na situação de solvência (estado em que o cliente honra com seus compromissos com os recursos que constituem seu patrimônio) do cliente que assumirá valores 0 ou 1, a depender de os dados procederem de um cliente adimplente. As variáveis independentes representam os fatores que se supõe influenciarem a inadimplência (descumprimento de um contrato) como, por exemplo, dados pessoais e financeiros. A probabilidade de inadimplência de uma determinada pessoa é dada por $\pi(\mathbf{x_i})$, que é a probabilidade condicional de Y assumir o valor dado suas características indivíduais. Os coeficientes estimados β representam a contribuição das variáveis explicativas x no logito da probabilidade.

O conjunto de atributos usados na regressão logística depende do tipo de modelo em desenvolvimento. Modelos de solicitação de crédito (*Aplication*), usados para decidir entre aceitar ou rejeitar um solicitante, costumam usar apenas informações pessoais sobre ele, dado que normalmente são as únicas informações disponíveis para o banco nesse estágio.

Andrade (2003) diz que uma vez desenvolvido o modelo, ele precisará ser testado com uma amostra para confirmar a solidez de seus resultados. Quando há dado o bastante, uma parte da amostra de desenvolvimento (amostra reservada) costuma ser destacada para o teste final do modelo. Mas um teste ideal exigiria também investigação de seu desempenho com uma amostra de outro tempo e outro universo.

2.5.1 *Estimação dos parâmetros*

Em Diniz e Louzada (2012) os parâmetros utilizados no modelo tem seus pesos estimados geralmente pelo método de máxima verossimilhança (Hosmer e Lemeshow, 2000). Neste método, os coeficientes são estimados de maneira a maximizar a probabilidade de se obter o conjunto de dados observados a partir do modelo proposto, em outras palavras, esse método estima melhor os valores dos coeficientes melhorando a inferência sob o modelo. Para a aplicão de tal método, primeiramente deve-se construír a função de verossimilhança que expressa a probabilidade dos dados observados, como função dos parâmetros $\beta_0, \beta_1, \ldots, \beta_k$. A maximização desta função fornece os estimadores de máxima verossimilhança para os parâmetros. No modelo de regressão logística, uma forma conveniente para

expressar a contribuição de um cliente $(y_i; x_i)$ para a função de verossimilhança é dada por:

$$\mathrm{L}(\beta; x_i) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{2.2}$$

Uma vez que as observações, ou seja, os clientes são considerados independentes, a função de verossimilhança pode ser obtida como produto dos termos na Equação (2.2), em que $\pi(\mathbf{x_i})$ é expresso na Equação (2.1)

$$\mathrm{L}(\beta; \mathbf{x}) = \prod_{i=1}^{n} \mathrm{L}(\beta; \mathbf{x}). \tag{2.3}$$

A partir do princípio da máxima verossimilhança, os valores das estimativas para β são aqueles que maximizam a equação (2.2). No entanto, pela facilidade matemática, é melhor trabalhar com o logaritmo dessa expressão, que é definida como:

$$\ell(\beta) = \log[\mathrm{L}(\beta)] = \sum_{i=1}^{n} \{y_i \log[\pi(x_i)] + (1 - y_i) \log[1 - \pi(x_i)]\} \tag{2.4}$$

Para obtenção dos valores dos coeficientes, que maximizam $\ell(\beta)$, calcula-se a derivada em relação a cada um dos parâmetros $\beta_1, \beta_2, \ldots, \beta_k$, sendo obtidas as seguintes equações

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

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0, \text{para } j = 1, 2, \ldots, k.$$

as quais, uma vez solucionadas via métodos numéricos, como por exemplo Newton-Raphson, fornecem as estimativas de máxima verossimilhança. Em geral, usa-se métdos mais elaborados, tais como escore de fisher dentre outros, em que é o mais comum de ser encontrado nos pacotes estatísticos é o de Newton-Raphson.

A partir do modelo ajustado pode-se predizer a probabilidade de novos candidatos a crédito serem maus pagadores. Esses valores preditos são utilizados, normalmente, para a aprovação ou não de uma linha de crédito, ou na definição de encargos financeiros de forma diferenciada. Além da utilização das estimativas dos parâmetros na predição do potencial de risco de novos candidatos a crédito, os estimadores dos parâmetros fornecem também a

informação, através da sua distribuição de probabilidade e do nível de significância, quais covariáveis estão mais associadas com o evento que está sendo modelado, ajudando na compreensão e interpretação do mesmo, no caso a inadimplência.

2.5.2 *Seleção de variáveis*

Uma vez escolhido o método de estimação dos parâmetros, o próximo passo para a construção do modelo é verificar se as covariáveis utilizadas e disponíveis para a modelagem são estatisticamente significantes com o evento modelado. Uma forma de testar a significância de uma determinada variável, é verificar se seus respectivos coeficientes são significantes. O modelo que inclui a covariável de interesse, fornece mais informação a respeito da variável resposta do que um modelo que não considera essa covariável, a ideia é que, se os valores preditos fornecidos pelo modelo com a covariável são mais precisos do que os valores preditos obtidos pelo modelo sem a covariável, há evidências de que essa covariável é importante. Então para essa seleção é utilizado o teste de wald a fim de verificar quais covariáveis são significantes para a construção do modelo, o julgamento é feito a partir do resultado do valor-p. Admitindo o nível de significância de 10%, valores-p inferiores há 10% são mantidos no modelo, pois o teste mostra que esse coeficiente poderá contribuir para o modelo.

Quando se tem variáveis *dummies* é observado que alguns coeficientes apresentam resultados de não significância, onde deve-se transformar esta classe na mesma classe de referência. Isso deve ser feito, pois o software R core team (2013) utiliza a primeira classe de cada categoria como referência e quando tem-se classes não significativas implica que não há diferença estatística dentre as classes desta categorias, podendo assim transforma-las em uma só classe. Portanto, encontrado todos os parâmetros chaves para explicar o fenômeno da inadimplência o modelo é construído e avaliado.

2.5.3 *Odds Ratio*

Considerando inicialmente o modelo logístico linear simples em que $\pi(x)$ é apresentado na Equação (2.1), é a probabilidade de "sucesso" dado o valor x de uma variável explicativa qualquer é definida como:

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \alpha + \beta x,$$

em que α e β são parâmetros desconhecidos. Esse modelo pode, por exemplo, ser aplicado para analisar a associação entre inadimplência e a ocorrência ou não de um fator particular, por exemplo sexo. Para uma variável dicotômica assumindo valores $(x = 1)$ e $(x = 0)$, obtêm-se que o odds é dado pelas equações (2.5) e (2.6), respectivamente.

$$\frac{\pi(1)}{1-\pi(1)} = \exp\{\alpha + \beta x\}, \tag{2.5}$$

$$\frac{\pi(0)}{1-\pi(0)} = \exp\{\alpha\}. \tag{2.6}$$

A razão entre os odds em (x = 1) e (x = 0) define o odds ratio, denotado por Ψ:

$$\Psi = \frac{\frac{\pi(1)}{(1-\pi(1))}}{\frac{\pi(0)}{(1-\pi(0))}} = \frac{\pi(1)(1-\pi(0))}{\pi(0)(1-\pi(1))} = \exp\{\beta\}$$

2.6 Implementação do Modelo

Quando um grande número de solicitantes ou clientes é encaminhado manualmente ao analista de crédito para verificar suas informações e aplicar regras de política de crédito, perde-se a maior parte dos benefícios associados ao uso de modelo de risco de crédito. Por outro lado, qualquer modelo de risco de crédito tem uma área "cinzenta" onde não é possível discriminar com nível aceitável de confiança entre os clientes "bons" e os possivelmente "ruins" pois essa área é a intersecção entre os dois grupos. O principal desafio dos gestores de risco de crédito é definir os limiares (pontos de corte) mais adequados e eficientes para cada modelo de risco. Para maximizar os benefícios de um modelo de risco de crédito, o corte ideal deve ser restabelecido levando em conta o custo de erro de classificação relacionado às taxas de erro de tipo I (conceder crédito, dado que o cliente é um inadimplente) e tipo II (negar crédito, dado que o cliente é um adimplente). Ademais, acredita-se que o valor ideal de corte não possa ser encontrado sem avaliar cuidadosamente as especificidades de cada banco (como tolerância ao risco, objetivos de resultado, custos e eficiência do processo de recuperação e possíveis estratégias de marketing).

3 QUALIDADE DO AJUSTE

O uso de medidas estatísticas para avaliar e, principalmente, para comparar o poder discriminatório de modelos de discriminação é uma prática comum entre os analistas de risco de crédito. A ideia de utilizar essas medidas está tão enraizada entre os analistas que praticamente se transformam no critério de aceitação ou não de um modelo de escoragem. Isto é um sério problema, pois, apesar da importância destes indicadores, eles estão longe de serem a única forma adequada de analisar, comparar e selecionar modelos.

No Brasil, a medida provavelmente mais utilizada é o KS distância de Kolmogorov-Smirnov. Como demostrado em Tomazzela *et al* (2008), a curva ROC é mais eficiente que o KS. O uso do KS deve-se a facilidade de cálculo e de interpretação por parte dos analistas de créditos. As demais medidas utilizadas exigem cálculos mais complexos e de difícil interpretação. As medidas mais populares são KS, Curva ROC e Coeficiente de Gini.

3.1 Teste de Komolgorov-Smirnov (KS)

Em modelos de classificação, por exemplo, entre bons e maus clientes, a estatística de Kolmogorov-Smirnov (KS) é construída com o objetivo de encontrar a distância máxima entre as distribuições de probabilidade dos bons e dos maus clientes. A estatística de KS mensura a máxima separação entre a frequência relativa acumulada de maus clientes (F_m) e a frequência relativa acumulada de bons clientes (F_b), na variável escore.

Hipóteses:

$$\begin{cases} H_0: & F_m(s) = F_b(s) \\ H_1: & F_m(s) \neq F_b(s) \end{cases}$$

em que,

- $F_m(s)$ = frequência relativa acumulada de maus clientes até o escore s;
- $F_b(s)$ = frequência relativa acumulada de bons clientes até o escore s;

Estatística do teste:

$$KS = \mathrm{S}|F_m(s) - F_b(s)|$$

Os analistas que desenvolvem modelos de risco de crédito costumam basear-se em alguns valores críticos, na avaliação da eficácia dos modelos de escoragem. Esses valores referenciais variam de analista para analista. É importante comentar que valores de KS superiores a 75% são raros. Quando ocorrem, os analistas revisam o modelo em busca de redundâncias ou erros de análise que possam gerar esses valores elevados de KS. A Tabela 2 apresenta um resumo das avaliações do KS baseadas em opiniões de vários analistas, Lecumberi (2003).

Tabela 2 - Valores de referência do KS em modelos de aplicação de crédito

Valor de KS(%)	Nível de Discriminação
Abaixo de 25	Baixo
De 20 a 30	Baixo/Aceitável
De 30 a 40	Bom
De 40 a 50	Muito bom
De 50 a 60	Excelente
De 60 a 70	Valores poucos usuais

A seguir são apresentados as vantagens e desvantagens da estatística de KS.

Vantagens:

- Apresenta um bom desempenho se comparado com outros métodos, por exemplo, o método de Pearson, quando tamanho da amostra é pequeno;
- Tem poucas restrições, é simples de calcular, e o teste é muito versátil no sentido de poder ser utilizado em diversas situações.

Desvantagens:

- Há algumas situações em que o KS pode assumir um valor alto, quando o modelo não está separando realmente bem os bons clientes dos maus clientes. Isto ocorre quando: as funções de distribuição de probabilidade para os bons clientes e para os maus clientes possuem uma variância diferente, formato diferente, mas os escores tanto dos bons clientes quanto dos maus clientes estão sobrepostos.

3.2 Curva ROC

A curva ROC baseia-se em duas definições: sensibilidade e especificidade. A sensibilidade pode ser entendida como a capacidade de identificar os maus clientes, a especificidade pode ser entendida como a capacidade de identificar os bons clientes.

A sensibilidade, para um dado escore é medida pela proporção de maus classificados corretamente, ou seja, a proporção de maus cujo escore é superior ao ponto de corte. A especificidade, para um dado escore, é medida pela proporção de bons classificados corretamente, ou seja, a proporção de bons cujo o escore é menor ou igual a esse escore. O valor "1-especificidade" representa os bons classificados como maus. Em outras palavras a sensibilidade é a probabilidade de um cliente ser mau pagador, dado que realmente ele é mau e especificidade é a probabilidade do cliente ser bom pagador, dado que realmente ele é bom.

Considere a Tabela 3, onde é representado a classificação real, cruzado com a classificação estimada. A partir desta tabela, pode-se calcular os valores da sensibilidade e da especificidade:

Tabela 3 - Matriz de dupla entrada

Previsão do modelo	Situação real		Total
	Mau	Bom	
Mau	m_M	m_B	m
Bom	b_M	b_B	b
Total	M	B	n

em que,

n : número total de clientes na amostra;
b_B : número de bons clientes que foram classificados como Bons (acerto);
m_M : número de maus clientes que foram classificados como Maus (acerto);
m_B : número de bons clientes que foram classificados como Maus (erro);
b_M : número de maus clientes que foram classificados como Bons (erro);
B : número total de bons clientes na amostra;
M : número total de maus clientes na amostra;
b : número total de clientes classificados como bons na amostra;
m : número total de clientes classificados como maus na amostra.

Para o cálculo da sensibilidade e da especificidade basta tomar alguns dos valores da Tabela 3. Para a sensibilidade usa-se $\frac{m_M}{M}$ e enquanto para a especificidade usa-se $\frac{b_B}{B}$.

Da mesma forma que no KS, alguns valores da curva ROC são utilizados pelos analistas para caracterizar o poder discriminatório do modelo de risco de crédito. Por exemplo, alguns analistas consideram um modelo satisfatório um modelo cuja área sob a curva ROC, a AUROC é igual ou maior que 0,7 e excelente se essa medida for superior a 0,8.

Hosmer e Lemeshow (2000) apresentam na Tabela 4 a seguinte regra geral para avaliação do resultado da AUROC, aplicada a modelos de concessão de crédito.

Tabela 4 - Valores de referência da área da curva ROC em modelos de aplicação de crédito

Valor da AUROC	Nível de Discriminação
Abaixo de 0,7	Baixo
De 0,7 a 0,8	Aceitável
De 0,8 a 0,9	Excelente
Acima de 0,9	Excepcional

3.3 Coeficiente de Gini

O coeficiente de Gini é de certa forma equivalente à curva ROC. É o dobro da área entre a curva ROC e a linha diagonal correspondente à classificação aleatória.

$$\text{Gini} = 2 \times (\text{AUROC} - 0,5)$$

A partir do valor obtido pelo coeficiente de Gini, tem-se que Gini = 0, significa que o poder discriminador é nulo. A classificação perfeita ocorre quando Gini = 1. Em geral, modelos de escoragem apresentam um coeficiente de Gini entre 40% e 60%, enquanto que modelos comportamentais apresentam um coeficiente de Gini entre 70% e 80%.

3.4 Validação

Para validar o modelo, aconselha-se verificar o desempenho do modelo quanto sua classificação na amostra de desenvolvimento. Como, geralmente, nas amostras desenvolvimento, em que os modelos são avaliados, se conhece a resposta dos clientes em relação a sua condição de crédito, e estabelecendo critérios para classificar estes clientes em bons e maus, torna-se possível comparar a classificação obtida com a verdadeira condição creditícia dos

clientes. Para monitorar o modelo o procedimento de avaliar o desempenho é feito para uma amostra teste, portanto o objetivo é monitorar a forma de classificação do modelo nesse novo conjunto de dados. No caso da ocorrência deste monitoramento, a classificação seja boa, o modelo pode continuar em aplicação.

3.4.1 *Capacidade de acerto do modelo*

Para calcular a capacidade de acerto deve-se dar origem a uma tabela de dupla entrada, como foi apresentado na Tabela 3. A forma utilizada para estabelecer a matriz de dupla entrada, é determinar um ponto de corte no escore final dos modelos tal que, indivíduos com pontuação acima desse corte são classificados como bons, por exemplo, e abaixo desse valor como maus clientes e comparando essa classificação com a situação real de cada indivíduo. Essa matriz configura, portanto, uma tabulação cruzada entre a classificação predita através de um único ponto de corte e a condição real e conhecida de cada indivíduo, em que a diagonal principal representa as classificações corretas e valores fora dessa diagonal correspondem à erros de classificação. A seguir são listados algumas medidas de capacidade de acerto do modelo sugerido por Diniz e Louzada (2012).

- Capacidade de acerto total (CAT) $= \dfrac{b_B + m_M}{n}$
- Capacidade de acerto dos maus clientes (CAM) ou sensibilidade $= \dfrac{m_M}{M}$
- Capacidade de acerto dos bons clientes (CAB) ou especificidade $= \dfrac{b_B}{B}$
- Valor preditivo positivo (VPP) $= \dfrac{m_M}{m_M + m_B}$
- Valor preditivo negativo (VPN) $= \dfrac{b_B}{b_B + b_M}$
- Prevalência (PVL) $= \dfrac{b_M + m_M}{n}$
- Correlação de mathews (MCC) $= \dfrac{b_B m_M - b_M m_B}{\sqrt{(b_B + b_M)(b_B + m_B)(m_M + b_M)(m_M + m_B)}}$

Em que m_M, m_B, m, b_M, b_B, b, M, B e n são encontradas na Tabela 3.

A Prevalência, proporção de observações propensas à característica de interesse ou a probabilidade de uma observação apresentar a característica de interesse antes de o modelo ser ajustado, é uma medida de extrema importância, principalmente quando trata-se de eventos raros.

A Capacidade de acerto Total é também conhecida como Acurácia ou Proporção de acertos de um Modelo de Classificação. Esta medida também pode ser vista como uma média ponderada da sensibilidade e da especificidade em relação ao número de observações que apresentam ou não a característica de interesse de uma determinada população. É importante ressaltar que a acurácia não é uma medida que deve ser analisada isoladamente na escolha de um modelo, pois é influenciada pela sensibilidade, especificidade e prevalência. Além disso, dois modelos com sensibilidade e especificidade muito diferentes podem produzir valores semelhantes de acurácia, se forem aplicados a populações com prevalências muito diferentes.

Para ilustrar o efeito da prevalência na acurácia de um modelo, pode-se supor uma população que apresente 5% de seus integrantes com a característica de interesse. Se um modelo classificar todos os indivíduos como não portadores da característica, tem-se um percentual de acerto de 95%, ou seja, a acurácia é alta e o modelo é pouco informativo.

O Valor Preditivo Positivo (VPP) de um modelo é a proporção de observações representando o evento de interesse dentre os indivíduos que o modelo identificou como evento. Já o Valor Preditivo Negativo (VPN) é a proporção de indivíduos que representam o não evento dentre os identificados como não evento pelo modelo. Estas medidas devem ser interpretadas com cautela, pois sofrem a influência da prevalência populacional.

O MCC, proposto por Matthews (1975), é uma medida de desempenho que pode ser utilizada no caso de prevalências extremas. É uma adaptação do Coeficiente de Correlação de Pearson e mede o quanto as variáveis que indicam a classificação original da resposta de interesse e a que corresponde a classificação do modelo obtida por meio do ponto de corte adotado, ambas variáveis assumindo valores 0 e 1, tendem a apresentar o mesmo sinal de magnitude após serem padronizadas Baldi *et al* (2000).

O MCC retorna um valor entre -1 e 1. O valor 1 representa uma previsão perfeita, um acordo total, o valor 0 representa uma previsão completamente aleatória e -1 uma previsão inversa, ou seja, total desacordo.

4 APLICAÇÃO

A aplicação deste trabalho faz-se necessário para melhor explicar todo o processo de tratamento dos dados, análise exploratória, modelagem, avaliação do instrumento e validação dos modelos de risco de credito.

4.1 Apresentação dos dados

O Banco de dados utilizado nesta aplicação é chamado de *German Credit data*, e está disponível de forma gratuita no website *UCI Machine Learning Repository* no sítio http://archive.ics.uci.edu/ml/. O arquivo German Credit, possui 20 variáveis, sendo composta por 13 categóricas e 7 numéricas, possui 1000 observações, dividido entre 700 bons pagadores e 300 maus pagadores. Este conjunto de dados apresenta atributos numéricos convertidos a partir do conjunto de dados original fornecido pelo Professor Dr. Hofmann do Instituto de Estatística e Econometria da Universidade de Hamburgo, e tem fim pedagógico cujo propósito é difundir o estudo desses modelos de crédito. Vale salientar que bancos de dados nessa área são bastante escassos no mundo acadêmico, pois, devido à competitividade do mercado financeiro não permitir a sua divulgação, usando como motivo, o receio de ser utilizado por empresas concorrentes.

Nas Tabelas 5 e 6 é apresentada uma amostra do banco de dados. As variáveis são do tipo dicotômicas, categóricas e numéricas. As categorias foram divididas da seguinte maneira, a letra "A e o número vizinho i" representa a variável i e o "número vizinho j" representa a categoria j da variável i, por exemplo "A121 = Aij " = é a variável Q12 na categoria 1. Vale salientar que a numeração da categoria j receberá somente um algarismo, enquanto a numeração da variável i será expressa por um ou dois algarismos.

Tabela 5 - Banco de dados

Indivíduos	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10
1	A11	6	A34	A43	1169	A65	A75	4	A93	A101
2	A12	48	A32	A43	5951	A61	A73	2	A92	A101
3	A14	12	A34	A46	2096	A61	A74	2	A93	A101
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
997	A11	30	A32	A41	3857	A61	A73	4	A91	A101
998	A14	12	A32	A43	804	A61	A75	4	A93	A101
999	A11	45	A32	A43	1845	A61	A73	4	A93	A101
1000	A12	45	A34	A41	4576	A62	A71	3	A93	A101

Tabela 6 - banco de dados (Continuação)

q11	q12	q13	q14	q15	q16	q17	q18	q19	q20	q21
4	A121	67	A143	A152	2	A173	1	A192	A201	1
2	A121	22	A143	A152	1	A173	1	A191	A201	2
3	A121	49	A143	A152	1	A172	2	A191	A201	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	A122	40	A143	A152	1	A174	1	A192	A201	1
4	A123	38	A143	A152	1	A173	1	A191	A201	1
4	A124	23	A143	A153	1	A173	1	A192	A201	2
4	A123	27	A143	A152	1	A173	1	A191	A201	1

A Tabela 7 lista as descrições das variáveis de forma detalhada, contendo nas colunas: variável, tipo, classe e código. Esta tabela vem no formato original, pois não foi utilizado nenhum método estatístico para tratá-la. A forma em que o banco de dados está organizado, não é a melhor maneira para proceder com a modelagem, pois variáveis numéricas precisam ser transformadas em categóricas e as variáveis categóricas que não apresentam uma boa discretização devem passar pelo processo de fusão apresentado na Subseção 2.4.2. Após o tratamento foi proposto um novo banco de dados para realizar a modelagem. A estrutura de cada variável modificada pode ser verificada na Seção 4.3, onde cada variável foi discriminada em subseções. Para a modelagem, considera-se a variável situação final do requerente (Q21), como variável resposta, e as demais variáveis como co-variáveis que procuram explicar a probabilidade de um indivíduo ser inadimplente.

Tabela 7 - Descrição das variáveis

Variável	Tipo	Classe	Código
q1 - Estado da conta	Categórica	0 Unidade monetária	A11
		0-199 Unidade monetária	A12
		200 Unidade monetária por um ano	A13
		Sem conta corrente	A14
q2 - Tempo do credito em meses	Numérica	-	-
q3 - Histórico do credito	Categórica	Sem credito tomado	A30
		Crédito quitado devidademente	A31
		Créditos existentes pago devidamente até agora	A32
		Atraso no pagamento	A33
		Créditos existentes em outro banco	A34
q4 - Finalidade do crédito	Categórica	Finalidade de comprar um carro novo	A40
		Comprar um carro usado	A41
		Comprar móveis ou equipamentos	A42
		Comprar radio ou tv	A43
		Comprar aparelhos domésticos	A44
		Fazer reparos	A45
		Investir em educação	A46
		Investir em férias	A47
		Investir em cursos profissionalizantes	A48
		Investir em negócios	A49
		Outros	A410
q5 - Valor do crédito	Numérica	-	-
q6 - Saldo médio da conta poupança	Categórica	Menor de 100 Unidade monetária	A61
		100-500 Unidade monetária	A62
		500-1000 Unidade monetária	A63
		maior do que 1000 Unidade monetária	A64
		Desconhecido ou não possui conta poupança	A65
q7 - Tempo no emprego atual	Categórica	Desempregado	A71
		Empregado há um ano	A72
		Empregado de um a quatro anos	A73
		Empregado de quatro a sete anos	A74
		Empregado a mais de sete anos	A75
q8 - Taxa de parcelamento em%	Numérica	-	-
q9 - Sexo/Estado civil	Categórica	Masculino divorciado ou separado	A91
		Feminino disvorciado, separado ou casado	A92
		Masculino Solteiro	A93
		Masculino casado ou viúvo	A94
		Feminino solteiro	A95
q10 - Co-requerente	Categórica	Não deve a ninguém	A101
		Deve ao co-requerente	A102
		Deve ao fiador	A103
q11 - Tempo de permanência na residencia atual	Numérica	-	-
q12 - Bens/propriedades	Categórica	Imóvel	A121
		Seguro de vida	A122
		Carro ou outro	A123
		Desconhecido ou não tem propiedade	A124
q13- Idade	Numérica	-	-
q14 - Outras formas de parcelamento	Categórica	Banco	A141
		Lojas	A142
		Nenhum	A143
q15- Tipo de Habitação	Categórica	Mora em imóvel alugado	A151
		Mora em imóvel próprio	A152
		Mora de graça	A153
q16 - Quantidade de credito no banco	Numérica	-	-
q17- Trabalho	Categórica	Desempregado	A171
		Não qualificado	A172
		Empregado	A173
		Autônomo	A174
q18 -Número de dependentes	Numérica	-	-
q19- Telefone	Categórica	Não possui telefone	A191
		Possui telefone registrado no nome	A192
q20- Trabalho fora da região de origem	Categórica	Trabalha fora do país	A201
		Não trabalha fora do país	A202
q21 - Situação final requerente	Categórica	Bom cliente	1
		Mau cliente	2

4.2 Amostragem

Foi realizada a amostragem proporcional, onde dos 700 bons clientes foram retirados de forma aleatória 300 e dos 300 maus foram todos selecionados. A amostra para o desenvolvimento do modelo foi composta de 200 clientes bons e 200 clientes maus. A amostra de monitoramento do modelo foi composta por 100 clientes bons e 100 clientes maus.

4.3 Análise exploratória

A análise exploratória é muito importante, pois o modelo será tão bom quanto melhor for a qualidade dos dados quanto ao poder de discriminação. Para ter essa qualidade as variáveis foram submetidas a uma análise bem minuciosa e detalhista, com o objetivo de capturar incoerências. Para mensurar a qualidade da categorização foi utilizado o risco relativo, WOE e teste Qui-Quadrado. Para o teste Qui-quadrado é adotado um nível de significância de $\alpha = 5\%$.

4.3.1 *Estado da conta (Q1)*

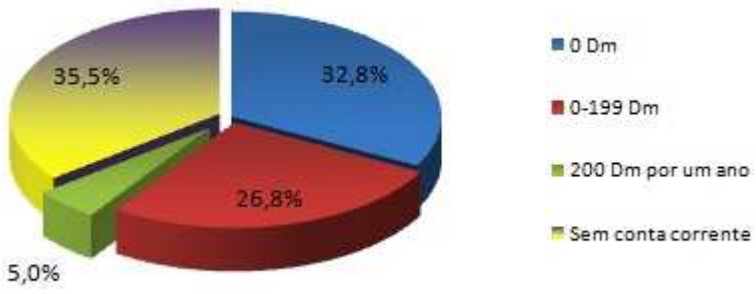
A Tabela 8 mostra o estado da conta em função da inadimplência, onde os valores do risco relativo e WOE têm medidas bem distantes entre si. Em vista disso, pode-se dizer que a categorização é boa. O teste de Qui-Quadrado verifica um valor-p inferior a 0,001, indicando que existe relação entre o estado da conta e inadimplência, assim pode-se dizer que o estado da conta pode auxiliar na explicação a inadimplência dos clientes. Em vista desses resultados favoráveis pode-se classificar essa variável como forte candidata a entrar para o modelo.

Tabela 8 - Estado da conta

Estado da conta	***Freq.***	***%***	**Bom(%)**	**Mau(%)**	***Risco relativo***	***WOE***
0 Um	131	32,75	9,75	23,00	0,42	- 0,37
0 ⊣ 199 Um	107	26,75	10,25	16,50	0,62	- 0,21
200 Um por um ano	20	5,00	2,75	2,25	1,22	0,09
Sem conta corrente	142	35,5	27,25	8,25	3,30	0,52

A Figura 1, mostra a proporção de cada categoria na variável estado da conta.

Figura 1 - Estado da conta



4.3.2 *Tempo de crédito em meses (Q2)*

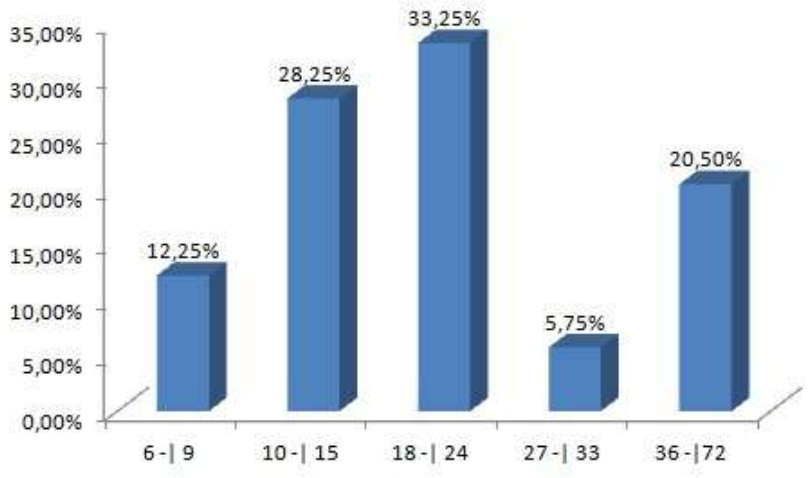
A Tabela 9 descreve o tempo de crédito em função da inadimplência. Essa variável foi discretizada com o objetivo de melhorar a variável para a análise. O resultado da discretização foi satisfatório, pois as medidas do risco relativo e WOE estão bem distantes entre si. O resultado do teste Qui-Quadrado calculou um valor-p inferior a 0,001, indicando que existe relação entre as variáveis e que o tempo de crédito pode auxiliar na explicação da inadimplência dos clientes. Dado esses fatores, é identificado que está variável é uma forte candidata a entrar para o modelo.

Tabela 9 - Tempo de crédito

Tempo de crédito em meses	**Freq.**	**%**	**Bom(%)**	**Mau(%)**	**Risco relativo**	**WOE**
6 ⊣ 9	49	12,25	8,25	4,00	2,063	0,314
10 ⊣ 15	113	28,25	17,25	11,00	1,568	0,195
16 ⊣ 24	133	33,25	17,00	16,25	1,046	0,020
25 ⊣ 33	23	5,75	2,25	3,50	0,643	-0,192
34 ⊣ 72	82	20,5	5,25	15,25	0,344	-0,463

A Figura 2, mostra a proporção de cada categoria na variável tempo de crédito.

Figura 2 - Tempo de crédito



4.3.3 *Histórico do crédito (Q3)*

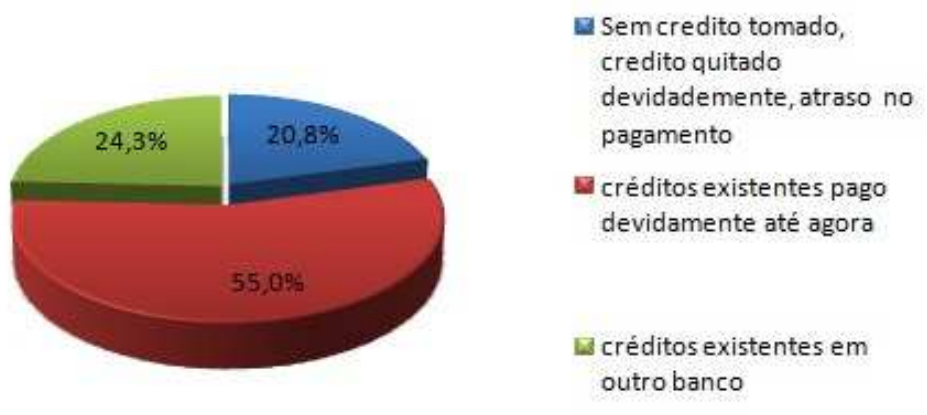
A Tabela 10 descreve o histórico da conta em função da inadimplência. A junção de categorias foi feito com o objetivo de tornar melhor a categorização da variável, com isto, as medidas do risco relativo e WOE são distantes entre si. No teste Qui-Quadrado, foi obtido um valor-p inferior a 0,001, sugerindo que existe associação entre as variáveis e que o histórico da conta pode auxiliar na explicação da inadimplência. Portanto, está variável é uma forte candidata a entrar para o modelo.

Tabela 10 - Histórico do crédito

Histórico do crédito	***Freq.***	***%***	**Bom(%)**	**Mau(%)**	***Risco relativo***	***WOE***
Sem credito tomado, credito quitado devidademente e atraso no pagamento	83	20,75	6,25	14,5	0,431	-0,365
créditos existentes pago devidamente até agora	220	55,00	26,00	29,0	0,897	-0,047
créditos existentes em outro banco	97	24,25	17,75	6,5	2,731	0,436

A Figura 3 mostra a proporção das categorias na variável histórico da conta.

Figura 3 - Histórico do crédito



4.3.4 *Finalidade do crédito (Q4)*

A Tabela 11 descreve a finalidade do crédito em função da inadimplência. A categorização apresentou valores de risco relativo e WOE distantes dos valores nos níveis, portanto a categorização é boa. No teste Qui-Quadrado, foi calculado um valor-p de 0,091, sugerindo que existe relação entre as variáveis e que a finalidade do crédito pode auxiliar na explicação da inadimplência. Dado esses fatores, é identificado que está variável é uma forte candidata a entrar para o modelo.

Tabela 11 - Finalidade do crédito

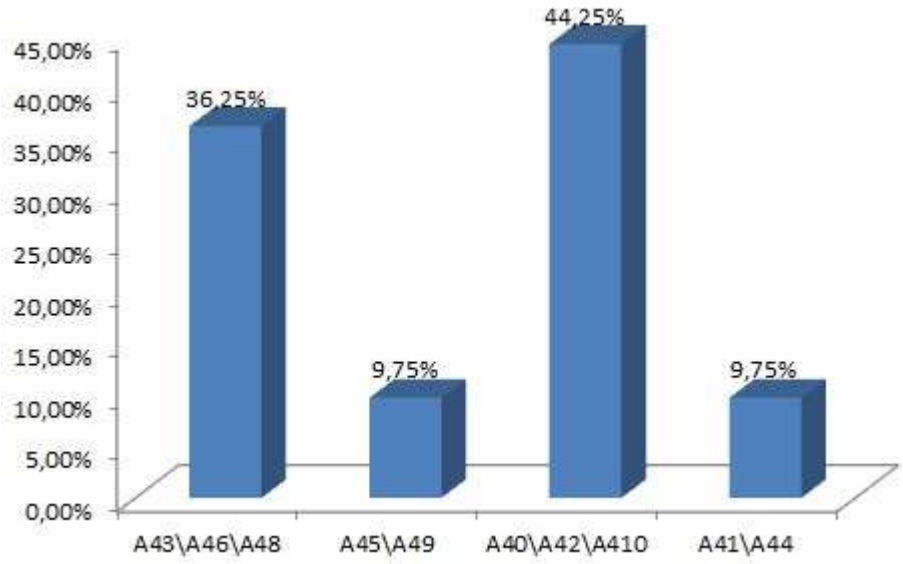
Finalidade do crédito	**Freq.**	**%**	**Bom(%)**	**Mau(%)**	**Risco relativo**	**WOE**
A43, A46 e A48	145	36,25	19,50	67,00	0,291	-0,536
A45 e A49	39	9,75	3,50	6,25	0,560	-0,252
A40, A42 e A410	177	44,25	21,00	23,25	0,903	-0,044
A41 e A44	39	9,75	6,00	3,75	1,600	0,204

Tabela 12 - Descrição da variável finalidade do crédito

Categoria	Código
Finalidade de comprar um carro novo	A40
Finalidade de comprar um carro usado	A41
Finalidade de comprar móveis ou equipamentos	A42
Finalidade de comprar radio ou tv	A43
Finalidade de comprar aparelhos domésticos	A44
Finalidade em fazer reparos	A45
Finalidade de investir em educação	A46
Finalidade de investir em férias	A47
Finalidade de investir em cursos profissionalizantes	A48
Finalidade de investir em negócios	A49
Outros	A410

A Figura 4 mostra a proporção das categorias na variável finalidade do crédito.

Figura 4 - Finalidade do crédito



4.3.5 *Valor do crédito (Q5)*

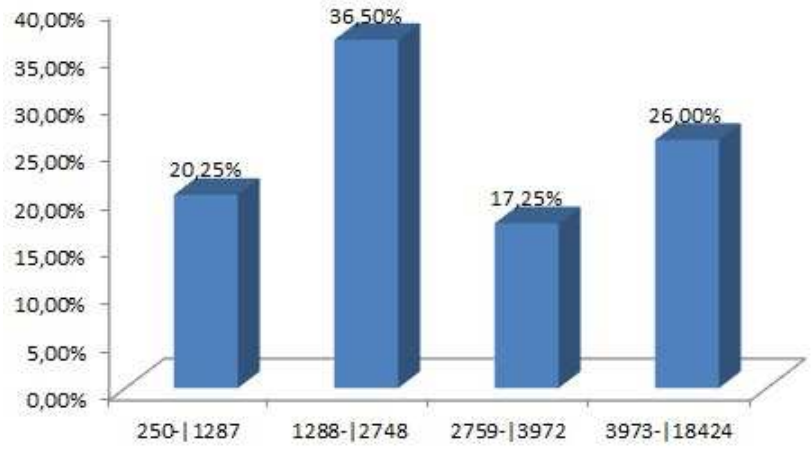
A Tabela 13 descreve o valor do crédito em função da inadimplência. Essa variável foi discretizada e o resultado da discretização foi satisfatório, pois às medidas do risco relativo e WOE estão bem distantes entre si, vale ressaltar que a medida que o risco decresce no último nível, isso indica que quanto maior o crédito tomado maior o risco do indivíduo tornar-se um mau cliente. O resultado do teste Qui-Quadrado calculou um valor-p inferior a 0,001, indicando que existe relação entre as variáveis e que o valor do crédito pode auxiliar na explicação da inadimplência dos clientes. Dado esses fatores, é identificado que está variável é uma forte candidata a entrar para o modelo.

Tabela 13 - Valor do crédito

Valor do crédito	Freq.	%	Bom(%)	Mau(%)	Risco relatico	WOE
250 ⊣ 1287	81	20,25	10,0	10,3	0,976	-0,011
1288 ⊣ 2748	146	36,50	21,0	15,5	1,355	0,132
2759 ⊣ 3972	69	17,25	11,0	6,3	1,760	0,246
3973 ⊣ 18424	104	26,00	8,0	18,0	0,444	-0,352

A Figura 5 mostra a proporção das categorias na variável valor do crédito.

Figura 5 - Valor do crédito



4.3.6 *Saldo médio da conta poupança (Q6)*

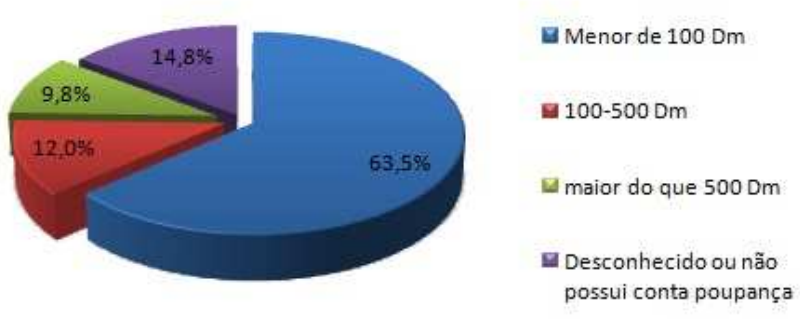
A Tabela 14 descreve o saldo da conta poupança em função da inadimplência. A categorização foi boa, pois as medidas para o risco relativo e WOE estão distantes entre si e o valor-p para o teste Qui-Quadrado é inferior a 0,001, indicando que a hipótese de não associação é rejeitada. Devido esses resultados observados, toma-se essa variável como uma forte candidata para estar presente no modelo.

Tabela 14 - Saldo médio da conta poupança

Saldo médio da conta poupança	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
Menor de 100 Um	254	63,50	26,75	36,75	0,728	-0,138
100-500 Um	48	12,00	6,00	6,00	1,000	0,000
maior do que 500 Um	39	9,75	7,75	2,00	3,875	0,588
Desconhecido ou não possui conta poupança	59	14,75	9,50	5,25	1,810	0,258

A Figura 6 abaixo apresenta a proporção nas categorias na variável saldo médio da conta poupança.

Figura 6 - Saldo médio da conta poupança



4.3.7 *Tempo no emprego atual (Q7)*

A Tabela 15 descreve o tempo no emprego atual em função da inadimplência. A categorização da variável é boa devido às medidas do risco relativo e WOE que estão distantes entre si. No teste Qui-Quadrado, foi calculado um valor-p de 0,022, esse resultado implica que existe associação entre as variáveis. Dado esses fatores, é identificado que está variável é uma forte candidata a compor o modelo.

Tabela 15 - Tempo no emprego atual

Tempo no emprego atual	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
Desempregado e empregado há um ano	97	24,25	9,50	14,75	0,64	-0,19
Empregado de um a quatro anos	135	33,75	16,00	17,75	0,90	-0,05
Empregado de quatro a sete anos	65	16,25	9,25	7,00	1,32	0,12
Empregado a mais de sete anos	103	25,75	15,25	10,5	1,45	0,16

A Figura 7 apresenta a proporção das categorias na variável tempo no emprego atual.

Figura 7 - Tempo no emprego atual

4.3.8 *Taxa de parcelamento em porcentagem (Q8)*

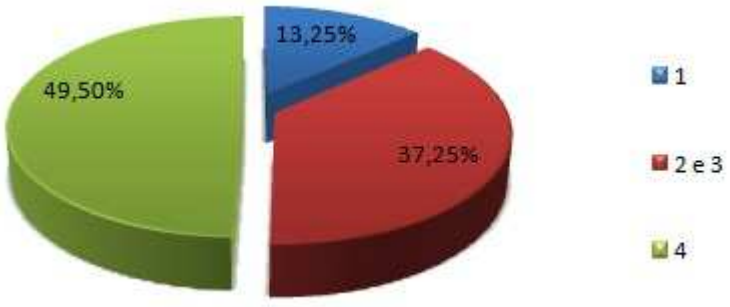
A Tabela 16 descreve a taxa de parcelamento em função da inadimplência. As medidas do risco relativo e WOE apresentam medidas distantes entre os níveis da variável categórica. No teste Qui-Quadrado, foi calculado um valor-p de 0,157, determinando que não existe associação entre as variáveis. Baseado nesses fatores, não utilizaremos esta variável na construção do modelo.

Tabela 16 - Taxa de parcelamento em porcentagem

Taxa de parcelamento em porcentagem	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
1	53	13,25	7,75	5,5	1,41	0,15
2 e 3	149	37,25	19,75	17,5	1,13	0,05
4	198	49,50	22,50	27,0	0,83	-0,08

A Figura 8 apresenta a proporção das categorias na variável taxa de parcelamento em porcentagem .

Figura 8 - Taxa de parcelamento em porcentagem



4.3.9 *Sexo/Estado civil (Q9)*

A Tabela 17 descreve o Sexo/Estado civil em função da inadimplência. A junção de categorias melhorou a categorização da variável e assim as medidas do risco relativo e WOE se tornaram mais significativas, pois as medidas estão distantes entre si. No teste Qui-Quadrado, foi calculado um valor-p de 0,016, determinando que existe associação entre as variáveis. Com base nesses resultados, julga-se está variável como uma forte candidata a compor o modelo.

Tabela 17 - Sexo/Estado civil

Sexo/Estado civil	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
Masculino divorciado ou separado, Feminino disvorciado, separado ou casado, e Masculino casado ou viúvo	182	45,5	19,75	25,75	0,766	-0,115
Masculino Solteiro	218	54,5	30,25	24,25	1,247	0,096

A Figura 9 apresenta a proporção das categorias na variável sexo/estado civil.

Figura 9 - Sexo/Estado civil

4.3.10 *Co-requerente (Q10)*

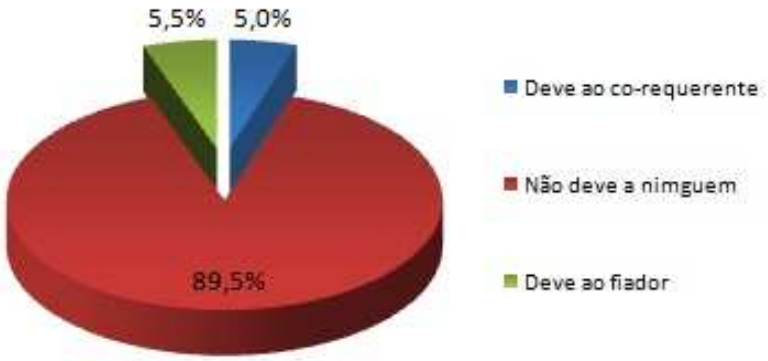
A Tabela 18 descreve o Co-requerente em função da inadimplência. A categorização é boa, pois as medidas para o risco relativo e WOE estão distantes entre si e o valor-p para o teste Qui-Quadrado é de 0,041, indicando que existe associação entre as variáveis e o saldo da conta poupança pode auxiliar na explicação da inadimplência. Devido esses resultados observados, toma-se essa variável como uma forte candidata para estar presente no modelo.

Tabela 18 - Co-requerente

Co-requerente	Freq.	%	Bom(%)	Mau(%)	Risco Relativo	WOE
Deve ao co-requerente	20	5,0	1,75	3,25	0,538	-0,269
Não deve a nimguem	358	89,5	44,25	45,25	0,978	-0,010
Deve ao fiador	22	5,5	4,00	1,50	2,667	0,426

A Figura 10 apresenta a proporção das categorias na variável co-requerente. Pode-se notar que a categoria com maior frequência é “não deve a ninguém”, por um lado esse resultado é satisfatório, pois a proporção de indivíduos que não deve é baixa, todavia é bom que a frequência seja razoável para cada categoria para poder discriminar bem todos os grupos.

Figura 10 - Co-requerente



4.3.11 *Tempo de permanência na residência atual (Q11)*

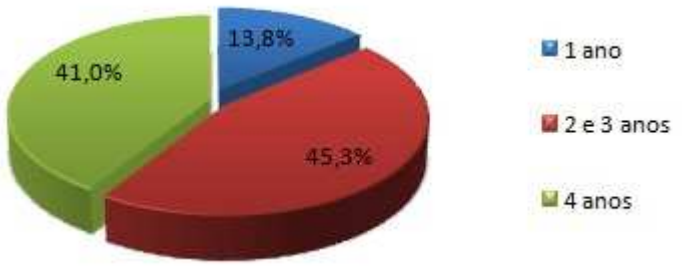
A Tabela 19 descreve o tempo de permanência na residência atual em função da inadimplência. A maneira como estava categorizada não credenciava a variável para a utilização no modelo, isso é decorente do resultado observado no risco relativo, onde os valores estavam próximos. A partir disto foi feito a união de categorias e não houve melhora na categorização da variável, pois como se pode observar nas medidas do risco relativo e WOE que a distância são próximas entre os níveis. No teste Qui-Quadrado, foi calculado um valor-p de 0,411, determinando que não existe associação entre as variáveis. Dado esses fatores, está variável não será utilizada na construção do modelo.

Tabela 19 - Tempo de permanência na residencia atual

Tempo de permanência na residencia atual	**Freq.**	**%**	**Bom(%)**	**Mau(%)**	**Risco relativo**	**WOE**
1 ano	55	13,75	7,50	6,25	1,20	0,08
2 e 3 anos	181	45,25	21,00	24,25	0,87	-0,06
4 anos	164	41,00	21,50	19,50	1,10	0,04

A Figura 11 apresenta a proporção das categorias na variável tempo de permanência na residência atual.

Figura 11 - Tempo de permanência na residencia atual



4.3.12 *Bens/Propriedades (Q12)*

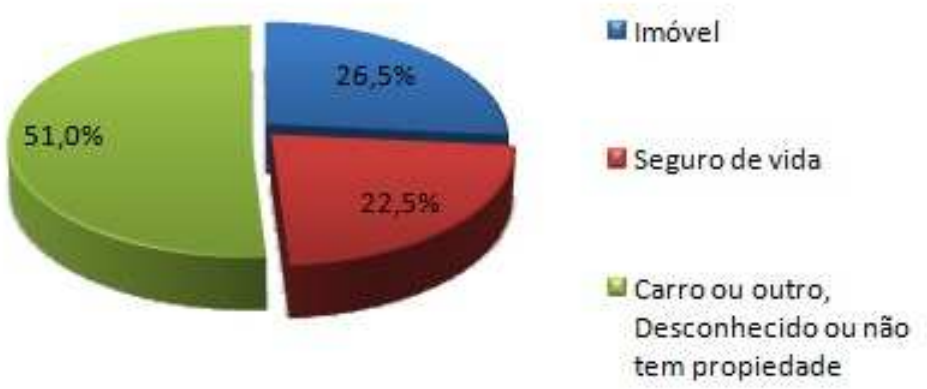
A Tabela 20 mostra Bens/propriedades em função da inadimplência, onde os valores do risco relativo e WOE têm medidas bem distantes entre si. Em vista disso, pode-se dizer que a categorização é boa. O teste de Qui-Quadrado verifica um valor-p de 0,085, indicando que existe relação entre o Bens/propriedades e inadimplência, assim pode-se dizer que o Bens/propriedades pode auxiliar na explicação da inadimplência dos clientes. Em vista desses resultados favoráveis pode-se classificar essa variável como forte candidata a compor o modelo.

Tabela 20 - Bens/propriedades

Bens/propriedades	Freq.	%	Bom(%)	Mau(%)	Risco Relativo	WOE
Imóvel	106	26,50	15,00	11,50	1,304	0,115
Seguro de vida	90	22,50	12,25	10,25	1,195	0,077
Carro ou outro, desconhecido ou não tem propiedade	204	51,00	22,75	28,25	0,805	-0,094

A Figura 12 apresenta a proporção das categorias na variável bens/propriedades.

Figura 12 - Bens/propriedades



4.3.13 *Idade (Q13)*

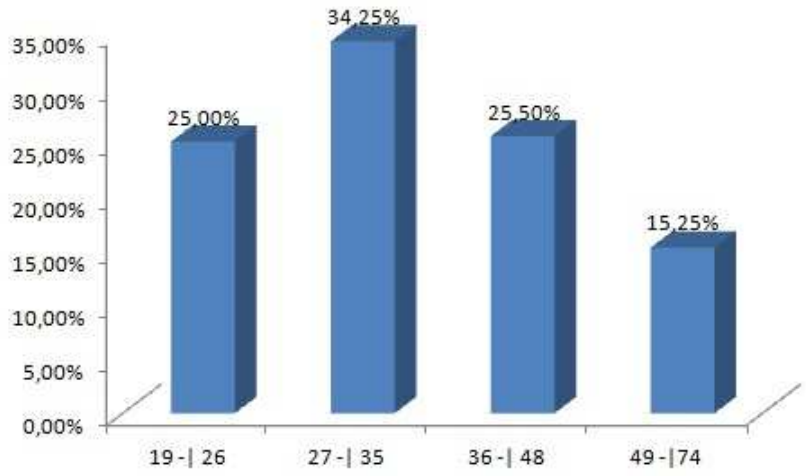
A Tabela 21 descreve a idade em função da inadimplência. Essa variável foi discretizada e o resultado da discretização foi satisfatório, pois as medidas do risco relativo e WOE estão bem distante entre os níveis. O resultado do teste Qui-Quadrado calculou um valor-p inferior a 0,012, indicando que existe relação entre as variáveis e que a idade pode auxiliar na explicação da inadimplência dos clientes. Baseado nesses informações, é identificado que está variável é uma forte candidata para estar presente no modelo.

Tabela 21 - Idade

Idade	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
19 ⊣ 26	100	25,00	18,75	6,25	3,000	0,477
27 ⊣ 35	137	34,25	18,25	16,00	1,141	0,057
36 ⊣ 48	102	25,50	9,75	15,75	0,619	-0,208
49 ⊣ 74	61	15,25	3,25	12,00	0,271	-0,567

A Figura 13 apresenta a proporção das categorias na variável idade. Pode-se notar que elas têm frequências razoáveis para a aplicação. Vale ressaltar a classe de 27 a 35 anos de idade, a qual apresentou maior frequência, correspondendo a 34,25%.

Figura 13 - Idade



4.3.14 *Outras formas de parcelamentos (Q14)*

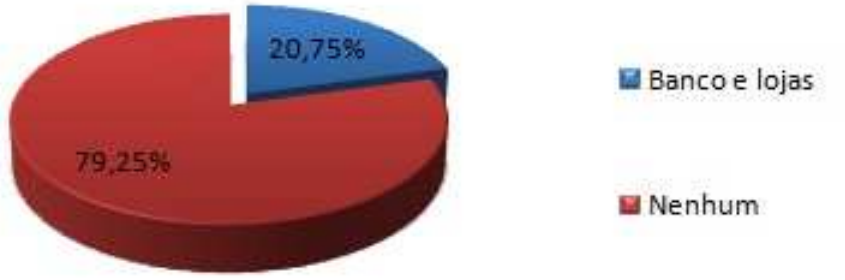
A Tabela 22 descreve outras formas de parcelamentos em função da inadimplência. A categorização é boa, pois as medidas para o risco relativo e WOE estão distantes entre si e o valor-p para o teste Qui-Quadrado é 0,010, indicando que existe associação entre as variáveis e que "outras formas de parcelamentos" pode auxiliar na explicação da inadimplência. Devido esses resultados, toma-se essa variável como uma forte candidata para estar presente no modelo.

Tabela 22 - Outras formas de parcelamentos

Outras formas de parcelamento	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
Banco e lojas	83	20,75	7,75	13,00	0,60	-0,22
Nenhum	317	79,3	42,25	37,00	1,14	0,06

A Figura 14 apresenta a proporção das categorias na variável outras formas de parcelamentos.

Figura 14 - Outras formas de parcelamentos



4.3.15 *Tipo de habitação (Q15)*

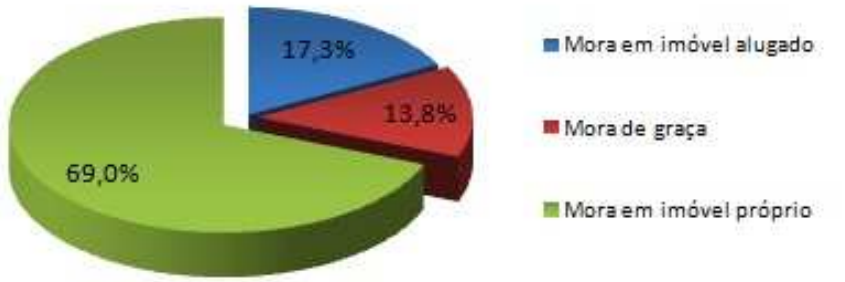
A Tabela 23 descreve o estado da habitação em função da inadimplência. A categorização da variável é boa, pois ao avaliar a medida do risco relativo e WOE, observa-se valores distantes entre os níveis. O teste Qui-Quadrado calculou um valor-p de 0,028, indicando que existe relação entre as variáveis. Logo essa variável é uma boa candidata a integrar o modelo.

Tabela 23 - Tipo de Habitação

Tipo de Habitação	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
Mora em imóvel alugado	69	17,25	6,50	10,75	0,60	-0,22
Mora de graça	55	13,75	6,00	7,75	0,77	-0,11
Mora em imóvel próprio	276	69,00	37,50	31,50	1,19	0,08

A Figura 15 apresenta a proporção das categorias na variável estado de habitação. Pode-se notar que elas têm frequências razoáveis para a aplicação. E que aproximadamente 70% dos indivíduos declararam morar em imóvel próprio.

Figura 15 - Tipo de Habitação



4.3.16 *Quatidade de crédito no banco (Q16)*

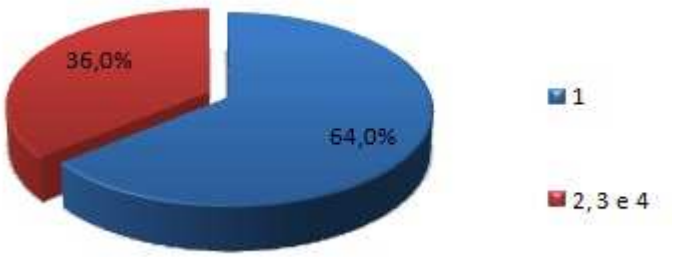
A Tabela 24 descreve a quantidade de crédito no banco em função da inadimplência. Essa variável foi discretizada e o resultado da discretização foi satisfatório, pois as medidas do risco relativo e WOE estão bem distante entre si. O teste Qui-Quadrado apresentou um valor-p de 0,037, indicando que existe relação entre as variáveis. Dado esses fatores, é identificado que está variável é uma forte candidata a participar do modelo.

Tabela 24 - Quatidade de crédito no banco

Quantidade de credito no banco	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
1	256	64	29,50	34,50	0,86	-0,07
2, 3 e 4	144	36	20,50	15,50	1,32	0,12

A Figura 16 apresenta a proporção das categorias na variável quantidade de crédito no banco.

Figura 16 - Quatidade de crédito no banco



4.3.17 *Trabalho (Q17)*

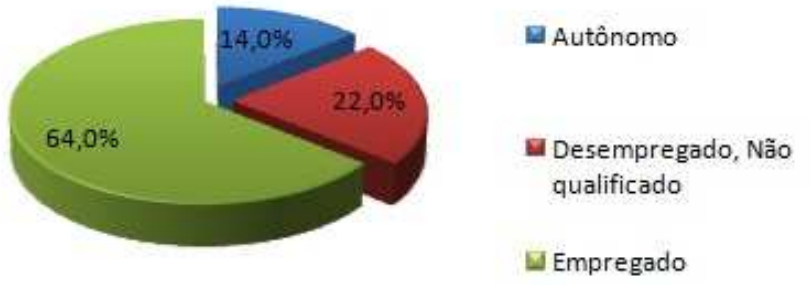
A Tabela 25 descreve o estado da habitação em função da inadimplência. A categorização da variável é boa, pois ao avaliar a medida do risco relativo e WOE, têm valores distantes entre si. O teste Qui-Quadrado calculou um valor-p de 0,084, indicando que existe relação entre as variáveis e que o trabalho pode auxiliar na explicação da inadimplência. Logo essa variável é uma boa candidata a integrar o modelo.

Tabela 25 - Trabalho

Trabalho	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
Autônomo	56	14,00	5,25	8,75	0,60	-0,22
Desempregado, Não qualificado	88	22,00	10,50	11,50	0,91	-0,04
Empregado	256	64,00	34,25	29,75	1,15	0,06

A Figura 17 apresenta a proporção das categorias na variável trabalho.

Figura 17 - Trabalho



4.3.18 *Número de dependentes (Q18)*

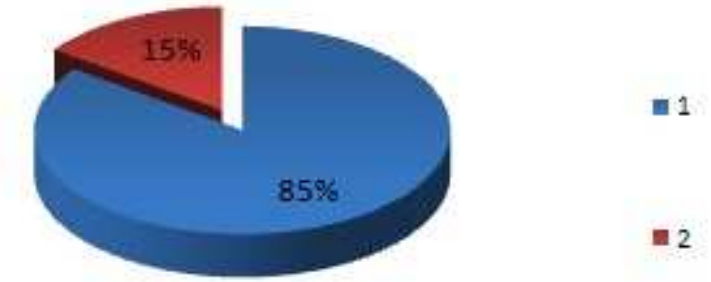
A Tabela 26 descreve o número de dependentes em função da inadimplência. Observa-se que as medidas do risco relativo e WOE, estão relativamente distantes entre si, indicando que a categorização é boa, no entanto como o valor-p calculado pelo o teste Qui-Quadrado foi de 0,575, indicando que as duas variáveis não tem relação e a parti disto o número de dependentes não pode explica a inadimplência dos clientes. Em vista desses resultados, essa variável não fará parte do modelo.

Tabela 26 - Número de dependentes

Número de dependentes	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
1	340	85	43	42	1,023	0,010
2	60	15	7	8	0,875	-0,057

A Figura 18 apresenta a proporção das categorias na variável número de dependente. As frequências nos níveis são razoáveis, com 85% dos indíviduos tem dois filhos.

Figura 18 - Número de dependentes



4.3.19 *Telefone (Q19)*

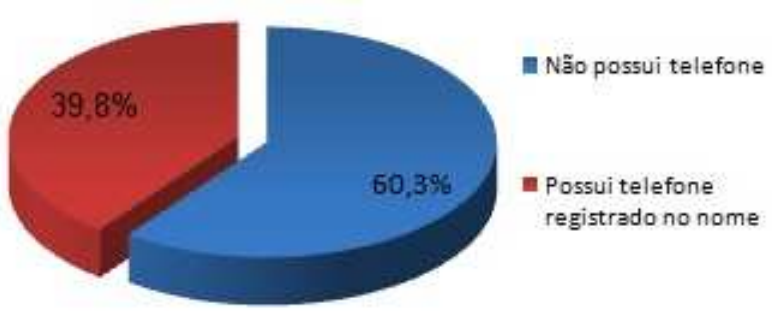
A Tabela 27 descreve a variável telefone em função da inadimplência. Observa-se que as medidas do risco relativo e WOE, estão bem próximas entre os demais níveis, indicando que a categorização não é boa e a união das mesmas não é interessante, pois a resposta tem opiniões diferentes. Pelo valor-p calculado para o teste Qui-Quadrado de 0,474, indicando que as duas variáveis não tem associação e a parti disto o número de clientes que usam telefone pode não auxiliar na explicação da inadimplência dos clientes. Em vista desses resultados, essa variável não fará parte do modelo.

Tabela 27 - Telefone

Telefone	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
Não possui telefone	241	60,3	29,25	31	0,943	-0,025
Possui telefone registrado no nome	159	39,8	20,75	19	1,092	0,038

A Figura 19 apresenta a proporção das categorias na variável telefone.

Figura 19 - Telefone



4.3.20 *Trabalho fora da região de origem (Q20)*

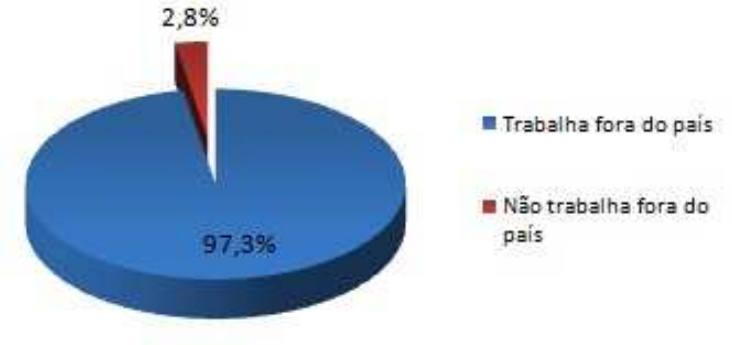
A Tabela 28 descreve o trabalho fora da região de origem em função da inadimplência. Observa-se que as medidas do risco relativo e WOE, estão distantes entre si, indicando que a categorização é boa, mas deve ser ressaltado que tem casela com frequência inferior a cinco e devido à baixa frequência o poder de discriminar a variável seja influenciado. A união dos dois níveis não é interessante, pois as classes desta variável tem bastante divergência de opinião. Não é correto tomar como referência o valor-p do teste Qui-Quadrado, pois para a utilização dele as suposições não foram satisfeitas. A conclusão é que número de pessoas que trabalham fora da região de origem pode não auxiliar na explicação da inadimplência dos clientes. Em vista desses resultados, essa variável não fará parte da construção do modelo.

Tabela 28 - Trabalho estrangeiro

Trabalho estrangeiro	Freq.	%	Bom(%)	Mau(%)	Risco relativo	WOE
Trabalha fora do país	389	97,3	47,75	49,50	0,96	-0,02
Não trabalha fora do país	11	2,8	2,25	0,50	4,50	0,65

A Figura 20 apresenta as proporções bastantes discrepantes entre as categorias. A classe "Trabalha fora do país" possui baixa frequência, logo essa variável não é interessante para a modelagem.

Figura 20 - Trabalho fora da região de origem



4.3.21 *Pré-Seleção de variáveis*

Após a análise exploratória em todas as variáveis, pode-se notar a priori quais variáveis participarão da contrução do modelo que tentará explicar a probabilidade de um determinado proponente ser inadimplente. Essa decisão dar-se à através da análise do resultado calculado no teste Qui-Quadrado, pois a partir destes resultados tem-se a classificação de quais variáveis tem associação com a variável de interesse. As variáveis que não apresentaram associação foram: Taxa de parcelamento (Q8), Tempo de permanência na residência atual (Q11), Número de dependentes (Q18), Telefone (Q19) e Trabalho fora da região de origem (Q20). Essas variáveis não irão compor o modelo, pois apresentam valores-p maiores que o nível de significância $\alpha = 10\%$.

4.4 Modelo de risco crédito

4.4.1 *Ajuste do modelo*

O modelo completo foi ajustado a partir da triagem feita na análise exploratória. Variáveis que não foram significativas pelo teste Qui-quadrado, foram excluídas do pro-

cesso de construção do modelo. A Tabela 29 apresenta os coeficientes estimados deste modelo.

Tabela 29 - Modelo completo

Parâmetros	Estimativa	Erro padrão	wald	valor-p
(Intercept)	3,3176	1,0739	3,09	0,0020***
A12	-0,5260	0,3562	-1,48	0,1398
A13	-0,9916	0,6309	-1,57	0,1160
A14	-2,2827	0,3814	-5,98	0,0000***
A22	0,3551	0,4727	0,75	0,4524
A23	0,7613	0,4877	1,56	0,1185
A24	0,7787	0,6953	1,12	0,2627
A25	1,4093	0,6241	2,26	0,0239**
A32	-0,7091	0,4057	-1,75	0,0805*
A34	-1,8829	0,4682	-4,02	0,0001***
A42	-0,9878	0,5440	-1,82	0,0694*
A43	0,1013	0,3212	0,32	0,7526
A44	0,0051	0,5075	0,01	0,9920
A52	0,1841	0,3994	0,46	0,6448
A53	-0,5457	0,5125	-1,06	0,2869
A54	1,0536	0,6076	1,73	0,0829*
A62	-0,4881	0,4456	-1,10	0,2734
A63	-1,3246	0,5326	-2,49	0,0129**
A65	-0,5495	0,4173	-1,32	0,1879
A73	-0,2176	0,3739	-0,58	0,5606
A74	-0,8112	0,4474	-1,81	0,0698*
A75	-0,0956	0,4382	-0,22	0,8273
A823	0,4157	0,4519	0,92	0,3576
A84	0,9487	0,4798	1,98	0,0480**
A93	-0,7828	0,3177	-2,46	0,0137**
A102	-0,2135	0,6067	-0,35	0,7249
A103	-1,6433	0,8686	-1,89	0,0585*
A112-3	1,1954	0,4245	2,82	0,0049***
A114	0,3988	0,4461	0,89	0,3713
A132	-0,8033	0,3827	-2,10	0,0358**
A133	-0,7499	0,4195	-1,79	0,0739*
A134	-0,7272	0,5198	-1,40	0,1618
A143	-0,9249	0,3401	-2,72	0,0065***
A152	-0,5799	0,5791	-1,00	0,3167
A153	-0,8473	0,4290	-1,97	0,0483**
A162-3-4	0,1673	0,3961	0,42	0,6728
A173	-0,8172	0,3684	-2,22	0,0265**
A174	-0,6236	0,5239	-1,19	0,2339

em que, *0,10; **0,05; ***<0,01, representam respectivamente, significância a 10%, 5% e 1%.

Pode-se notar na Tabela 29 que existem alguns parâmentros que não são significativos a um nível de 10%, vale ressaltar que o software R core team (2013) faz o ajuste do modelo utilizando caselas de referência, isso implica que o modelo toma para o intercepto todos as primeiras categorias de cada variável que não são significantes para o modelo, por exemplo a variável Q1 apresenta as categorias A11, A12, A13 e A14, pode-se notar que a categoria A11 esta no intercepto com as demais categorias de todas as variáveis, as categorias A12 e A13 não foram significantes e a categoria A14 é significante. O fato das categorias A12 e A13 não serem significantes, implica que elas não tem diferença estatística entre si, onde seus coeficientes não terão nenhuma contribuição para o modelo e portanto podem e devem ser renomeadas como A11. A11 será uma categoria que receberá respostas dela

mesma, A12 e A13. Pode-se concluir com o exemplo que categorias não significativas serão renomeadas com o nome da primeira categoria de cada variável.

A Tabela 30 mostra a renomeação, vale salientar que houve variáveis que não são significativas para o modelagem, essas variáveis foram excluídas. As variáveis selecionadas foram: Estado da conta (Q1), Tempo de crédito em meses (Q2), Histórico de crédito (Q3), Finalidade do crédito (Q4), Valor do crédito (Q5), Saldo médio da conta poupança (Q6), Tempo no emprego atual (Q7), Sexo/Estado civil (Q9), Co-requerente (Q10) e Outras formas de parcelamento (Q14).

Tabela 30 - Modelo final

Parâmetros	Estimativa	Erro padrão	wald	Odds Ratio	valor-p
(Intercepto)	1,9604	0,3569	5,49	7,10	0,0000***
A14	-1,6867	0,2741	-6,15	0,18	0,0000***
A25	0,7811	0,3758	2,08	2,18	0,0377**
A34	-1,0271	0,2997	-3,43	0,35	0,0006***
A42	-1,0431	0,4460	-2,34	0,35	0,0193**
A54	0,9336	0,3520	2,65	2,54	0,0080***
A63	-1,2356	0,4785	-2,58	0,29	0,0098***
A74	-0,6632	0,3359	-1,97	0,51	0,0484**
A93	-0,6276	0,2470	-2,54	0,53	0,0111**
A103	-1,3028	0,5470	-2,38	0,27	0,0172**
A143	-1,0393	0,3115	-3,34	0,35	0,0008***

em que, *0,10; **0,05; ***<0,01, representam respectivamente, significância a 10%, 5% e 1%.

Vale salientar que a variável Tempo de permanência na residência atual (Q11) teve sua estimativa significativa mas foi excluida devido a falta de lógica na interpretação. Há retirada dessa variável influenciou de forma negativa nas variáveis Taxa de parcelamento (Q8) e idade (Q13), pois elas estavam quase no limite para serem rejeitadas e o fato da retirada da variável Q11 do modelo aumentou o valor-p dessas duas variáveis, desqualificando-as para a modelagem. Pode-se notar na Tabela 30 que o ajuste final apresenta todas as variáveis significativas para o modelo, os próximos passos são avaliar a qualidade do modelo, validar o modelo quanto a sua previsão dos escores de cada indivíduo e interpretar cada parâmetro.

4.4.2 *Avaliação da qualidade do ajuste*

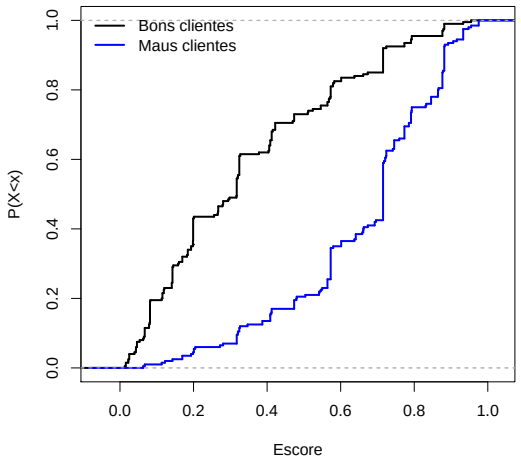
A avaliação da qualidade do ajuste utilizando as medidas, Kolmogorov-Smirnov, Curva ROC e Coeficiente de Gini, tem a importância de avaliar a qualidade do ajuste e a

partir disto definir se o modelo é adequado ou não. Vale salientar que essas medidas não são as únicas empregadas na validação do modelo.

4.4.2.1 Estatística de KS

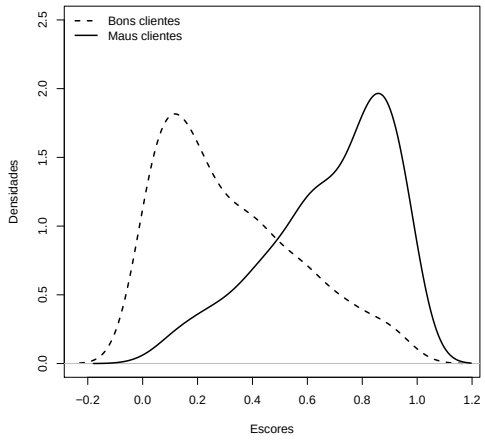
A estatística KS mensura quão distantes são as duas distribuições acumuladas, a distribuição dos bons clientes com a distribuição dos maus clientes. A estatística calculada $KS = 54\%$, indica que o afastamento da distribuição é excelente com base na Tabela 2, apresentada na página 22. O fato de ser excelente implica que o modelo discrimina bem os bons clientes dos maus clientes, por isso a importância dessa estatística. Pode-se notar esse afastamento através da Figura 21 das distribuições de probabilidades acumuladas e da Figura 22 com as distribuições densidades de probabilidade.

Figura 21 - Distribuição acumulada



Pela Figura 22, pode-se notar o ponto de intersecção entre as duas densidades. Esse ponto de intersecção é chamado de ponto corte e é aproximadamente 0,5. Esse ponto é o ponto que melhor divide as duas distribuições de probabilidade. Portanto é o melhor ponto para discriminar os dois grupos. Esse ponto tem por objetivo classificar os clientes de acordo com seus respectivos escores, pois quem tiver escore inferior a 0,5 será classificado como bom cliente e caso contrário serão classificados como mau cliente. Na realidade o ponto de corte não necessariamente é o ponto de interseção entre os dois grupos, pois de acordo com a política de crédito das financeiras esse ponto pode ser alterado a fim de conseguir maximizar o lucro da empresa, por exemplo.

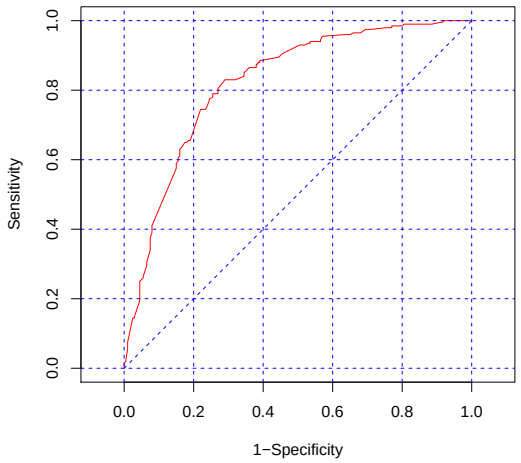
Figura 22 - Densidade



4.4.2.2 Curva ROC

A Figura 23 descreve a taxa de verdadeiro positivo em relação a taxa de falso positivo. Quando se tem a curva bem próxima do canto superior esquerdo do gráfico, diz-se que o modelo descrimina bem os dois grupos, embora essa análise seja visual, há indícios de que o modelo discrimina bem, pelo procedimento que consiste em calcular a área sob a curva ROC, denomidado de AUROC, a área encontrada foi de 0,82, determinado que o modelo descrimina os grupos de forma excelente, veja Tabela 4, página 23.

Figura 23 - Curva ROC



Na Tabela 31 pode-se notar como o modelo classificou os indivíduos que participaram da modelagem, pode-se perceber que o modelo realizou uma boa classificação, pois classificou corretamente 308 observações dentre os 400 amostrados.

Tabela 31 - Matriz de dupla entrada

Previsão do modelo	Situação real		Total
	Mau	Bom	
Mau	157	49	206
Bom	43	151	194
Total	200	200	400

Na Tabela 32, pode-se notar que a capacidade de acerto total foi de 0,77, indicando que o modelo classificou corretamente 77% dos 200 observados. Analisando a capacidade de acerto de maus e bons clientes, tem-se que respectivamente o modelo acerta 78,5% dos maus e 75,5% dos bons.

Tabela 32 - Capacidade de acerto do modelo na amostra de desenvolvimento

Indicadores	Medidas
Capacidade de acerto total	0,770
Especificidade	0,785
Sensibilidade	0,755
Valor preditivo positivo	0,762
Valor preditivo negativo	0,778
Prevalência	0,500
Correlação	0,540

Segundo os resultados da Tabela 32, verifica-se o valor preditivo positivo do modelo de 76,2%, ou seja, a probabilidade de um indivíduo ser inadimplente, dado que o modelo o classificou como tal é de 76,2%. De maneira análoga, o valor preditivo negativo é de 77,8%, isto é, a probabilidade de um indivíduo não ser inadimplente, dado que o modelo o classificou desta forma, é de 77,8%. A prevalência de 50% mostra que proporções de bons e maus clientes estão príximas e a correlação de Mathews de 0,54 mostra uma correlação positiva, além de diferir de zero, indicando que a previsão não é aleatória.

4.4.2.3 Coeficiente de Gini

O coeficiente de GINI é uma medida de discriminação do modelo, o resultado calculado de 0,64 indica que o modelo discrimina bem o grupo dos bons clientes dos maus clientes, como visto na Seção 3.3.

4.4.3 *Capacidade de acerto do modelo na amostra teste*

Na Tabela 33, pode-se notar a proporção de acertos e erros da classificação do modelo final em um banco de dados “teste”. Para o teste foram usadas 200 observações divididas

em igual proporção. A Tabela 33 apresenta a classificação dos clientes na amostra de teste.

Tabela 33 - Matriz de dupla entrada

Previsão do modelo	Situação real		Total
	Mau	Bom	
Mau	69	39	108
Bom	31	61	92
Total	100	100	200

Na Tabela 34 são apresentados as medidas referentes a amostra de teste, pode-se notar que a capacidade de acerto total foi de 0,65, indicando que o modelo classificou corretamente 65% dos 200 observados. Analisando a capacidade de acerto de maus e bons clientes, tem-se que respectivamente o modelo acertou 69% dos maus e 61% dos bons comparando com os 100 indivíduos de cada grupo.

Tabela 34 - Capacidade de acerto do modelo em amostra teste

Indicadores	Medidas
Capacidade de acerto total	0,650
Especificidade	0,690
Sensibilidade	0,610
Valor preditivo positivo	0,638
Valor preditivo negativo	0,663
Prevalência	0,500
Correlação	0,300

De acordo com os resultados da Tabela 34, verifica-se o valor preditivo positivo do modelo de 63,8%, ou seja, a probabilidade de um indivíduo ser inadimplente, dado que o modelo o classificou como tal é de 63,8%. De maneira análoga, o valor preditivo negativo é de 66,3%, isto é, a probabilidade de um indivíduo não ser inadimplente, dado que o modelo o classificou desta forma, é de 66,3%. A prevalência de 50% mostra que as proporções de bons e maus clientes estão divididas igualmente e a correlação de Mathews de 0,3 mostra uma correlação positiva, indicando que a previsão não é aleatória.

4.4.4 *Comparação da amostra de desenvolvimento e teste*

Os valores calculados para a capacidade de acerto da amostra teste foram menores do que da amostra de desenvolvimento, era de se esperar, pois o modelo foi construido utilizando os dados da amostra de desenvolvimento. Essa diferença nos calculos foram

mínimas, onde implica que os dados de teste tiveram boa resposta ao modelo. Portanto o modelo é validado e a análise é concluida.

4.4.5 *Interpretação dos Odds Ratio dos coeficientes estimados*

- **Estado da conta (Q1) - Sem conta corrente (A14):** o fato do cliente não possuir conta corrente reduz o risco de apresentar algum problema de crédito, em que o valor do odds ratio é 0,18. Em regressão logística indica que a chance de ocorrer algum problema para os clientes que não possui conta corrente é aproximadamente 18% do que para os que possuem.

- **Tempo de crédito em meses (Q2) - 36 a 72 meses (A25):** o fato do cliente ter um tempo de crédito de 36 a 72 meses faz com que o risco de crédito aumente 2,18 vezes em relação aos clientes que possuem crédito em um tempo inferior a 36 meses. O risco de observar inadimplência é mais provável nessa categoria.

- **Histórico do crédito (Q3) - Crédito existente em outro banco (A34):** essa categoria reduz o risco de apresentar algum problema de crédito, em que o valor do odds ratio é 0,35. Esse valor para o odds ratio indica que a chance de ocorrer algum problema para os clientes que possuem crédito em outro banco é aproximadamente 35% da chance dos que não posuem crédito tomado, crédito quitado devidamente e atraso no pagamento.

- **Finalidade do crédito (Q4) - Investir em negócios e fazer reparos (A42):** o fato do cliente pedir o crédito com a finalidade de investir em negócios e fazer reparos existente reduz o risco de apresentar algum problema de crédito, em que o valor do odds ratio é 0,35.

- **Valor do crédito (Q5) - 4000 a 18000 (54):** nesta categoria o cliente que tem valor de crédito de 4000 a 18000 faz com que o risco de crédito aumente 2,54 vezes em relação aos clientes que possuem um crédito menor que 4000. Logo quem tem crédito alto tem mais chance de ser inadimplente em relação aos clientes que se enquadram nas outras categorias.

- **Saldo médio da conta poupança (Q6) - Maior que 500 Um (63):** o fato do cliente ter uma conta poupança com mais de 500 faz com que ocorra uma redução no risco de crédito, em que o valor do odds ratio é 0,29. Logo nesta categoria é menos provável a acorrência de clientes inadimplentes.

- **Tempo no emprego atual (Q7) - A mais de quatro anos (74):** o fato do cliente ter maior estabilidade no emprego, com quatro a sete anos, oferece um risco menor de crédito, em que o valor do odds ratio é 0,51. A chance de ocorrer algum problema para cliente que está empregado a mais quatro anos é aproximadamente 51% menor que clientes com menos de quatro anos de emprego.

- **Sexo/Estado civil (Q9) - Masculino solteiro(A93):** o fato do cliente ser do sexo masculino e estado civil solteiro oferece um risco menor de crédito, em que o valor do odds ratio é 0,53. A chance de observar a inadimplência em clientes masculinos e solteiro é de 53% menor da chance dos que não pertencem a essa categoria.

- **Co-requerente (Q10) - Deve ao fiador (103):** essa categoria oferece um risco menor de crédito, em que o valor do odds ratio é 0,27. A chance de ocorrer algum problema para os clientes que deve ao fiador é 27% do que para os que não devem a ninguém e deve ao co-requerente.

- **Outras formas de parcelamento (Q14) - Nenhum (A143):** o fato do cliente não ter nenhum parcelamento faz com que o risco de crédito seja menor, em que o valor do odds ratio é 0,35. A chance de observar o fenômeno da inadimplência para os clientes que não tem nenhum parcelamento é de 35% dos clientes que tem parcelamento em lojas ou bancos.

5 CONCLUSÃO

A análise de crédito é uma técnica bastante comum na área financeira. Quando uma instituição financeira vende um crédito a um cliente ela estará comprando um risco, onde esse risco é a probabilidade do cliente não cumprir com suas obrigações. A fim de prever o risco em uma operação de crédito, foi desenvolvido o modelo de risco de crédito por meio de análise de regressão logística. A partir desse modelo as instituições financeiras vendem crédito a um risco menor maximizando os lucros com segurança.

O Banco de dados utilizado nesta aplicação é chamado de *German Credit data*, foi produzido pela Universidade de Hamburgo, pelo Professor Dr. Hofmann do Instituto de Estatística e Econometria. Este banco passou por um tratamento de dados a fim de ter mais qualidade e a partir disto foi submetido a métodos estatísticos para a mineração das informações.

A análise exploratória é uma técnica fundamental para indicar quais variáveis poderiam participar da construção do modelo de risco de crédito. No modelo final, todos os coeficientes foram significativos, indicando assim que todas as variáveis selecionadas explicam a probabilidade de inadimplência de um indivíduo.

A avaliação da qualidade do modelo apresentou resultados satisfatórios a respeito da discriminação dos grupos. Os índices que avaliaram a capacidade de acerto do modelo, tanto na amostra de desenvolvimento, quanto na amostra de validação, obtiveram bons resultados em seus respectivos índices. As variáveis selecionadas para explicar o risco de uma operação de crédito foram: Estado da conta; Tempo de crédito em meses; Histórico da conta; Finalidade do crédito; Valor do crédito; Saldo médio da conta poupança; Tempo no emprego atual; Taxa de parcelamento em porcentagem; Sexo/Estado civil; Co-requerente; Tempo de permanência na residência atual; Idade; Outras formas de parcelamento; e Estado de habitação. Mediante a análise de todos os indicadores, concluiu-se que o modelo proposto satisfaz todas as condições, portanto a análise foi bem sucedida.

Para os alunos do DEMA - UFC deixo este trabalho como contribuição a fim de

que motive novos alunos a desenvolver novas pesquisas nesse segmento, pois essa área é bastante promissora tanto no meio acadêmico, quanto no mercado de trabalho. A presente monografia possibilitou em mim o aperfeiçoamento nos modelos lineares generalizados, conhecer uma forma muito útil de aplicar a técnica e além facilitar o ingresso no mercado de trabalho.

REFERÊNCIAS

ANDRADE, F. W. M.**Modelos de risco de crédito:** tecnologia de crédito. n. 38, 2003.

BALDI, P. et al. **Assessing the accuracy of prediction algorithms for classification:** an overview. Bioinformatics, v. 16, n. 4, p. 412-424, 2000.

BAPTISTA, José Galvão. **O custo de intermediação financeira em Cabo Verde:** fatores condicionantes. Praia: Banco de Cabo Verde, 2006.

BOLFARINE, H. ; BUSSAB, W. O. **Elementos de amostragem**. São Paulo: Edgard Blucher, 2005.

CAMPOS, Humberto de. Teste de Chi-Quadrado e teste de Kolmogorov-Smirnov. In: **Estatística experimental não-paramétrica**. 4. ed. Piracicaba: USP/ESALQ, 1983. cap. 2, p. 38-47.

CAMPOS, P. S. S. **Estimação Bayesiana em modelos de regressão logística**. Dissertação de mestrado. UFPA, (2007).

CAQUETE, J. ; ALTMAN, E. ; NARAYANAM, P. **Gestão do risco de crédito:** o próximo grande desafio financeiro. Rio de Janeiro: Qualitymark, 1999.

DINIZ, Carlos ; LOUZADA, Francisco. **Modelagem estatística para risco de crédito**. Paraíba: /s. n./, 2012. (20 SINAPE).

HOSMER, D. ; LEMESHOW, S. **Applied logistic regression.** 2nd ed. New York: Wiley-Interscience, 2000.

JOOS, P. et al. **Credit classification:** a comparison of logit models and decision trees. In: PROCEEDINGS NOTES OF THE WORSHIP ON APPLICATION OF MACHINE LEARNING AND DATA MINING IN FINANCE, 1988. Belgium. Conference Belgium, 1988. p. 59-72.

LECUMBERRI, L. F. L. ; DUARTE, A M. **Uma metodologia para o gerenciamento de modelos de escoragem em operações de crédito de varejo no Brasil.** Economia Aplicada, São Paulo, v. 7, n. 4, p. 795-818, 2003.

LEWIS, E. M. **An introduction to credit scoring.** San Rafael : Isaac and Co., 1992.

MATTHEWS, B. W. **Comparison of the predicted and observed secondary structure of t4 phageiysozyme.** Biochimica et Biophysica Acta, v. 405, n. 2, p. 442-451, out. 1975.

MCCULLAGH, P. ; NELDER, J. **Generalized linear models.** 2nd ed. Boca Raton, Flórida: Chapman & Hall/CRC, 1989.

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

R Core Team (2013). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

SICSU, Abraham Laredo. **Credit scoring:** desenvolvimento, implantação, acompanhamento. São Paulo: Blucher, 2010.

THOMAS, L. C. **A survey of credit and behavioral scoring:** forecasting financial risk of lending to consumers. International Journal of Forecasting, Edinburg, v. 16, n. 2, p. 149-172, 2000.

THOMAS, L. ; EDELMAN, D. ; CROOK, J. N. **Credit scoring and its applications.** Philadelphia: Siam, 2002.

THOMAS, Lyn C. **Consumer credit models:** pricing, profit and portfolios. New York: Oxford University, 2009.

TOMAZELA, S. ; SICSU, A . L. ; LIMA, A. C. P. **Análise empírica dos indicadores KS e ROC**. São Paulo: FGV/EAESP, 2008.

ANEXO

Análise de diagnóstico

Figura 24 - Quantis de probabilidade

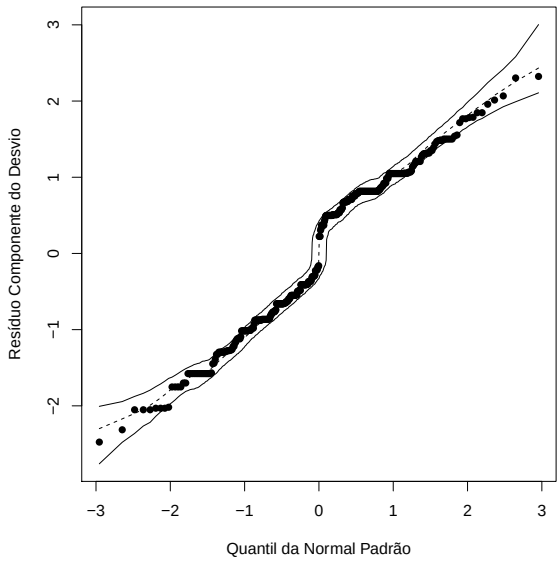


Figura 25 - Distância de cook

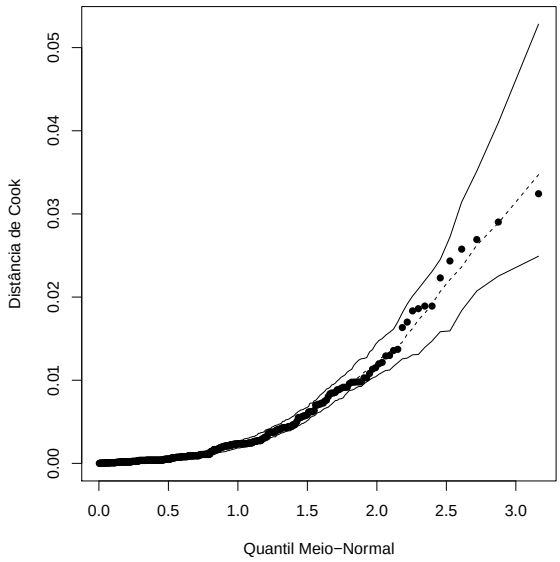


Figura 26 - Análise de Sensibilidade