UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Comparação do desempenho de modelos de *Credit Scoring* utilizando diferentes composições amostrais de grupos de clientes

Autora: Pricila Duarte Moreira

Orientadora: Professora Dra. Lisiane Priscila Roldão Selau

Porto Alegre, dezembro de 2014

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Comparação do desempenho de modelos de *Credit Scoring* utilizando diferentes composições amostrais de grupos de clientes

Autora: Pricila Duarte Moreira

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professora Dra. Lisiane Priscila Roldão Selau
Professor MSc. Eduardo de Oliveira Horta

Porto Alegre, dezembro de 2014.

DEDICATÓRIA

Dedico este trabalho, com todo meu amor,
aos meus queridos e incomparáveis pais, Paulo Ricardo e Jucelaine,
e ao meu doce e companheiro noivo, Bruno.

“Viva como se você fosse morrer amanhã.

Aprenda como se você fosse viver para sempre.”

Mahatma Gandhi

AGRADECIMENTOS

Ao incentivo e amor incondicional dos meus pais, Paulo e Jucelaine. Vocês são responsáveis por tudo isso ter sido possível. Obrigada por tudo que vocês fazem por mim, a cada dia.

À amizade, apoio e bagunça dos meus irmãos, Paola e Ricardo. Mana, tu, mesmo sem saber, trouxe as palavras certas quando eu precisava. Mano, mesmo negando, a cosquinha no final dos abraços fez toda a diferença.

Ao meu amor, Bruno, que iniciou esta caminhada ao meu lado como namorado e estamos concluindo juntos, como noivos. Teu carinho e compreensão me fortaleceram e confortam.

Pai, mãe, mana, mano e Bruno, amo vocês demais!

Aos meus familiares e amigos, que sempre estiveram na torcida e entenderam minha ausência.

Aos meus quatro avós, por sempre transmitirem confiança, mesmo o avô que só pode estar de corpo presente até a fase do vestibular.

Às minhas madrinhas, por todo carinho e por apoiarem minhas decisões.

À Universidade Federal do Rio Grande do Sul (UFRGS), por me proporcionar um ensino de qualidade em um campus maravilhoso com animais adoráveis.

Aos colegas de Estatística, pela parceria e aprendizado nesses anos de convívio.

A minha orientadora Prof. Dra. Lisiane Selau pelos ensinamentos, ternura e dedicação durante a realização deste trabalho.

A todos os professores da UFRGS que em algum momento do curso tive a oportunidade de ter o conhecimento compartilhado.

Muito obrigada!

RESUMO

Com o crescente incentivo ao consumo e aumento na concessão de crédito, os modelos de previsão de risco passam a ter grande importância como ferramenta para tomada de decisão. Para a construção destes modelos o impacto de diferentes composições amostrais de grupos de clientes é questão importante para investigar. Nesse sentido, este trabalho tem como objetivo comparar o desempenho de modelos com distintas proporções amostrais dos grupos bons e maus clientes por meio da técnica estatística de regressão logística. Todas as etapas de construção do modelo são descritas detalhadamente, sendo o método dividido em sete grandes etapas: (i) Planejamento e definições; (ii) Identificação das variáveis previsoras; (iii) Amostragem e coleta dos dados; (iv) Análise dos dados; (v) Análise dicotômica; (vi) Obtenção da fórmula preliminar; (vii) Acurácia e validação do modelo. Os três esquemas de amostragem comparados foram aplicados no mesmo banco de dados real, proveniente de uma rede de farmácias com crediário próprio. Os resultados obtidos demonstram que as diferentes composições amostrais interferem na previsão do perfil do grupo de interesse, pois quanto maior a proporção dos indivíduos que são minoria na população, neste caso os maus pagadores, maior a taxa de acerto no grupo.

Palavras-chave: Análise de Crédito; Previsão de Risco; Regressão Logística; Composição Amostral.

ABSTRACT

With the increasing boost consumption and increase in lending, risk prediction models now have great importance as a tool for decision making. For the construction of these models the impact of different sample compositions customer groups is important issue to investigate. Thus, this study aims to compare the performance of models with different sample proportions of good groups and bad clients through logistic regression statistical technique. All model building steps are described in detail, the method divided into seven major steps: (i) Planning and definitions; (ii) Identification of the predictor variables; (iii) Sampling and data collection; (iv) analysis of data; (v) dichotomous analysis; (vi) providing the preliminary formula; (vii) Accuracy and validation of the model. The three compared sampling schemes were applied in the same actual database, from a network of pharmacies with customer credit. The results show that the different sample compositions interfere in predicting the profile of the group of interest, because the higher the proportion of individuals who are a minority in the population, in this case the bad payers, the higher the hit rate in the group.

Keywords: Credit Analysis; Risk Forecast; Logistic Regression; Sample Composition.

SUMÁRIO

1. Introdução 10
2. Referencial Teórico 12
2.1. Crédito e Modelos de Crédito 12
2.2. Regressão Logística 14
3. Método e Sistemática da Modelagem 15
3.1. Planejamento e Definições 16
3.2. Identificação das Variáveis Previsoras 17
3.3. Amostragem e Coleta dos Dados 17
3.4. Análise dos Dados 18
3.5. Análise dicotomizada 19
3.6. Obtenção da Fórmula Preliminar 20
3.7. Acurácia e Validação do Modelo 21
4. Resultados 22
4.1. Planejamento e Definições 22
4.1.1. Definição do produto e do mercado 22
4.1.2. Finalidade do modelo e composições a serem comparadas 22
4.1.3. Definição da inadimplência / diferentes percentuais 22
4.1.4. Definição do horizonte de previsão do modelo e verificação do histórico de crédito 23
4.2. Identificação das Variáveis Previsoras 23
4.3. Amostragem e Coleta dos Dados 24
4.3.1. Separação das amostras de análise, validação e teste 24
4.3.2. Definição do período e tamanho da amostra para cada modelo 24
4.4. Análise dos Dados 26
4.4.1. Avaliação da consistência e preenchimento dos dados 26
4.4.2. Escolha das variáveis que integrarão o modelo 26
4.5. Análise dicotômica 27
4.5.1. Agrupamento de atributos de variáveis 27
4.5.2. Criação das variáveis *dummies* 27
4.6. Obtenção da Fórmula Preliminar 27
4.6.1. Escolha de técnicas estatísticas multivariadas e Determinação do *software* a ser usado 27

4.6.2. Seleção de variáveis independentes e Verificação da suposição das técnicas 28
4.7. Acurácia e Validação do Modelo 31
4.7.1. Medição de índices de desempenho (KS, Curva ROC) 31
4.7.2. Percentual de classificações corretas 32
4.7.3. Comparação de desempenho dos esquemas de composição amostral propostos 33
5. Conclusões 34
REFERÊNCIAS 36
APÊNDICE A - AGRUPAMENTO DE CEP RESIDENCIAL 39
APÊNDICE B - AGRUPAMENTO DE CEP COMERCIAL 42
APÊNDICE C - AGRUPAMENTO DE NATURALIDADE 45
APÊNDICE D - AGRUPAMENTO DE PROFISSÕES 48

1. Introdução

As mudanças ocorridas no cenário financeiro mundial a partir dos anos 90, como a desregulamentação das taxas de juros e câmbio, o aumento de liquidez, o aumento da concorrência bancária e a estabilização do real, geraram um significativo aumento de competição financeira e propiciaram um maior incentivo ao crédito. Contudo, atrelado ao aumento da concessão de crédito está o risco de crédito, cujo gerenciamento é um grande desafio, definido por Morgan (2013) como a possibilidade de perdas devido ao não pagamento, por parte do cliente, dos valores devidos ou a incapacidade da contraparte em honrar os compromissos assumidos com o credor.

A gestão de risco passou a ocupar, nos últimos tempos, posição de destaque na administração financeira, especialmente em consequência da expansão do crédito, do crescimento do mercado e da globalização (BUENO, 2003). Qualquer erro na decisão de concessão pode significar que em uma única operação haja a perda do ganho obtido em dezenas de outras bem-sucedidas (STEINER *et al*., 1999). Entende-se, portanto, que a implementação de técnicas robustas para controle e gerenciamento do risco é de grande importância para a melhoria na capacidade de distinção entre bons e maus pagadores, auxiliando em decisões mais assertivas e menos subjetivas na concessão de crédito e objetivando o maior aceite de clientes considerados propícios à adimplência e recusa de clientes potencialmente inadimplentes.

Nesse sentido, ao invés de visar a adesão de todos os clientes à sua carteira de crédito, as empresas devem ser seletivas, oferecendo incentivos financeiros apenas àqueles que apresentem maior probabilidade de quitar suas dívidas com base no seu histórico de consumidor e comportamento de compras. Quando há o acesso a estas informações, muitas delas podem não ser fidedignas, completas ou claras. Em ambientes dinâmicos, onde as decisões devem ser tomadas de forma rápida, estas análises não podem se dar de maneira subjetiva, sem critérios definidos.

Esta gestão do risco é uma tarefa muito trabalhosa, demorada e que tem um impacto significativo nos lucros das empresas. As decisões devem ser muito criteriosas, pois decisões erradas podem provocar não apenas prejuízos financeiros à instituição de crédito, mas também prejuízos financeiros e morais aos seus clientes (MENDES FILHO *et al.*, 1996).

A utilização de modelos de previsão de risco de crédito pode ser um fator fundamental para o sucesso das empresas, pois é delineado, com base nas informações dos atuais usuários, o perfil comportamental de clientes inadimplentes, gerando, a partir daí, uma regra que indique previamente a chance de não pagamento de um futuro cliente. Assim, a decisão de concessão de crédito fica facilitada e mais confiável, substituindo-se argumentos subjetivos por argumentos quantitativos. Segundo Thomas (2000), instituições financeiras estão intensificando e aperfeiçoando a utilização de ferramentas estatísticas para a concessão de crédito, comparando, dentre as técnicas existentes, a que trará resultados mais positivos para o negócio da empresa.

Contudo, o impacto de diferentes composições amostrais também deve ser fator de investigação. Brown e Mues (2012) comentam que diversos estudos têm sido realizados para comparar o desempenho das várias técnicas existentes, mas eles não incidem especificamente sobre a forma como estas técnicas possam ser influenciadas por amostras fortemente desequilibradas, ou até que ponto tal comparação é afetado pelo problema de desequilíbrio de classe.

É plausível pensar que a diferença na taxa de acerto do modelo de classificação de risco pode variar ao utilizar-se composições amostrais onde a proporção dos grupos bons e maus é igual entre si, igual à da população, com maioria bons, ou inversa à da população, com maioria maus. Tendo em vista que na grande maioria das carteiras o número de inadimplentes é consideravelmente menor que o número de pagadores em dia, a extração de amostras destes dois grupos em tamanhos iguais, como frequentemente utilizado, pode não ser o método com melhores resultados. Nesse sentido, o objetivo deste trabalho é verificar como o desbalanceamento das composições amostrais de bons e maus clientes pode interferir nas taxas de acerto de um modelo de *Credit Scoring* – modelo estatístico para obter-se a probabilidade do futuro cliente pagar o crédito fornecido.

A justificativa para a realização deste trabalho está embasada na importância do crédito a pessoas físicas no atual contexto econômico do país, no qual a utilização de crédito é maciçamente incentivada e combustível para o desenvolvimento financeiro de pessoas economicamente menos favorecidas. Ainda, é importante salientar que os resultados deste trabalho poderão contribuir para o aumento de lucro – ou diminuição do prejuízo – das empresas, visto que quaisquer melhorias na detecção de bons e maus pagadores podem provocar uma melhoria significativa nos ganhos.

Este trabalho está organizado em cinco seções. Primeiro, a introdução, já apresentada nesta seção, contextualizando a importância de utilizações de modelos de previsão para a área de crédito e a necessidade de investigarmos o impacto da utilização de composições amostrais diferentes. Após, a segunda seção trará a fundamentação teórica, na qual serão expostos o referencial sobre modelos de previsão de risco de crédito e a Regressão Logística, técnica utilizada para a construção do modelo com três diferentes composições amostrais, nas seguintes proporções de bons e maus pagadores, respectivamente: 64/36, 50/50 e 36/64. Na terceira seção é detalhada a sistemática proposta para a construção do modelo. A quarta seção apresentará o modelo construído, bem como a discussão dos resultados obtidos para cada composição amostral em um mesmo banco de dados relacionados a clientes de uma rede de farmácia com unidades no Rio Grande do Sul. Na quinta e última seção, serão apresentadas as principais conclusões deste trabalho juntamente com as considerações finais deste estudo.

2. Referencial Teórico

O risco é um fator intrínseco em quase totalidade das tomadas de decisão. Sua correta análise é tema de grande relevância nas instituições financeiras, visto ser motivo de impulsionamento aos ganhos esperados, conforme Selau (2012). Entender e administrar esse risco auxiliam no processo decisório e em uma tomada de decisão mais assertiva, pois, em posse de informações acerca das alternativas existências e prevendo resultados esperados, poderemos assumi-lo objetivando retornos mais satisfatórios.

A recusa de crédito a todo e qualquer cliente que apresente risco pode ser financeiramente errônea. É sabido que provavelmente nem todos os devedores pagarão suas contas em dia. Porém, algumas pessoas podem gerar um lucro grande à empresa por meio do pagamento de suas parcelas acrescidas de juros, ou ainda, taxas e multas devido a atrasos. Os ganhos obtidos por estes clientes poderão ser superiores aos ganhos obtidos pelos clientes que quitam suas contas até a data de vencimento, comenta Pacheco (2012). Contudo, é imprescindível conhecer e delimitar o limite tênue onde a inadimplência passe a ser prejudicial à instituição. A empresa toma decisões de concessão de crédito de forma que a composição da carteira visa maximizar o lucro esperado.

Segundo Steiner *et al.* (1999), a correta decisão de concessão de crédito é essencial para a sobrevivência das instituições financeiras. Por vezes pode ser preferível aceitar menor número de bons pagadores a fim de diminuir, juntamente, o número de maus pagadores recebendo crédito, assim, os ganhos podem ser maximizados pela redução das perdas com inadimplência.

O controle e administração eficaz do risco de crédito se tornaram indispensáveis. Sendo assim, as instituições financeiras passaram a primar pela constante melhoria das ferramentas utilizadas para tal fim, por meio do desenvolvimento e aperfeiçoamento de métodos objetivos e quantitativos.

2.1. Crédito e Modelos de Crédito

Na atual dinâmica financeira vista no mundo, e em particular no Brasil, a concessão de crédito é um dos pilares do desenvolvimento econômico. Com as políticas governamentais de aumento de renda e incentivo ao microcrédito adotadas nos últimos anos, muitas pessoas saíram da linha de pobreza, passando a integrar uma parcela da população economicamente ativa e consumidora, alvo das instituições financeiras e do mercado varejista. Sem a alternativa do crédito, a quantidade de compradores poderia ser muito menor e, por consequência, o lucro do fabricante também seria reduzido (SILVA, 2003).

Não apenas no mercado financeiro, mas também no comércio, o crédito assume um papel de facilitador. Possibilita não apenas o aumento do poder de compra do cliente, mas também um incremento das vendas para o ofertante. Porém, nos grandes varejos para pessoa física, um ponto desfavorável é a necessidade de fornecer crédito em grande escala, o que demanda elevado

número de analistas e as empresas não dispõem para suprir essa concessão em massa (PACHECO, 2012).

Apesar de a análise de crédito feita por um analista levar em consideração todo seu conhecimento técnico e empírico sobre a empresa e o ramo do negócio, a decisão se dá de forma subjetiva, podendo variar a decisão se tomada por diferentes analistas com equiparável expertise. Ainda, a mensuração do risco não se dá de forma quantitativa, de tal modo que se faz necessária a utilização de ferramentas sofisticadas para auxiliar a quantificar o risco das operações na tomada de decisão, como comenta Moraes (2012).

Neste sentido, a fim de agregar vantagem competitiva e proporcionar uma maior rentabilidade e segurança nas relações comerciais, muitas empresas vêm adotando os modelos estatísticos de *Credit Scoring*. Modelos de previsão de risco de crédito vêm sendo amplamente estudados e ganhando espaço devido a sua importância para a saúde de instituições financeiras, já que o sucesso dessas instituições está diretamente relacionado a sua capacidade de gerir os riscos (GHODSELAHI, 2011).

Os modelos de *Credit Scoring* são sistemas que atribuem pontuações às variáveis de decisão de crédito de um proponente, mediante a aplicação de técnicas estatísticas. Esses modelos visam a segregação de características que permitam distinguir os bons dos maus créditos (LEWIS, 1992). Diante do exposto, é notável que a utilização de modelos estatísticos para criação de escore de risco de crédito propicia benefícios voltados a diversos aspectos, entre eles a automatização da apreciação de concessão, capacidade de aumentar o volume de análises, objetividade nas decisões, gerenciar eficaz do risco da carteira de crédito e consistência nos critérios de escolha.

Assim, a ideia essencial dos modelos de *Credit Scoring* é identificar certos fatores-chave que influenciam na adimplência ou inadimplência dos clientes, permitindo a classificação dos mesmos em grupos distintos e, como consequência, a decisão sobre a aceitação ou não do crédito em análise. (ARAUJO; CAMONA, 2007). Uma pequena melhora na precisão da classificação dos clientes pode resultar em uma grande redução do risco e gerar significativa economia para a instituição.

O desenvolvimento desses modelos se dá por meio do entendimento e detecção, através de um procedimento técnico, de quais características estão mais significativamente ligadas com os atributos dos clientes bons pagadores, conforme abordado por Louzada (2008). A partir disto, obtém-se uma estimativa para a probabilidade de pagamento por parte dos solicitantes de crédito, criando, então, uma classificação de perfil de clientes ranqueados de acordo com sua chance de inadimplência.

Autores atentam ao fato de que nenhum modelo de previsão de risco é capaz de detectar todas as características relevantes para a distinção do perfil de cada indivíduo. Caso fosse possível, custo, trabalho e tempo tornariam o procedimento potencialmente não satisfatório. Assim sendo, não estão

eliminadas as chances de se recusar para a base de clientes um bom pagador, assim como de se aceitar um mau pagador. A calibragem do método deve se dar de tal maneira que os custos esperados da empresa sejam minimizados, tendo conhecimento de qual situação acarretará maior prejuízo aos negócios. Deve-se enfatizar que a modelagem é uma ferramenta de apoio à análise capaz de agregar diversos benefícios, porém não é capaz de substituir o conhecimento dos profissionais da área.

Os modelos de *Credit Scoring* são geralmente baseados em técnicas de análise estatística multivariada. Na sequência será descrita a regressão logística, abordagem utilizada neste trabalho.

2.2. Regressão Logística

Quando se tem o interesse na avaliação da influência de fatores sobre uma resposta dicotômica, a regressão logística é o método usualmente utilizado (HOSMER; LEMESHOW, 1989). A análise Logit ou regressão logística consiste em uma técnica estatística utilizada na separação de dois grupos, que visa obter a probabilidade de que uma observação pertença a um conjunto determinado, em função do comportamento das variáveis explicativas (HAIR *et al.*, 2005).

A utilização da técnica de regressão logística é adequada em muitas situações porque permite que se analise o efeito de uma ou mais variáveis explicativas (discretas ou contínuas) sobre uma variável resposta dicotômica, representando a presença (1) ou ausência (0) de uma característica (HOSMER; LEMESHOW, 1989). Sua aplicação, por intermédio de um modelo, é frequentemente utilizada para análises cujos dados possuam resposta binária ou dicotômica a fim de realizar a predição ou explicar a ocorrência da variável resposta, relacionando-a com os fatores que influenciem positiva ou negativamente – variáveis preditoras – a probabilidade da ocorrência de determinado fenômeno.

Assim, na regressão logística, a variável dependente, uma vez que possui caráter não-métrico, é inserida através do uso de variáveis *dummy* (dicotômica ou binária), que assumem valor 0 para indicar a ausência de um atributo e 1 para indicar a presença de um atributo (GUJARATI, 2000).

O modelo de previsão da regressão logística contextualizado na análise de risco de crédito é utilizado para mensuração da probabilidade de inadimplência de determinado grupo de clientes, assumindo ser logisticamente distribuída, com resultado binomial 0 ou 1. Supondo Y a variável resposta binária e X_1, X_2, ..., X_k as variáveis independentes preditoras de Y, a função de distribuição pode ser expressa pela Equação 01:

$$P_i = Pr\,[\,Y_i = 1 \mid X = X_i\,] = 1 / (1+e^{-z}) \qquad (01)$$

onde:

P_i : Representa a probabilidade de um evento ocorrer;

$Z= b_0 + b_1X_1 + b_2X_2 + ... + b_iX_i$;

X_i : Representam cada i-ésima variável preditora;

b_i : Coeficientes a serem estimados para cada uma das variáveis.

A variável Z, conhecida como *log odds*, varia de $-\infty$ a $+\infty$. Assim, a função probabilidade P_i normaliza a saída do modelo para o intervalo [0,1], informando o valor de interesse, a probabilidade de ocorrência do evento estudado. Para testar a significância dos coeficientes estimados utilizaremos a estatística de Wald, como sugerido em Hair *et al.* (2005), por fornecer a significância estatística para cada coeficiente estimado.

Apesar de sua flexibilidade, existe o pressuposto importante da baixa correlação entre as variáveis explicativas, já que o modelo de regressão logística é sensível à colinearidade entre as variáveis (HAIR *et al*., 2005). A utilização de variáveis altamente correlacionadas para a estimação do modelo pode ocasionar estimativas extremamente inflacionadas dos coeficientes de regressão (HOSMER; LEMESHOW, 1989).

Há métodos automáticos para selecionar as variáveis que irão compor um modelo, baseados num algoritmo pré-existente que testa a importância de cada variável. Esta seleção pode se dar a partir da inclusão ou exclusão de variáveis fundamentada em uma regra de decisão. Segundo Corrar *et al.* (2007), o método *stepwise* é considerado como contendo ações corretivas para os problemas de multicolinearidade. Desta forma, ao desconsiderar variáveis que apresentem sinais de multicolinearidade mantemos no modelo apenas as variáveis que agregam na explicação da variável resposta por possuir maior significância estatística.

3. Método e Sistemática da Modelagem

Para a estimação do modelo serão utilizados dados reais, provenientes de uma farmácia que concede crédito próprio para parcelamento das compras. Tendo em vista o desejo de comparar qual a melhor proporção de bons e maus clientes na amostra para se obter um modelo mais eficiente, serão criados três composições amostrais com os seguintes percentuais de bons e maus clientes, respectivamente: 50/50 – mesma quantidade de ambos perfis -, 64/36 – refletindo a proporção real do negócio - e 36/64 – invertendo a proporção dos dois grupos.

O método proposto será uma adaptação, a partir do descrito por Moraes (2012) baseado em Sicsú (1998) e Sicsú (2010), composto das seguintes fases:

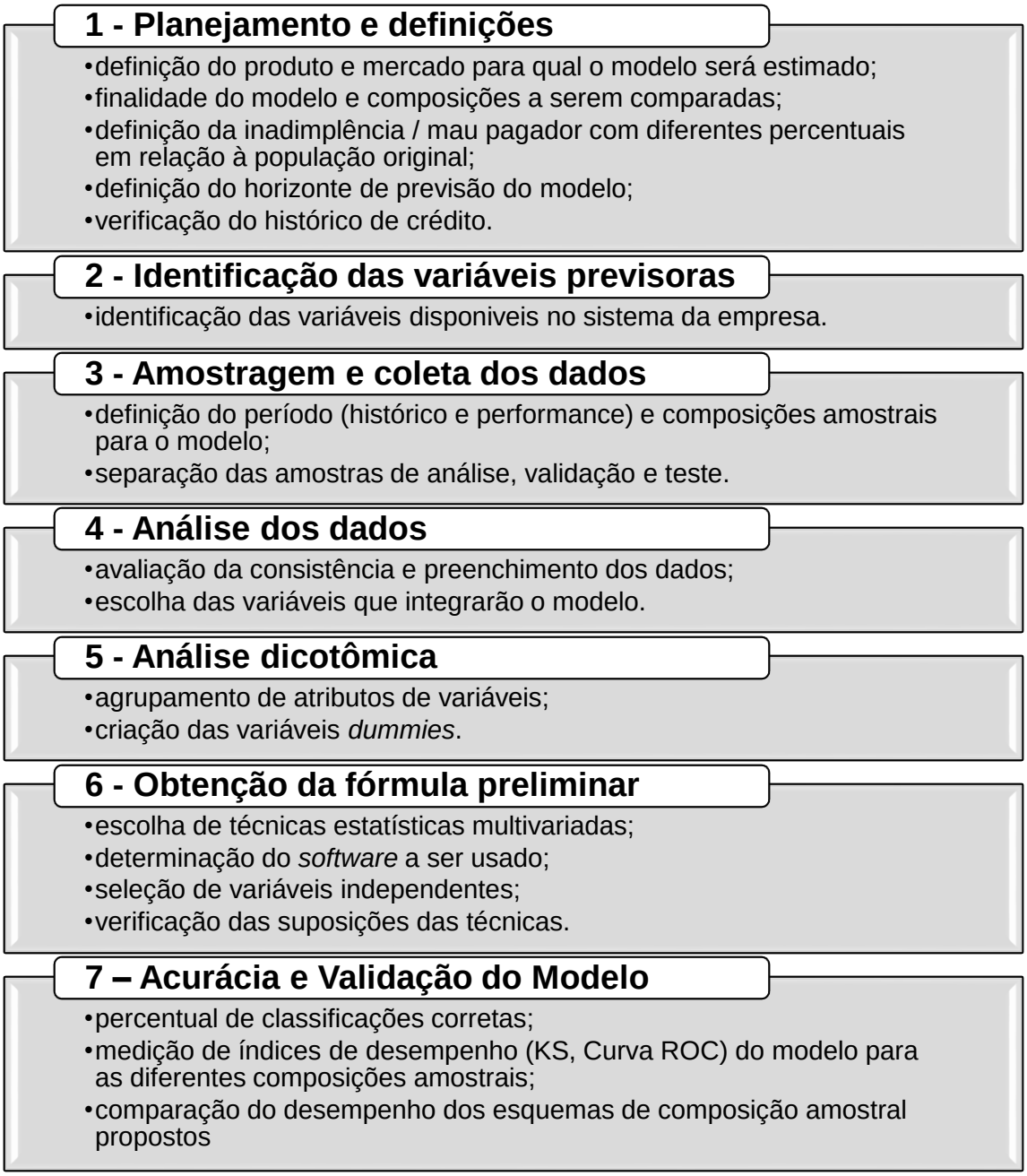

Figura 1 - Etapas para o desenvolvimento do modelo.

3.1. Planejamento e Definições

A clara e objetiva definição de quaisquer itens e etapas são fundamentais e as mesmas devem se manter inalteradas ao longo do procedimento de análise. Deste modo, deve-se objetivar o mercado para o qual o modelo será estimado e, caso haja uma gama grande de produtos de crédito fornecidos pela instituição, também deverá ser realizada a delimitação.

A formulação de três distintas propostas de composição amostral possui a finalidade de inferir qual a melhor forma de dividir a amostra para obtenção de resultados mais satisfatórios. Amostra com igual proporção de clientes bons e

maus pagadores pode trazer um equilíbrio na identificação de ambos perfis. Quando se mantém a proporção da base de dados real pode-se ter informações mais fidedignas. Já no último caso, em que a proporção dos dois grupos é inversa em relação a original, podemos estar coletando mais informações do perfil mais crítico para o lucro da empresa, os inadimplentes.

A definição de inadimplência varia de acordo com as regras pré-existentes em cada instituição, conforme relevância dos atrasos. Há, ainda, os clientes que devem ser excluídos do modelo devido à falta de histórico de utilização ou por apresentarem padrão indefinido em relação ao pagamento (ARAUJO; CAMONA, 2007).

Existem, portanto, 4 grupos que devem ser separados no total de créditos concedidos: (i) os clientes que nunca utilizaram o crédito – sem uso; (i) os clientes com pouco ou nenhum atraso – bons; (iii) os clientes em faixas de atrasos intermediárias – indeterminados; (iv) os clientes com atrasos consideráveis – maus. (SELAU, 2008). A suposição básica para construir um modelo de previsão de crédito é que os clientes têm o mesmo padrão de comportamento ao longo do tempo (GOUVÊA; GONÇALVES, 2006).

3.2. Identificação das Variáveis Previsoras

A análise das informações constantes no banco de dados da empresa pode ser feita pela observação detalhada da proposta de crédito, questionário respondido pelo cliente. Por meio dela é possível identificar as variáveis que poderão fazer parte do modelo final e, portanto, devem ser listadas na amostra de estudo. Dentre as possíveis informações selecionadas, chamadas também de variáveis demográficas, pode-se citar: sexo, idade, escolaridade, estado civil, tipo de ocupação, tipo de residência, tempo no emprego atual, entre outras (SELAU; RIBEIRO, 2009).

Desta forma, há a necessidade de definir os perfis a partir de eventos passados por meio da identificação das características dos clientes que pagam ou não o crédito tomado diante da empresa, utilizando as variáveis já existentes a fim de analisar o histórico fornecido pela instituição.

Caso haja a necessidade de propor o entendimento de atributos que não estejam contemplados pela proposta de crédito preenchida com as informações dos indivíduos, estas não serão contempladas pelo modelo, visto não haver histórico na base de dados, poderão apenas ser incluídas para melhoria da análise em momento futuro.

3.3. Amostragem e Coleta dos Dados

Ao aplicar técnicas estatísticas multivariadas é necessário atentar ao número de variáveis que serão incluídas na análise a fim de garantir um tamanho

de amostra adequado. Hair *et al.* (2005) sugerem utilizar uma proporção de ao menos 20 observações para cada variável explicativa incluída no modelo.

Em relação ao tempo de observação, devemos ter um período hábil para que o perfil de pagamento se revele e para que seja feita a classificação do indivíduo. Um tempo de 12 a 18 meses após a concessão do crédito para que se verifique o desempenho de pagamento dos clientes pode ser suficiente, conforme proposto por Lawrence (1992) e Selau (2012).

Após selecionadas todas as variáveis e unidades da amostra, serão construídos três bancos de dados. Esta etapa se dará pela necessidade de criar proporções diferentes de clientes para proceder com a comparação do modelo com as três distintas composições amostrais. Todos os procedimentos subsequentes serão executados em cada um dos três bancos de dados criados.

De acordo com Crone (2011), as amostras são construídas usando uma amostragem aleatória sem substituição referente à variável alvo com um número igual de bons e maus ou com igual proporção em relação a cada um dos grupos na população. Neste trabalho analisaremos estas duas configurações e, adicionalmente, uma composição em que o grupo de maus clientes é visivelmente maior que dos bons, tendo a proporção inversa da apresentada na população. Ainda por Crone (2011), o fator limitante para o tamanho das amostras é muitas vezes o número de maus, pois, excetuando algumas situações, os casos de maus clientes ocorrem em menor volume que os casos de bons.

Por fim, necessita-se do particionamento da amostra, sendo reservada uma parte dela para estimação e outra para a validação/teste do modelo treinado, que será usada na avaliação do poder preditivo do modelo (MORAES, 2012). Não há na literatura uma regra fixa de qual percentual da amostra deva ser utilizado para a construção do modelo. Porém, por ser a etapa fundamental da análise, deve-se demandar um maior número de observações para este momento. A amostra de análise costuma ter cerca de 70% a 80% das observações, podendo o percentual restante ser reservado apenas em uma amostra, de validação, (possui a mesma configuração da amostra de análise, porém com observações não constantes na amostra de análise, logo, desconhecidas pelo modelo) ou em outras duas, sendo uma de validação, com cerca de 20% do percentual e outra de teste (possui configuração igual à do banco de dados, contendo inclusive clientes dos grupos indefinido e sem uso), com os 10% restantes.

3.4. Análise dos Dados

Estatísticas descritivas são importantes ferramentas para a avaliação das variáveis, sendo muito úteis para criticar a proporção de *missings* nas variáveis e para detectar dados incorretos e valores discrepantes - os *outliers* (PACHECO, 2012). A detecção da presença de valores discrepantes é muito importante, pois mesmo sendo reais, podem comprometer a estimativa dos pesos das variáveis

(SICSÚ, 2010). Além disso, devem ser avaliados os valores faltantes (*missing*), podendo efetuar exclusão de observações. Variáveis com muitas observações *missings* devem ser descartadas da análise (MORAES, 2012).

Certas variáveis do banco de dados podem necessitar algum tipo de tratamento, como, por exemplo, classificar o cliente em um dos tipos, dependendo do número de dias de atraso, ou ainda, a partir de duas ou mais delas proceder com a criação de uma nova variável a fim de atender às necessidades e objetivo do modelo, como a utilização da data do cadastro e data de nascimento para a criação da variável idade. Nessa etapa é importante algum conhecimento prévio na área específica de aplicação do modelo para a escolha dessas variáveis (PACHECO, 2012).

3.5. Análise dicotomizada

Esta verificação é realizada objetivando avaliar o poder discriminatório individual de cada variável preditora por meio da avaliação de relação de cada uma delas com a variável que define o evento de interesse. Para isto, cada variável explicativa é confrontada com os dois grupos extremos que definem o modelo, ou seja, os bons e os maus pagadores. As variáveis que apresentarem maior diferença entre a porcentagem de bons e maus contribuirão mais para a separação dos perfis e terão maior capacidade preditora.

O agrupamento das classes será pelo critério do risco relativo (RR), que é calculado com o uso das tabelas de contingência, dividindo-se o percentual de bons pagadores pelo percentual de maus pagadores de cada atributo possível em cada variável, por exemplo, feminino e masculino são os atributos da variável sexo e cada uma delas possui seu respectivo risco relativo.

Quanto mais os percentuais de bons e maus diferirem para os atributos de uma mesma variável, maior será a utilidade dessa variável para o prognóstico de desempenho futuro (LEWIS, 1992). Podemos perceber pela Figura 2 que, quanto mais extremo o valor, afastando-se do ponto neutro, maior o poder discriminante da variável (ou classe), pois tem-se mais acentuadas as características que identificam os clientes.

As variáveis agrupadas terão como regra os valores apresentados na Figura 2, de modo que tenhamos classes homogêneas dentro e heterogêneas entre si, pois estarão sendo reunidos atributos com risco semelhante.

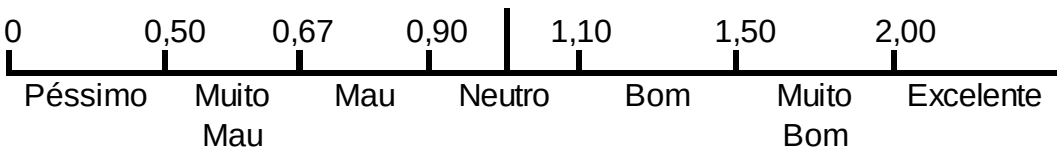

Figura 2 – Classes do risco relativo (RR).

Após a avaliação dos melhores níveis de cada variável, deve-se criar a variável *dummy* para cada atributo que fará parte da análise múltipla. Essa variável assumirá apenas o valor 0 ou 1 (ex.: estado civil solteiro = 1, caso contrário = 0). Dessa forma, os problemas de não linearidade serão evitados, como valores esperados indo de -∞ a +∞ (SELAU, 2011). As *dummies* são as variáveis indicadoras, utilizadas na construção do modelo.

3.6. Obtenção da Fórmula Preliminar

Com o banco de dados pronto e as variáveis tratadas, inicia-se a modelagem. Nesta etapa é escolhida a técnica estatística a ser utilizada, determinado o *software* estatístico em que se darão os procedimentos e, finalmente, selecionadas as variáveis explicativas.

A estimação dos parâmetros, os pesos, de cada atributo para o cálculo do score final da probabilidade estimada para cada indivíduo é realizada com base em uma técnica estatística, neste caso, por meio da regressão logística binária, como anteriormente descrito na Equação 01.

Neste estudo será utilizado o SPSS, cujos métodos disponíveis de seleção automática das variáveis que irão compor o modelo de regressão são *forward*, *backward* e *stepwise*.

Na regressão logística deve-se atentar para a verificação da ausência de multicolinearidade, e para isso pode-se lançar mão do método *stepwise*, já presentes em muitos *softwares* estatísticos. Assim as variáveis preditoras serão incorporadas no modelo automaticamente (MORAES, 2012). Este método agrega à equação de regressão logística as variáveis mais explicativas e, a cada inclusão, verifica a correlação da variável de entrada com as já acrescentadas anteriormente. Caso apresente alta correlação com as demais, haverá a retirada de uma delas a fim de assegurar o pressuposto da técnica multivariada.

A construção de um modelo é uma tarefa complexa. É necessária, por exemplo, a avaliação de variáveis que devem entrar ou sair da análise para evitar problemas de multicolinearidade. Esse cuidado é importante porque, muitas vezes, devido à presença de alta correlação entre as variáveis explicativas, podem ocorrer trocas de sinais dos pesos no modelo final (HAIR *et al.*, 2005).

A seleção das variáveis se dará pelo método *stepwise*. Contudo, apenas a utilização de um método automático não é o suficiente, por melhor que ele possa parecer. A verificação deve ser feita utilizando os RR's calculados juntamente com os sinais e pesos atribuídos a cada atributo e se a relação é condizente com os bons e maus pagadores. Um atributo com risco na categoria péssimo ou excelente não deve ter menor peso, consequentemente relevância, que um atributo com risco na categoria neutro, por exemplo.

3.7. Acurácia e Validação do Modelo

Três medidas de desempenho podem ser usadas para avaliar o melhor modelo (i) percentual de classificações corretas; (ii) o valor do teste de Kolmogorov-Smirnov (KS) para duas amostras; (iii) área abaixo da curva ROC (*receiver operating characteristic)*.

O percentual de acerto nas classificações deve ser avaliado pela comparação dos resultados observados e previstos pelo modelo. Desta forma, a taxa de acerto é medida pela divisão da quantidade de clientes corretamente classificados pelo total de clientes que fizeram parte da análise. Especialistas consideram satisfatórios os modelos com taxa de acerto superior a 65% (SELAU; RIBEIRO, 2009).

O índice KS tem por objetivo determinar a diferença máxima entre duas distribuições acumuladas. Esse indicador é calculado com base nas funções de distribuição acumulada dos bons e dos maus pagadores, sendo representado pela maior diferença entre essas funções (SICSÚ, 2010). Quanto maior o valor da medida, que é apresentada em percentual, melhor o ajuste do modelo. Tomando um valor de referência para o teste KS, Picinini *et al*. (2003, p. 465) sugerem: “o teste de Kolmogorov-Smirnov (KS) é utilizado no mercado financeiro como um dos indicadores de eficiência de modelos de credit scoring, sendo que o mercado considera um bom modelo aquele que apresente um valor de KS igual ou superior a 30”.

Segundo Sicsu (2010) e Pacheco (2012), a área sob a curva ROC se baseia na sensibilidade e na especificidade calculadas em relação a um escore, sendo a sensibilidade a capacidade de identificação dos maus créditos, e a especificidade a capacidade de identificação dos bons créditos. Ou seja, a proporção de classificações corretas do modelo, considerando-se toda a distribuição.

Porém, avaliar os resultados obtidos utilizando as variáveis que foram utilizadas em sua criação é tendencioso, visto o modelo já as conhecer. Estes índices de desempenho possivelmente sejam superestimados, não dando uma real ideia da assertividade do instrumento. Devido ao exposto, o modelo será novamente avaliado, contudo, por seu desempenho na amostra de validação, que são novas para a equação modelada. Refere-se à verificação do ajuste dentro da amostra comparado à previsão (fora da amostra). Em um momento posterior, faz-se nova medição dos índices de desempenho em uma situação mais próxima da realidade, quando a amostra de teste será utilizada, esta contendo a real configuração da base e dos solicitantes ao crédito, contendo não apenas clientes considerados bons ou maus, mas também os indefinidos e os que não utilizam o crédito.

4. Resultados

Nesta sessão serão apresentados os passos realizados para obtenção dos três modelos propostos e suas comparações, seguindo a ordem descrita pelo método na sessão anterior a fim de evidenciar realização de todas as etapas. Para todos os procedimentos práticos, desde tratamento do banco de dados até as análises e definição da equação de regressão, foi utilizado o *software* SPSS nas versões 18 e 21, tendo a ferramenta Excel do Pacote Office 2013 como apoio.

4.1. Planejamento e Definições

4.1.1. Definição do produto e do mercado

Os dados fornecidos para esta pesquisa são provenientes de uma rede de farmácias com unidades no estado do Rio Grande do Sul. As informações são de clientes para os quais foi oferecido o cartão de crédito do próprio grupo com o objetivo de facilitar a compra de produtos nas lojas da rede.

Sabendo haver um único produto financeiro, ou seja, apenas o cartão de crédito da loja, todas as informações constantes na base de todos os clientes que preencheram propostas serão, inicialmente, consideradas aptas para o estudo.

4.1.2. Finalidade do modelo e composições a serem comparadas

Ao se criar modelos preditivos de concessão de crédito espera-se que estes sejam um instrumento eficiente no auxílio à tomada de decisão, diminuindo o risco de conceder crédito a um cliente potencialmente inadimplente e, consequentemente, as perdas da empresa.

O estudo busca comparar três diferentes composições amostrais. Esta comparação é avaliada a partir de evidência empíricas baseada na capacidade preditiva / ajuste fora da amostra de cada uma das três especificações.

4.1.3. Definição da inadimplência / diferentes percentuais

Com o auxílio e conhecimento técnico de analistas da instituição embasados com a determinação de qualidade e ganhos/perdas a cada dia de atraso no pagamento, os clientes foram divididos em quatro grupos de acordo com o número de dias de atraso e utilização do crédito, são eles: (i) bons clientes, aqueles que quitam suas contas assumidas junto à empresa em dia ou com um atraso máximo de 30 dias; (ii) maus clientes, ou inadimplentes, aqueles que tiveram uma ou mais parcelas que, ou foram quitadas com mais de 90 dias de

atraso, ou após 90 dias ainda não foram quitadas; (iii) os clientes com atrasos entre 30 e 90 dias são classificados como indefinidos; (iv) o último grupo é formado pelo clientes que nunca utilizaram o crédito da rede no período de abrangência do estudo. Os clientes incluídos na amostra foram divididos em dois grupos: (i) e (ii). Os clientes dos itens (iii) e (iv) foram excluídos da amostra utilizada para a estimação do modelo para que haja um maior poder discriminativo.

Neste momento foram construídos três bancos de dados, para que seja possível proceder com a criação de três composições amostrais distintas, com percentuais diferentes de cada. Os percentuais serão os seguintes: (i) proporcional à população, em que serão mantidas as proporções de bons (63,7%) e maus (36,3%) clientes; (ii) com proporção de 50% de cada um dos dois grupos; (iii) proporção inversa da apresentada na base de dados, em que haverá 36,3% bons e 63,7% maus.

4.1.4. Definição do horizonte de previsão do modelo e verificação do histórico de crédito

Para avaliar o histórico dos usuários é necessária a espera de algum tempo para que se distinga seu comportamento e desempenho como pagador. Foram selecionados os cadastros que foram incluídos no banco de dados da empresa há pelo menos 12 meses para se adequar a este critério.

4.2. Identificação das Variáveis Previsoras

As variáveis a serem utilizadas no modelo são as já disponíveis na proposta de crédito da loja: sexo, naturalidade, data de nascimento, estado civil, profissão, grau de instrução, tipo de ocupação, tipo de residência, CEP residencial e comercial, data de admissão, paga pensão, possui filhos, data do cartão, crédito de terceiros e tipo de salário.

Há, ainda, a inclusão de variáveis computadas a partir das pré-existentes ou informadas pela empresa. São elas: idade, calculada a partir da data do cadastro juntamente com a data de nascimento; tempo de serviço, calculada a partir da data do cadastro juntamente com data de admissão; tipo de cliente, calculada a partir dos dias de atraso nos pagamentos; desconto no pagamento; maior atraso.

Ficaram disponíveis para utilização no modelo, após ajustes e inclusões, as variáveis apresentadas na Figura 3.

As variáveis renda, pensão, tipo de renda e crédito terceiros serão desconsideradas da análise, visto possuir nível baixo de preenchimento.

SEXO	PENSÃO
NATURALIDADE	POSSUI FILHOS
ESTADO CIVIL	DATA CARTÃO
DATA DE NASCIMENTO	CRÉDITO TERCEIROS
GRAU DE INSTRUÇÃO	TIPO DE SALÁRIO
PROFISSÃO	DESCONTO PAGAMENTO
TIPO DE OCUPAÇÃO	MAIOR ATRASO
TIPO DE RESIDÊNCIA	IDADE DO CLIENTE NO CADASTRO
CEP RESIDENCIAL	TEMPO SERVIÇO
CEP COMERCIAL	TIPO DE CLIENTE
DATA ADMISSÃO	

Figura 3: Variáveis aptas a fazer parte do modelo

4.3. Amostragem e Coleta dos Dados

4.3.1. Separação das amostras de análise, validação e teste

A amostra de maior importância é a utilizada para a estimação dos parâmetros do modelo, assim, a amostra de análise contará com 70% das observações disponíveis no banco de dados. Faz-se uma amostragem aleatória simples sem reposição da totalidade do banco e, após, a exclusão dos clientes considerados indefinidos ou sem utilização de crédito.

A amostra de validação tem por objetivo assegurar o funcionamento do modelo para observações novas, diferentes das já conhecidas pelo modelo. Para esta etapa são destinadas 20% das observações. Mantém-se o procedimento de excluir informações dos clientes dos grupos indeterminados e sem uso.

Os 10% restantes serão reservados para um teste com a configuração real da população. Neste momento os grupos excluídos nas demais amostras serão mantidos, a fim de entender o comportamento e proceder com medição de índices de qualidade do modelo em uma situação semelhante à originalmente existente.

4.3.2. Definição do período e tamanho da amostra para cada modelo

O tamanho amostral terá como referência o número de clientes do tipo mau, visto ser o grupo com menos observações.

Composição 64 - 36

A seleção dos indivíduos é feita de forma simples, apenas coletando por meio de amostragem aleatória simples sem reposição a porcentagem do banco de dados que ficará em cada amostra mantendo o mesmo percentual existente na base de dados. Apresentadas na Tabela 1 as amostras resultantes seguindo este processo.

Tabela 1: Amostras com percentual igual ao do banco de dados

		Composição 64 - 36			
		Análise	Validação	Teste	Total
Tipo	MAU	2861	715	397	3973
	BOM	5014	1254	696	6964
	IND	0	0	275	275
	S.USO	0	0	332	332
Total		7875	1969	1700	11544

Composição 50 - 50

A composição de igual tamanho é criada tendo como parâmetro o grupo de menor tamanho, neste caso, tipo de cliente mau. Tendo este número como referência, seleciona-se, a partir de uma amostra aleatória simples, o mesmo número de indivíduos do grupo maior, tipo de cliente bom.

A amostra de teste é a única que mantém-se com mesma proporção do banco de dado em todos os casos, pois, como anteriormente mencionado, objetiva refletir a situação mais próxima do original. Seguem resultados obtidos na Tabela 2.

Tabela 2: Amostras com igual percentual dos grupos bom e mau

		Composição 50 - 50			
		Análise	Validação	Teste	Total
Tipo	MAU	2861	715	397	3973
	BOM	2861	715	696	4272
	IND	0	0	275	275
	S.USO	0	0	332	332
Total		5722	1430	1700	8852

Composição 36 - 64

A composição amostral 36 - 64 deve ser calculada a partir da amostra de tamanhos iguais. Mantém-se, novamente, o grupo de maus clientes como referência, porém, neste momento, diminuindo o número de bons a fim de torna-lo menor. O valor fixado deve ser calculado por meio das proporções originais da população e invertendo nos grupos, teremos, assim, um grupo de bons com 36,3% das unidades selecionadas e o grupo de maus totalizará os 63,7% restantes nas amostras de análise e de validação. Na Tabela encontram-se 3 as quantidades obtidas em cada grupo.

Tabela 3: Amostras com percentual inverso ao do banco de dados

		Composição 36 - 64			
		Analise	Validacao	Teste	Total
Tipo	MAU	2861	715	397	3973
	BOM	1632	408	696	2736
	IND	0	0	275	275
	S.USO	0	0	332	332
Total		4493	1123	1700	7316

4.4. Análise dos Dados

4.4.1. Avaliação da consistência e preenchimento dos dados

Houve a necessidade do tratamento de problemas no banco, como *missings*, *outliers* e valores inconsistentes. Os *missings* foram padronizados em sua representação, visto ausência de resposta e ausência da característica serem informações distintas. As variáveis contínuas foram categorizadas em classes, limitando valores *outliers* e facilitando na interpretação das análises e criação de RR's. Os valores inconsistentes foram tratados por meio de agrupamento, quando possível, ou transformação em *missing*, como nos casos de valores absurdos, como idade negativa.

4.4.2. Escolha das variáveis que integrarão o modelo

Todas as varáveis disponíveis, apresentadas na Figura 3, subseção 4.2, foram consideradas aptas para a modelagem e entraram no processamento de criação de *dummies*, quando necessário.

4.5. Análise dicotômica

4.5.1. Agrupamento de atributos de variáveis

Com o uso de tabelas de contingência, calcula-se o risco relativo (RR) de cada variável, conforme método descrito na subseção 3.5, ao dividir o percentual de bons clientes pelo de maus, para que seja possível o agrupamento de atributos semelhantes. Seguindo a escala demonstrada na Figura 2, há variáveis com grande número de atributos, como naturalidade, simplificadas em no máximo sete.

Seguindo a ilustração utilizando a variável naturalidade, todas as cidades com risco semelhante, por exemplo, inferior a 0,5, estarão em uma mesma categoria (péssimo). Nesta classe estarão apenas as cidades com grande maioria de maus pagadores.

4.5.2. Criação das variáveis *dummies*

Foram geradas variáveis *dummies*, assumindo os valores 0 ou 1 (o cliente possui ou não a característica), para todas as variáveis categorizadas que possuíam mais de duas categorias, tendo, então, uma *dummy* para cada atributo. Essa sutileza auxilia na prevenção de problemas decorrentes da não linearidade dos atributos.

4.6. Obtenção da Fórmula Preliminar

4.6.1. Escolha de técnicas estatísticas multivariadas e Determinação do *software* a ser usado

O *software* SPSS foi utilizado nas análises com auxílio da ferramenta Excel. A estimação dos parâmetros de cada atributo para cada modelo foi realizada com base na regressão logística binária pelo método *stepwise*, com níveis de significância para a entrada e saída de variáveis do modelo de 5% e 10%, respectivamente, em conjunto com os RR's para validação dos resultados obtidos, seguindo uma coerência lógica. Não podemos, por exemplo, ter variáveis em grupos mais extremos de risco, como péssimo, apresentando menor influência na discriminação de perfil do que uma menos extrema, como neutro.

4.6.2. Seleção de variáveis independentes e Verificação da suposição das técnicas

Composição 64 - 36

Os atributos selecionados pelo método utilizado são verificados com o auxílio dos RR's calculados. Obteve-se a seleção de 29 variáveis independentes como significativas para discriminar o evento, dentre as 73 *dummies* relacionadas como possíveis.

Na Tabela 4 são apresentadas as variáveis significativas para a obtenção do escore do modelo na composição 64 - 36.

Tabela 4: Variáveis significativas para a composição 64 - 36

Atributo	**Variável *dummy***	**b estimado**	**p-valor**
Idade Classe 1	DIDAD1	-0,7401	0,000
Idade Classe 2	DIDAD2	-0,2549	0,003
Idade Classe 5	DIDAD5	0,3030	0,001
Idade Classe 6	DIDAD6	0,6100	0,000
Idade Classe 7	DIDAD7	0,9999	0,000
Idade Classe 8	DIDAD8	1,3323	0,000
Sexo Feminino	DSEXOF	0,3765	0,000
Ensino Fundamental	DFUN	-0,2386	0,000
Casado	DCASADO	0,4022	0,000
Separado	DSEPAR	-0,3620	0,025
Solteiro	DSOLTE	-0,1830	0,015
Tempo Serviço Classe 7	DTSERV7	0,5022	0,001
Tempo Serviço Classe 8	DTSERV8	0,6407	0,000
Possui Filho	DFILHO	-0,2671	0,000
Residência Alugada	DRES_ALU	-0,4166	0,000
Residência Outros	DRES_OUT	0,2910	0,002
Naturalidade Classe 1	DCIDNA1	-0,4555	0,019
Naturalidade Classe 2	DCIDNA2	-0,4343	0,000
Naturalidade Classe 3	DCIDNA3	-0,3592	0,002
Naturalidade Classe 7	DCIDNA7	0,5069	0,002
CEP Residencial Classe 2	DGCEPRE2	-0,6492	0,000
CEP Residencial Classe 3	DGCEPRE3	-0,2600	0,000
CEP Residencial Classe 6	DGCEPRE6	0,4660	0,000
CEP Residencial Classe 7	DGCEPRE7	0,5917	0,006
CEP Comercial Classe 1	DGCEPCO1	-0,4116	0,009
CEP Comercial Classe 7	DGCEPCO7	0,5353	0,012
Profissão Classe 1	DGPROF1	-0,8504	0,000
Profissão Classe 2	DGPROF2	-0,4070	0,000
Profissão Classe 3	DGPROF3	-0,1812	0,002
	Constante	0,5122	0,000

Composição 50 - 50

As variáveis selecionadas para a equação apresentaram alguns pontos de atenção que devem ser verificados: não consta no modelo preliminar a *dummy* mais extrema da variável CEP Residencial e constam outros níveis. O mesmo ocorre em CEP comercial. Porém, não trata-se de incoerência visto não haver informações nas categorias desconsideradas. As únicas *dummies* que receberam tratamento foram dos níveis 6 e 7 da variável profissão, agrupadas. Apresentadas na Tabela 5, as variáveis significativas a composição 50 – 50.

Tabela 5: Variáveis significativas para a composição 50 - 50

Atributo	Variável *dummy*	b estimado	p-valor
Idade Classe 1	DIDAD1	-0,6831	0,000
Idade Classe 2	DIDAD2	-0,2125	0,032
Idade Classe 5	DIDAD5	0,2278	0,029
Idade Classe 6	DIDAD6	0,6346	0,000
Idade Classe 7	DIDAD7	0,9957	0,000
Idade Classe 8	DIDAD8	1,4754	0,000
Sexo Feminino	DSEXOF	0,3553	0,000
Ensino Fundamental	DFUN	-0,1886	0,002
Casado	DCASADO	0,5436	0,000
Estato Civil Outros	DOUTR	0,4289	0,000
Tempo Serviço Classe 4	DTSERV4	-0,3681	0,028
Tempo Serviço Classe 8	DTSERV8	0,5225	0,010
Tempo Serviço Classe 9	DTSERV9	0,5649	0,048
Possui Filho	DFILHO	-0,2870	0,000
Ocupação Aposentado	DOCUP_AP	-0,4161	0,003
Ocupação Autônomo	DOCUP_AU	-0,1776	0,012
Residência Alugada	DRES_ALU	-0,4812	0,000
Residência Outros	DRES_OUT	0,3694	0,001
Naturalidade Classe 1	DCIDNA1	-0,6913	0,003
Naturalidade Classe 2	DCIDNA2	-0,4888	0,000
Naturalidade Classe 3	DCIDNA3	-0,3108	0,025
Naturalidade Classe 7	DCIDNA7	0,6338	0,001
CEP Residencial Classe 2	DGCEPRE2	-0,3577	0,000
CEP Residencial Classe 3	DGCEPRE3	-0,2146	0,005
CEP Residencial Classe 7	DGCEPRE7	0,3764	0,011
CEP Comercial Classe 1	DGCEPCO1	-0,6712	0,018
CEP Comercial Classe 2	DGCEPCO2	-0,2711	0,018
CEP Comercial Classe 5	DGCEPCO5	0,1783	0,063
CEP Comercial Classe 7	DGCEPCO7	0,6091	0,001
Profissão Classe 1	DGPROF1	-0,7406	0,000
Profissão Classe 2	DGPROF2	-0,3995	0,000
Profissão Classes 6 e 7	DGPROF67	0,4023	0,000
	Constante	-0,1943	0,072

Composição 36 - 64

Da mesma forma que no modelo preliminar obtido com a composição 50 - 50, as variáveis selecionadas para a equação apresentaram alguns pontos de atenção que devem ser verificados: o peso da variável *dummy* naturalidade classe 3 foi menor que para a *dummy* classe 2 deste atributo, porém, o RR possui relação inversa; para a variável referente ao tempo de serviço, há *dummy* mais extrema que não foi selecionada para o modelo, incoerente por discriminar melhor que as *dummies* menos extremas que foram incluídas; da mesma forma para CEP residencial.

A única correção necessária foi o agrupamento das *dummies* dos grupos mais extremos do atributo naturalidade (1 e 2). Nas demais *dummies* citadas não constaram os grupos intermediários no modelo por não haver nenhuma observação nos mesmos.

Demonstra-se na Tabela 6 as variáveis significativas para a obtenção da probabilidade estimada de um proponente vir a ser um bom cliente:

Figura 6: Variáveis significativas para a composição 36 - 64

Atributo	**Variável *dummy***	**b estimado**	**p-valor**
Idade Classe 1	DIDAD1	-0,6049	0,000
Idade Classe 5	DIDAD5	0,4856	0,000
Idade Classe 6	DIDAD6	0,7665	0,000
Idade Classe 7	DIDAD7	1,0697	0,000
Idade Classe 8	DIDAD8	1,4625	0,000
Sexo Masculino	DSEXOM	-0,3826	0,000
Ensino Superior	DSUP	0,4163	0,001
Casado	DCASADO	0,5936	0,000
Separado	DSEPAR	-0,4598	0,033
Naturalidade Classes 1 e 2	DCIDNA12	-0,3732	0,000
Naturalidade Classe 3	DCIDNA3	-0,3673	0,008
Naturalidade Classe 7	DCIDNA7	0,6228	0,000
Tempo Serviço Classe 0	DTSERV0	-0,2021	0,030
Tempo Serviço Classe 2	DTSERV2	-0,6831	0,005
Tempo Serviço Classe 8	DTSERV8	0,6509	0,006
Possui Filho	DFILHO	-0,3196	0,000
Paga Pensão	DPENSAO	-0,5731	0,035
Ocupação Autônomo	DOCUP_AU	-0,1392	0,086
Residência Alugada	DRES_ALU	-0,4744	0,000
CEP Residencial Classe 3	DGCEPRE3	-0,5163	0,000
CEP Residencial Classe 7	DGCEPRE7	0,4763	0,003
CEP Comercial Classe 2	DGCEPCO2	-0,3035	0,004
Profissão Classe 1	DGPROF1	-0,7294	0,001
	Constante	-0,1865	0,093

O sinal dos coeficientes de cada uma das variáveis indica o sentido para a classificação do tipo de cliente, sendo um indicativo de uma característica para um cliente mau o sinal negativo, e de um cliente bom o sinal positivo (SELAU; RIBEIRO, 2009).

O método *stepwise*, utilizado para selecionar automaticamente quais variáveis integrarão o modelo a ser estimado, leva em conta o atendimento ao pressuposto de não multicolinearidade devido ao método ter como mecanismo a inclusão de variáveis independentes com alto poder discriminatório e pouco correlacionadas entre si.

4.7. Acurácia e Validação do Modelo

4.7.1. Medição de índices de desempenho (KS, Curva ROC)

O teste não-paramétrico KS compara se duas amostras vêm de uma mesma população. O intuito é inferir se as duas amostras de clientes (bons e maus) provêm de populações distintas. Desta forma obtém-se evidências que o modelo está cumprindo seu objetivo de separar os dois grupos. A comparação dos valores do teste KS para as três composições amostrais consideradas é apresentada na Tabela 7.

Tabela 7: Resultados do Teste KS

Composição	Amostra		
	Análise	**Validação**	**Teste**
64 - 36	36,28%	36,19%	33,11%
50 - 50	37,02%	32,87%	37,29%
36 - 64	36,55%	35,96%	30,22%

Todos os valores encontrados para KS nas amostras de cada composição são maiores que 30%, valor mínimo para considerar um modelo com bom poder de separação, conforme referências apresentadas na seção 3.7.

A área sob a curva ROC mensura a relação entre a sensibilidade e a especificidade do teste, ou seja, o grau de acerto ou de erro. Quanto maior este índice, maior o percentual de acerto do modelo. A comparação dos valores da área sob a curva ROC para as três composições do modelo construído é apresentada na Tabela 8.

Tabela 8: Área sob a Curva ROC

Composição	Amostra		
	Análise	**Validação**	**Teste**
64 - 36	74,25%	73,38%	71,46%
50 - 50	74,33%	70,56%	73,52%
36 - 64	74,38%	73,18%	68,97%

Todos os valores encontrados para área nas amostras de cada modelo são maiores ou próximo a 70%, valor utilizado para considerar um modelo com boa classificação. (PACHECO,2012).

4.7.2. Percentual de classificações corretas

Uma outra forma de verificar o poder de previsão do modelo proposto é a medição do percentual de acerto nas classificações do modelo em suas diferentes composições amostrais. Previamente é necessário definir um ponto de corte na escala de escores em que divide-se os clientes em supostamente bons ou maus pagadores, ou ainda, separe os clientes em um grupo que receberá e outro que não receberá o crédito da empresa. Será utilizado o ponto de corte igual a 0,5, ou seja, se a probabilidade de pagamento for menor que 0,5 não é concedido o crédito. Nas Tabelas 9, 10, e 11 estão apresentadas as matrizes de classificação dos clientes para as composições amostrais 64 - 36, 50 - 50 e 36 - 64, respectivamente, para a amostra de análise.

Tabela 9: Matriz de classificação da Amostra de Análise Composição 64 - 36

			Classificação		
			MAU	BOM	Total
Tipo	MAU	Contagem	1312	1549	2861
		% dentro de tipo	45,9%	54,1%	100,0%
	BOM	Contagem	742	4272	5014
		% dentro de tipo	14,8%	85,2%	100,0%
Total		Contagem	2054	5821	7875
		% dentro de tipo	26,1%	73,9%	100,0%

Tabela 10: Matriz de classificação da Amostra de Análise Composição 50 - 50

			Classificação		
			MAU	BOM	Total
Tipo	MAU	Contagem	1981	880	2861
		% dentro de tipo	69,2%	30,8%	100,0%
	BOM	Contagem	927	1934	2861
		% dentro de tipo	32,4%	67,6%	100,0%
Total		Contagem	2908	2814	5722
		% dentro de tipo	50,8%	49,2%	100,0%

Tabela 11: Matriz de classificação da Amostra de Análise Composição 64 - 36

			Classificação		
			MAU	BOM	Total
Tipo	MAU	Contagem	2420	441	2861
		% dentro de tipo	84,6%	15,4%	100,0%
	BOM	Contagem	891	741	1632
		% dentro de tipo	54,6%	45,4%	100,0%
Total		Contagem	3311	1182	4493
		% dentro de tipo	73,7%	26,3%	100,0%

Conforme aumenta a proporção de maus nas composições amostrais, maior o percentual de acerto nesse grupo. Esta melhora na discriminação dos clientes inadimplentes pode ser interpretada como extremamente positiva para a análise de concessão de crédito, visto os clientes desse grupo trazerem prejuízos. Esse achado corrobora o que afirmam Steiner et al. (1999), a não concessão a um futuro inadimplente pode trazer mais benefício que o fornecimento de crédito a diversos outros bons pagadores.

A taxa de acerto geral das amostras de análise, validação e teste do modelo com composição amostral 64 - 36 foi, respectivamente, 70,91%, 69,22% e 67,80%. As mesmas taxas para as amostras com composição 50 - 50 foram 68,42%, 65,73% e 66,15% e para a composição 36 – 64 foram 70,35%, 70,53% e 59,01%. Tais valores são considerados bons, segundo referências na área, para modelos de previsão de risco de crédito.

4.7.3. Comparação de desempenho dos esquemas de composição amostral propostos

Pela Tabela 12 é possível observar que as três composições do modelo apresentaram índices aceitáveis e semelhantes tanto para o teste KS quanto para a área sob a curva ROC. Porém, no percentual de acerto as amostras de

análise das composições amostrais 64 - 36 e 64 - 36 apresentaram performance levemente superior à da composição 50 - 50. Para a amostra de validação a composição 36 - 64 foi superior às duas outras.

Sabendo que quaisquer melhorias na assertividade das decisões de concessão de crédito são relevantes para a empresa e tendo, de forma geral, obtido bons resultados nos três modelos, é preciso escolher aquele que potencialmente traga qualquer incremento ao negócio.

Tabela 12: Índices de desempenho do modelo nas três composições

Composição / Amostra	% acerto			KS			Curva ROC		
	Análise	Validação	Teste	Análise	Validação	Teste	Análise	Validação	Teste
64 - 36	70,91%	65,73%	66,15%	0,3628	0,3619	0,3311	0,7425	0,7338	0,7146
50 - 50	68,42%	65,73%	66,15%	0,3702	0,3287	0,3729	0,7433	0,7056	0,7352
36 - 64	70,35%	70,53%	59,01%	0,3655	0,3596	0,3022	0,7438	0,7318	0,6897

Tendo, ainda, o objetivo de diminuir o fornecimento de crédito aos maus pagadores, é possível comparar pela Tabela 13 qual modelo traz mais benefícios conciliando este critério.

Tabela 13: Índices de acerto do modelo nas três composições

Composição	Tipo de Cliente	Amostra		
		Análise	Validação	Teste
64 - 36	MAU	45,86%	43,50%	40,05%
	BOM	85,20%	83,89%	83,62%
50 - 50	MAU	69,24%	65,73%	68,01%
	BOM	67,60%	65,73%	65,09%
36 - 64	MAU	84,59%	84,76%	84,38%
	BOM	45,40%	45,59%	44,54%

Ao comparar o modelo mais assertivo considerando especificamente o grupo de inadimplentes, percebe-se que, quanto maior a proporção deste grupo na composição amostral, melhores foram os resultados. O modelo tendo composição 64 - 36 discriminou melhor os bons pagadores, com composição 50 - 50 discriminou de forma equilibrada bons e maus pagadores, já utilizando composição 36 - 64 discriminou melhor os maus pagadores.

5. Conclusões

Neste trabalho foi apresentada a sistemática para a construção de três esquemas de amostragem para a estimação de um modelo de regressão

logística, utilizando, para cada composição amostral, um percentual de bons e maus clientes distinto. Com o objetivo de encontrar a composição com melhor desempenho na predição de maus clientes, foram testados modelos com os seguintes percentuais de clientes inadimplentes: 36,3% - igual ao encontrado na população, 50% - grupos de igual tamanho e 63,7% - proporção inversa à encontrada na população. O modelo aqui detalhado poderá servir de auxílio para pesquisadores que desejam desenvolver esquemas de amostragem cujo público de menor volume precise ser melhor identificado.

Os três esquemas desenvolvidos tiveram desempenho satisfatório com assertividade na predição geral na amostra de análise, igual a 70,91%, 68,42% e 70,35%, para as amostragens em percentual 64 - 36, 50 - 50 e 36- 64, respectivamente. Por este critério observou-se uma pequena superioridade na divisão dos grupos seguindo a proporção real da base de dados, porém, a proporção inversa teve resultado muito próximo.

Observando a amostra de análise, obteve-se resultados semelhantes para o modelo com composições amostrais 64 - 36 e 64 - 36 tanto no Teste KS, 36,28 e 36,55, quanto na Área sob a Curva ROC, 74,25% e 74,38%, respectivamente, com o modelo tendo a composição 64 - 36 tendo uma pequena melhora, o que não se manteve nas amostras de validação e de teste, apesar de se manter uma pequena diferença para os dois modelos. De forma geral, estas medidas de desempenho foram satisfatórias para o modelo utilizando as três composições amostrais, conforme valores de referência utilizados.

A grande contribuição deste trabalho foi encontrar evidências de que o esquema de amostragem influi na capacidade preditiva do modelo, e mais que isso, as diferentes proporções de clientes dos dois grupos, bons e maus, pode facilitar a previsão do perfil de clientes do menor grupo. Quanto maior a proporção dos indivíduos que são minoria na população, neste caso os maus pagadores, maior a taxa de acerto no grupo. Na amostra de análise obteve-se taxa de acerto no grupo de maus pagadores para o modelo com composição 64 - 36 de 45,86%, 50 - 50 de 69,24% e 36 - 64 de 84,59%, o que mostra um grande ganho da identificação dos inadimplentes. Em contrapartida, o grupo dos bons pagadores teve sua taxa de acerto diminuída. Porém, geralmente, o interesse maior está na discriminação dos clientes que não pagam, visto que um único mau pagador pode acarretar a perda do ganho obtido em diversas outras transações bem sucedidas.

Portanto, a utilização de modelos de previsão de risco de crédito com maior percentual de maus pagadores na amostra permite melhor identificação deste perfil de clientes, ainda que na população eles sejam a menor parte.

Entende-se que a implementação desses modelos preditivos é facilitadora na tomada de decisão de concessão de crédito por eliminar a subjetividade da análise possibilitando a criação de uma regra que diminua de forma considerável o tempo de resposta para cada proposta de crédito e reduza a inadimplência ao fornecer crédito a clientes melhores qualificados.

REFERÊNCIAS

ARAUJO, E. A.; CAMONA, C. U. M. **Desenvolvimento de Modelos Credit Scoring com Abordagem de Regressão Logística para a Gestão da Inadimplência de uma Instituição de Microcrédito.** 2007.

BROWN, I.; MUES, C. **An experimental comparison of classification algorithms for imbalanced credit scoring data sets.** University of Southampton, UK. 2012.

BUENO, V. F. F. **Avaliação de risco na concessão de crédito bancário para micros e pequenas empresas.** Florianópolis: UFSC, 2003. Dissertação (Mestrado em Engenharia da Produção), Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, 2003.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise multivariada: para cursos de administração, ciências contábeis e economia.** São Paulo: Atlas, 2007.

CRONE, S. F.; FINLAY, S. **Instance sampling in credit scoring: An empirical study of sample size and balancing.** Lancaster University, UK. 2012.

GHODSELAHI, A. A hybrid support vector machine ensemble model for credit scoring**. International Journal of Computer Applications**, v.17, n.5, p. 975-8887, March 2011.

GOUVÊA, M. A.; GONÇALVES, E. B. Análise de Risco de Crédito com o Uso de Modelos de Redes Neurais e Algoritmos Genéticos. In: IX SEMEAD – **Seminários em Administração** FEA-USP, 2006, São Paulo. Anais.

GUJARATI, D. N. – **Econometria Básica**, Makron Books, 2000.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise multivariada de dados.** 5.ed. Porto Alegre: Bookman, 2005.

HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression.** New York: John Wiley & Sons, 1989.

LAWRENCE, D. **Handbook of consumer lending.** New Jersey: Prentice Hall. 1992.

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

LOUZADA, F.; AMARAL, G. J. A.; GUIRADO, L.; SILVA, P. H. F.; ABREU, H. J.; FERREIRA, M. R. P. Medidas estatísticas da capacidade preditiva de modelos de classificação em credit scoring. **P@rtes** (São Paulo), v. 68, p.7-28, 2008.

MENDES FILHO, E. F.; CARVALHO, A. C. P. L. F.; MATIAS, A. B. Utilização de redes neurais artificiais na análise de risco de crédito a pessoas físicas. In: **III Simpósio Brasileiro de Redes Neurais**, Recife. Anais. 1996.

MORAES, L. G. **Uma Abordagem Alternativa de Behavioral Scoring Usando Modelagem Híbrida de Dois Estágios: Regressão Logística e Redes Neurais.** Porto Alegre: UFRGS. 2012.

MORGAN, J. P.; **Relatório de Gerenciamento de Riscos (Pilar III).** Prospecto de 31 de Março de 2013.

PACHECO, J. S. **Utilizando análise de sobrevivência no crédito: um modelo de behavioral scoring.** Porto Alegre: UFRGS. 2012.

PICININI, R.; OLIVEIRA, G. M. B.; MONTEIRO, L. H. A. Mineração de Critério de Credit Scoring Utilizando Algoritmos Genéticos. In: **VI Simpósio Brasileiro de Automação Inteligente**, 2003, Bauru.

SELAU, L. P. R. **Construção de modelos de previsão de risco de crédito.** Porto Alegre: UFRGS, 2008. Dissertação (Mestrado em Engenharia da Produção), Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Sul, 2008.

SELAU, L. P. R. **Modelagem para Concessão de Crédito a Pessoas Físicas em empresas Comerciais: Da Decisão Binária para a Decisão Monetária.** Porto Alegre: UFRGS, 2012. Dissertação (Doutorado em Administração), Programa de Pós-Graduação em Administração, Universidade Federal do Rio Grande do Sul, 2012.

SELAU, L. P. R., RIBEIRO, J. L. D. Uma sistemática para construção e escolha de modelos de previsão de risco de crédito. **Gestão Produção**, v. 16, n. 3, p. 398-413, jul.-set. 2009.

SELAU, L. P. R.; RIBEIRO, J. L. D. Systematic approach to construct credit risk forecast models. **Pesquisa Operacional**, v.31, p.1-17, 2011.

SICSÚ, A. L. Credit scoring: desenvolvimento de um sistema de credit scoring – Parte II. **Tecnologia de Crédito**. São Paulo: Serasa, n.5, 1998.

SICSÚ, A. L. **Desenvolvimento, implantação, acompanhamento.** São Paulo: Blucher, 2010.

SILVA, J. P. **Gestão e análise de risco de crédito.** 4 ed. São Paulo: Atlas, 2003.

STEINER, M. T. A.; CARNIERI, C.; KOPITTKE, B. H.; STEINER NETO, P. J. Sistemas especialistas probabilísticos e redes neurais na análise do crédito bancário. **Revista de Administração**, v.34, n.3, p.56-67, 1999.

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

APÊNDICE A - AGRUPAMENTO DE CEP RESIDENCIAL

Composição 64 - 36

		2 PRIMEIRAS POSIÇÕES	3 PRIMEIRAS POSIÇÕES
1	**Péssimo Desempenho**		
2	**Muito Mau Desempenho**		908 PORTO ALEGRE 919 PORTO ALEGRE 923 PORTO ALEGRE
3	**Mau Desempenho**		906 PORTO ALEGRE 912 PORTO ALEGRE 915 PORTO ALEGRE 917 PORTO ALEGRE 925 GUAIBA 932 SAPUCAIA DO SUL 934 NOVO HAMBURGO 940 GRAVATAI 948 ALVORADA 961 PELOTAS
4	**Desempenho Neutro**		902 PORTO ALEGRE 911 PORTO ALEGRE 914 PORTO ALEGRE 935 NOVO HAMBURGO 941 GRAVATAI 944 VIAMAO 949 CACHOEIRINHA 956 TAQUARA 957 BENTO GONCALVES 975 URUGUAIANA
5	**Bom Desempenho**		933 NOVO HAMBURGO 945 VIAMAO 955 OSORIO 960 PELOTAS 964 BAGE 967 SAO JERONIMO
6	**Muito Bom Desempenho**	99 PASSO FUNDO - RS	922 PORTO ALEGRE 924 PORTO ALEGRE 937 CAMPO BOM 938 SAPIRANGA 950 CAXIAS DO SUL 958 ESTRELA 962 RIO GRANDE
7	**Excelente Desempenho**		965 CACHOEIRA DO SUL 966 RIO PARDO 968 STA CRUZ DO SUL 973 SAO GABRIEL 988 SANTO ANGELO 989 SANTA ROSA

Composição 50 - 50

		2 PRIMEIRAS POSIÇÕES	3 PRIMEIRAS POSIÇÕES
1	**Péssimo Desempenho**		
2	**Muito Mau Desempenho**	90 PORTO ALEGRE - RS	917 PORTO ALEGRE
3	**Mau Desempenho**		912 PORTO ALEGRE 915 PORTO ALEGRE 919 PORTO ALEGRE 944 VIAMAO 948 ALVORADA 949 CACHOEIRINHA 961 PELOTAS
4	**Desempenho Neutro**	92 CANOAS - RS	932 SAPUCAIA DO SUL 934 NOVO HAMBURGO 940 GRAVATAI 941 GRAVATAI 967 SAO JERONIMO
5	**Bom Desempenho**		933 NOVO HAMBURGO 935 NOVO HAMBURGO 938 SAPIRANGA 955 OSORIO 956 TAQUARA 957 BENTO GONCALVES 960 PELOTAS
6	**Muito Bom Desempenho**	99 PASSO FUNDO - RS	950 CAXIAS DO SUL 962 RIO GRANDE
7	**Excelente Desempenho**	97 SANTA MARIA - RS 98 CRUZ ALTA - RS	937 CAMPO BOM 965 CACHOEIRA DO SUL 966 RIO PARDO 968 STA CRUZ DO SUL

Composição 36 - 64

		2 PRIMEIRAS POSIÇÕES	3 PRIMEIRAS POSIÇÕES
1	**Péssimo Desempenho**		
2	**Muito Mau Desempenho**		
3	**Mau Desempenho**	90 PORTO ALEGRE - RS 91 PORTO ALEGRE - RS 92 CANOAS - RS 94 GRAVATAI - RS	932 SAPUCAIA DO SUL 934 NOVO HAMBURGO 961 PELOTAS
4	**Desempenho Neutro**		935 NOVO HAMBURGO 967 SAO JERONIMO
5	**Bom Desempenho**		933 NOVO HAMBURGO 938 SAPIRANGA 955 OSORIO 960 PELOTAS 975 URUGUAIANA
6	**Muito Bom Desempenho**		956 TAQUARA 957 BENTO GONCALVES
7	**Excelente Desempenho**	98 CRUZ ALTA - RS	950 CAXIAS DO SUL 965 CACHOEIRA DO SUL 973 SAO GABRIEL

APÊNDICE B - AGRUPAMENTO DE CEP COMERCIAL

Composição 64 - 36

		2 PRIMEIRAS POSIÇÕES	3 PRIMEIRAS POSIÇÕES
1	**Péssimo Desempenho**		912 PORTO ALEGRE 948 ALVORADA
2	**Muito Mau Desempenho**		915 PORTO ALEGRE 919 PORTO ALEGRE 932 SAPUCAIA DO SUL 961 PELOTAS
3	**Mau Desempenho**		900 PORTO ALEGRE 902 PORTO ALEGRE 906 PORTO ALEGRE 908 PORTO ALEGRE 910 PORTO ALEGRE 911 PORTO ALEGRE 913 PORTO ALEGRE 917 PORTO ALEGRE 925 GUAIBA 949 CACHOEIRINHA
4	**Desempenho Neutro**		901 PORTO ALEGRE 904 PORTO ALEGRE 905 PORTO ALEGRE 933 NOVO HAMBURGO 935 NOVO HAMBURGO 940 GRAVATAI 944 VIAMAO 956 TAQUARA 967 SAO JERONIMO 975 URUGUAIANA
5	**Bom Desempenho**		934 NOVO HAMBURGO 938 SAPIRANGA 955 OSORIO 957 BENTO GONCALVES 960 PELOTAS 962 RIO GRANDE
6	**Muito Bom**	99 PASSO FUNDO - RS	958 ESTRELA
7	**Excelente Desempenho**		937 CAMPO BOM 950 CAXIAS DO SUL 966 RIO PARDO 968 STA CRUZ DO SUL 973 SAO GABRIEL 988 SANTO ANGELO 989 SANTA ROSA

Composição 50 - 50

		2 PRIMEIRAS POSIÇÕES	3 PRIMEIRAS POSIÇÕES
1	**Péssimo Desempenho**		912 PORTO ALEGRE 915 PORTO ALEGRE
2	**Muito Mau Desempenho**		906 PORTO ALEGRE 908 PORTO ALEGRE 910 PORTO ALEGRE 932 SAPUCAIA DO SUL 948 ALVORADA 961 PELOTAS
3	**Mau Desempenho**	92 CANOAS - RS	900 PORTO ALEGRE 901 PORTO ALEGRE 904 PORTO ALEGRE 913 PORTO ALEGRE 917 PORTO ALEGRE 919 PORTO ALEGRE 944 VIAMAO 949 CACHOEIRINHA 967 SAO JERONIMO
4	**Desempenho Neutro**		902 PORTO ALEGRE 911 PORTO ALEGRE 935 NOVO HAMBURGO 940 GRAVATAI
5	**Bom Desempenho**		905 PORTO ALEGRE 933 NOVO HAMBURGO 934 NOVO HAMBURGO 938 SAPIRANGA 955 OSORIO 956 TAQUARA 957 BENTO GONCALVES 960 PELOTAS
6	**Muito Bom Desempenho**		
7	**Excelente Desempenho**	97 SANTA MARIA - RS 98 CRUZ ALTA - RS 99 PASSO FUNDO - RS	937 CAMPO BOM

Composição 36 - 64

		2 PRIMEIRAS POSIÇÕES	3 PRIMEIRAS POSIÇÕES
1	**Péssimo Desempenho**		948 ALVORADA
2	**Muito Mau Desempenho**	91 PORTO ALEGRE - RS	900 PORTO ALEGRE 906 PORTO ALEGRE 932 SAPUCAIA DO SUL 949 CACHOEIRINHA
3	**Mau Desempenho**	92 CANOAS - RS	902 PORTO ALEGRE 908 PORTO ALEGRE 935 NOVO HAMBURGO 944 VIAMAO
4	**Desempenho Neutro**		933 NOVO HAMBURGO 961 PELOTAS 967 SAO JERONIMO
5	**Bom Desempenho**		901 PORTO ALEGRE 904 PORTO ALEGRE 905 PORTO ALEGRE 938 SAPIRANGA 960 PELOTAS
6	**Muito Bom**	95 CAXIAS DO SUL 97 SANTA MARIA - RS	
7	**Excelente Desempenho**	98 CRUZ ALTA - RS	

APÊNDICE C - AGRUPAMENTO DE NATURALIDADE

Composição 64 - 36

1	**Péssimo Desempenho**	ALVORADA CRUZ ALTA ESTEIO SAPUCAIA DO SUL
2	**Muito Mau Desempenho**	GRAVATAI IJUI PORTO ALEGRE TRAMANDAI
3	**Mau Desempenho**	CAMAQUA CAMPO BOM CANOAS GUAIBA NOVO HAMBURGO RIO GRANDE SANTANA DO SAO LEOPOLDO URUGUAIANA
4	**Desempenho Neutro**	ALEGRETE SAO BORJA TRES PASSOS
5	**Bom Desempenho**	CACAPAVA DO SUL MONTENEGRO OSORIO PASSO FUNDO PELOTAS SANTA CRUZ DO SUL SANTA ROSA SANTO ANGELO SAO GABRIEL SAO JERONIMO SAO LOURENCO DO SUL SAO LUIZ GONZAGA SAPIRANGA VIAMAO
6	**Muito Bom Desempenho**	BAGE BUTIA CACHOEIRA DO SUL CAXIAS DO SUL HORIZONTINA PALMEIRA DAS MISSOES RIO PARDO SANTA MARIA SAO FRANCISCO DE
7	**Excelente Desempenho**	CANGUCU ENCRUZILHADA DO SUL GIRUA ROLANTE SANTA VITORIA DO SANTO ANTONIO DA SAO SEPE TAPES TAQUARA TORRES TRES DE MAIO

Composição 50 - 50

1	**Péssimo Desempenho**	ALVORADA
2	**Muito Mau Desempenho**	CRUZ ALTA ESTEIO PORTO ALEGRE SANTANA DO LIVRAMENTO SAO LEOPOLDO URUGUAIANA
3	**Mau Desempenho**	ALEGRETE CAMAQUA GRAVATAI RIO GRANDE
4	**Desempenho Neutro**	CANOAS GUAIBA NOVO HAMBURGO PELOTAS SAPIRANGA
5	**Bom Desempenho**	CAXIAS DO SUL OSORIO PALMEIRA DAS MISSOES SANTO ANGELO SAO FRANCISCO DE PAULA SAO JERONIMO VIAMAO
6	**Muito Bom Desempenho**	BAGE CACHOEIRA DO SUL MONTENEGRO RIO PARDO SANTA MARIA SANTA ROSA SAO GABRIEL TAQUARA TRES DE MAIO
7	**Excelente Desempenho**	CANGUCU SANTO ANTONIO DA PATRUL SAO LOURENCO DO SUL SAO SEPE

Composição 36 - 64

1	**Péssimo Desempenho**	ALVORADA ESTEIO GRAVATAI SAO LEOPOLDO
2	**Muito Mau Desempenho**	PORTO ALEGRE
3	**Mau Desempenho**	CANOAS NOVO HAMBURGO SANTANA DO LIVRAMENTO SAO JERONIMO
4	**Desempenho Neutro**	PELOTAS SAO BORJA URUGUAIANA VIAMAO
5	**Bom Desempenho**	BAGE CAMAQUA SANTA CRUZ DO SUL SANTA MARIA SANTO ANGELO SAPIRANGA
6	**Muito Bom Desempenho**	ALEGRETE CACHOEIRA DO SUL MONTENEGRO PALMEIRA DAS MISSOES
7	**Excelente Desempenho**	CANGUCU SANTA ROSA SANTO ANTONIO DA PATRUL SAO LOURENCO DO SUL SAO SEPE

APÊNDICE D - AGRUPAMENTO DE PROFISSÕES

Composição 64 - 36

1	**Péssimo Desempenho**	ALMOXARIFE CABELEIREIRO GERENTE PINTOR PROMOTOR VENDAS TELEFONISTA
2	**Muito Mau Desempenho**	AUX COZINHA BABA MANICURE MECANICO PEDREIRO PORTEIRO RECEPCIONISTA VENDEDOR
3	**Mau Desempenho**	AUTONOMO AUX ADMINISTRATIVO AUX PRODUCAO AUX SERVICOS GERAIS AUXILIAR COMERCIANTE COZINHEIRO MOTORISTA VIGILANTE
4	**Desempenho Neutro**	ATENDENTE CAIXA DIARISTA DOMESTICA SERVENTE TEC ENFERMAGEM
5	**Bom Desempenho**	AGENTE BALCONISTA COSTUREIRO DO LAR INDUSTRIARIO OPERADOR PENSIONISTA SECRETARIA
6	**Muito Bom Desempenho**	COMERCIARIO FAXINEIRA
7	**Excelente Desempenho**	AGRICULTOR APOSENTADO PROFESSOR

Composição 50 - 50

1	**Péssimo Desempenho**	AUX COZINHA BABA GERENTE PEDREIRO PORTEIRO
2	**Muito Mau Desempenho**	CABELEIREIRO COZINHEIRO SERVENTE VENDEDOR
3	**Mau Desempenho**	AUTONOMO AUX PRODUCAO AUX SERVICOS GERAIS AUXILIAR CAIXA DIARISTA VIGILANTE
4	**Desempenho Neutro**	ATENDENTE AUX ADMINISTRATIVO COMERCIANTE DOMESTICA TEC ENFERMAGEM
5	**Bom Desempenho**	COMERCIARIO COSTUREIRO INDUSTRIARIO MOTORISTA SECRETARIA
6	**Muito Bom Desempenho**	AGRICULTOR BALCONISTA DO LAR OPERADOR PENSIONISTA PROFESSOR
7	**Excelente Desempenho**	APOSENTADO

Composição 36 - 64

1	**Péssimo Desempenho**	ATENDENTE GERENTE PEDREIRO PORTEIRO
2	**Muito Mau Desempenho**	AUX ADMINISTRATIVO COZINHEIRO DIARISTA VENDEDOR
3	**Mau Desempenho**	AUTONOMO AUX PRODUCAO AUX SERVICOS GERAIS SECRETARIA VIGILANTE
4	**Desempenho Neutro**	BALCONISTA CAIXA DOMESTICA MOTORISTA OPERADOR
5	**Bom Desempenho**	AUXILIAR COMERCIANTE COSTUREIRO DO LAR PENSIONISTA
6	**Muito Bom Desempenho**	
7	**Excelente Desempenho**	APOSENTADO PROFESSOR