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

Combinação em série e em paralelo de modelos de Redes Neurais e Regressão Logística – Um estudo de caso em *Cross-Selling*

Autora: Sabrina Zanatta Grebin
Orientadora: Professora Dra. Lisiane Priscila Roldão Selau

Porto Alegre, 16 de janeiro de 2013.

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

Combinação em série e em paralelo de modelos de Redes Neurais e Regressão Logística – Um estudo de caso em *Cross-Selling*

Autora: Sabrina Zanatta Grebin

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professora Dra. Lisiane Priscila Roldão Selau
Professora Dra. Liane Werner

Porto Alegre, 16 de janeiro de 2013.

Dedico este trabalho aos meus pais, Mirtes e Valdemar, aos meus irmãos, Tiago e Mairus, e ao meu namorado, Fábio.

Agradecimentos

Aos meus pais, Mirtes e Valdemar e aos meus irmãos, Tiago e Mairus: com certeza a parte mais difícil foi ficar longe de vocês, muito obrigada por tudo que vocês representam na minha vida, toda a força que eu precisei veio de vocês. Mãe, tu és o meu exemplo, muito obrigada por nunca medir esforços para nos fazer felizes. Ao meu namorado, Fábio, por estar sempre ao meu lado em todos os momentos, acompanhando todos os meus erros e acertos, por todo o cuidado, carinho, amor e incentivo: Obrigado por fazer parte da minha vida, e por torná-la ainda mais feliz. Mãe, Pai, Tiago, Mairus e Fábio, amo vocês!

Aos meus sogros, Tânia e Gilmar, e ao meu cunhado, Bruno, por toda a paciência e por terem me acolhido com muito carinho. Aos colegas da Estatística, em especial à Natalia Giordani, Pricila Henckes e Tássia Henckes, por terem se tornado muito mais do que colegas, dividindo momentos de muita alegria, muito estudo e muitas dores de barriga pré-provas também. À Andriara, Flávia, Andrigo, aos ex-colegas de LOPP, em especial à Verinha e ao Belleza, aos colegas de trabalho e a todos que, de alguma forma, contribuíram para o meu crescimento e para a concretização de mais esta etapa.

Aos professores do Departamento de Estatística por todo o conhecimento transmitido, em especial à professora Lisiane, pela orientação neste trabalho e por toda paciência e confiança, e às professoras Liane e Marcia pelos ensinamentos durante as bolsas de iniciação científica. À instituição que forneceu os dados para este trabalho e ao SAS Institute, pela disponibilização da licença do *software* para que este trabalho pudesse ser executado.

Resumo

Como resultado ao crescente desenvolvimento tecnológico nas últimas décadas, computadores cada vez mais potentes tornam possível o armazenamento diário de grande quantidade de dados. As técnicas de mineração de dados surgem como uma alternativa inteligente e eficaz para transformar essa grande massa de dados em conhecimento. Este trabalho se propôs a resolver um problema real de *cross-selling* de uma instituição financeira brasileira e, com o objetivo de contribuir para o desenvolvimento das técnicas de *data mining*, realizou-se uma comparação entre duas técnicas consagradas na área, regressão logística e redes neurais, e entre duas formas de combinação das mesmas, em série (*hybrid*), onde a regressão logística é utilizada para selecionar as variáveis que irão entrar na rede neural, e em paralelo (*ensemble*), onde os resultados das técnicas individuais são combinados com base em suas decisões. As comparações entre o desempenho das técnicas individuais e dos métodos de combinação indicam que, para este estudo, as duas técnicas e os dois métodos utilizados obtiveram desempenho similar, porém os dois métodos de combinação apresentaram melhor desempenho.

Palavras-chave: *Cross-Selling*, Regressão Logística, Redes Neurais, *Hybrid*, *Ensemble*.

Sumário

Este artigo será submetido à “REVISTA BRASILEIRA DE ESTATÍSTICA”

1. INTRODUÇÃO

Como resultado ao crescente desenvolvimento tecnológico nas últimas décadas, computadores cada vez mais potentes tornam possível o armazenamento diário de grande quantidade de dados. Empresas dos mais diversos setores e de todos os tamanhos buscam transformar essa grande massa de dados em informação útil para a tomada de decisão. Para tanto, são necessárias ferramentas e técnicas capazes de extrair esse conhecimento. Devido a crescente concorrência do mercado, empresas buscam antecipar as necessidades e preferências dos seus clientes. Segundo Barry e Linoff (2004), para melhorar o relacionamento com os clientes é preciso aprender o seu comportamento, para que com base nesse conhecimento seja possível aumentar a rentabilidade do negócio, de modo a se obter clientes mais satisfeitos e fiéis.

Desde 1960, o processamento de dados vem sendo migrado para sistemas cada vez mais sofisticados e poderosos, evoluindo, na década de 1970, para sistemas relacionais e seguindo em direção ao desenvolvimento de sistemas de banco de dados avançados (*advanced database systems*), armazenamento de dados (*data warehousing*) e mineração de dados (*data mining*), surgindo, na década de 1980, com análises avançadas de dados (HAN *et al*., 2012).

Simultaneamente ao avanço tecnológico, o constante aumento dos conjuntos de dados e, consequentemente, das variáveis que estão sujeitas a se relacionarem, vêm tornando os métodos tradicionais de análise menos eficientes e, por vezes, inadequados (SILVA, 2009). As técnicas de *data mining* surgem então como uma alternativa inteligente e eficaz de identificar padrões de comportamento dos dados, transformando-os em conhecimento. Segundo Berry e Linoff (2004), elas tem por finalidade a classificação, estimação, previsão, agrupamento e descrição.

Segundo Fayyad *et al.*(1996), dentre as finalidades da mineração de dados, as duas mais utilizadas na prática são: a previsão, onde o valor previsto para determinada variável de interesse é explicado pelas variáveis presentes no banco de dados; e a descrição, onde são descritos os padrões existentes nas relações entre elas. Não raro, o objetivo está tanto em encontrar o padrão de relacionamento entre as variáveis quanto em prever um valor futuro para as mesmas. Alguns exemplos são: modelos de venda para estratégias de marketing, modelos de risco de crédito, segmentação de clientes, previsão de inadimplência e controle de fraudes.

Neste estudo, o foco da estratégia de *data mining* é a venda-cruzada (*cross-selling*) que, segundo Dyche (2001), se caracteriza pela venda de um produto ou serviço adicional como resultado de uma compra anterior e, quando corretamente realizada, significa vender o produto certo ao cliente certo. É consenso no marketing de relacionamento que é menos dispendioso e mais rentável investir mais nos clientes atuais do que adquirir um novo cliente.

Entre as técnicas mais utilizadas para a previsão e descrição estão a regressão logística, análise discriminante, árvores de decisão e redes neurais. Sendo que regressão logística e redes neurais têm sido as mais utilizadas (SELAU, 2012). Com o intuito de se obter ganhos no desempenho e na explicação do modelo obtido através dessas técnicas, autores vêm estudando formas de combiná-las. Dentre elas, a combinação em série, também conhecida como modelo híbrido *(hybrid*), e a combinação em paralelo, ou simplesmente combinação de previsões (*ensemble*). Segundo Selau (2012), a combinação em série consiste em utilizar em sequência duas técnicas distintas, com o intuito de melhorar o seu poder preditivo. A combinação em paralelo, que, segundo Werner (2005) vem sendo estudada desde Bates e Granger (1969), é um método bastante utilizado para a previsão em séries temporais com o intuito de minimizar os seus erros e consiste basicamente em combinar as previsões obtidas através de duas ou mais técnicas individuais.

Nesse contexto, o objetivo deste trabalho é comparar o desempenho individual das técnicas de regressão logística e redes neurais e de duas formas de combinação: em série, onde a regressão logística é utilizada para selecionar as variáveis que irão entrar na rede neural, e em paralelo, onde os resultados das técnicas individuais são combinados com base em suas decisões. Para tanto, este trabalho se propõe a resolver um problema real de *cross-selling* de uma instituição financeira brasileira.

Este trabalho está estruturado em cinco seções, sendo a primeira a introdução já exposta, onde foram apresentadas as considerações iniciais a cerca do problema de pesquisa. Na segunda seção são apresentados os conceitos e a fundamentação teórica dos modelos de *cross-selling*, das técnicas de modelagem e dos métodos de combinação utilizados no desenvolvimento deste trabalho. Na terceira seção é descrita a metodologia de pesquisa, bem como as etapas propostas para a construção dos modelos. A seção quatro ilustra os resultados e comparações obtidas através do método proposto. Por fim, as principais conclusões e considerações são apresentadas na seção cinco, além de sugestões para trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção é apresentada uma revisão, com base na literatura, sobre as técnicas e métodos a serem utilizados no desenvolvimento deste trabalho.

2.1 Modelos de *cross-selling*

Para Berry e Linoff (2004), é preciso que as grandes empresas imitem as pequenas, criando um relacionamento com seus clientes, baseado no conhecimento sobre eles, para melhor atendê-los. Segundo HAN *et al*. (2012), no processo de descoberta de conhecimento em base de dados, conhecido por KDD (*Knowledge Discovery in Databases*), primeiramente os dados são preparados para mineração (que inclui limpeza dos dados, combinação com outras bases de dados, seleção e transformação de variáveis), então é realizada a mineração de dados e os padrões descobertos são transformados em conhecimento.

Para se estabelecer um relacionamento de longo prazo com o mercado, duas técnicas são utilizadas no Marketing de Relacionamento: *up-selling* e *cross-selling*. O conceito de *up-selling* é aumentar o valor da venda de um mesmo produto, já os modelos de *cross-selling*, traduzidos na literatura como "venda casada de produtos" ou simplesmente "venda-cruzada", buscam nos dados o conhecimento necessário para identificar o perfil de clientes mais propensos a adquirir um ou mais produtos adicionais (BERRY e LINOFF, 2004).

O objetivo ao se realizar uma modelagem de *cross-selling* é concentrar mais esforços nos clientes atuais, cujo custo de aquisição é menor em relação a clientes novos, porque os clientes atuais já possuem um relacionamento com a empresa. O uso adequado da técnica consiste em identificar que produto ou serviço oferecer a qual cliente e em que momento, com o intuito de que estes venham a adquirir mais produtos ou serviços com a empresa, se tornando clientes fiéis. Consequentemente, o bom emprego de *cross-selling* implica em um aumento da satisfação dos clientes, da sua fidelidade e lucratividade (KAMAKURA *et al*., 1991).

Assim como nas outras aplicações de *data mining*, os trabalhos em *cross-selling* também utilizam com bastante frequência as técnicas de regressão logística e de redes neurais. Como, por exemplo, Kisahleitner (2008) que compara o desempenho das técnicas de regressão logística, árvores de decisão e redes neurais na aquisição de

clientes não correntistas para cartão de crédito de uma empresa do setor financeiro. Os resultados mostram que, para este estudo, as três técnicas utilizadas obtiverem desempenho similar. Silva (2009) e Adorno e Bueno (2011) fazem uso das técnicas de redes neurais e regressão logística que são comparadas para desenvolver um modelo de propensão ao consumo de um produto de crédito pessoal. Ambos destacam que as diferenças entre os melhores modelos originados de cada técnica não são muito significativas, porém o modelo de redes neurais obteve desempenho melhor. Knott *et al*. (2002) utilizaram-se de dados de um banco de varejo com o objetivo de aumentar a quantidade de produtos adquiridos por clientes, prevendo qual é o próximo produto que cada cliente possui maior probabilidade de comprar. Os autores utilizaram regressão logística, logit multinomial, análise discriminante e redes neurais com o intuito de desenvolver um modelo de c*ross-selling* com a abordagem de *next-product-to-buy* (NPTB).

2.2 Técnicas de Modelagem Individuais

2.2.1 *Regressão Logística*

O modelo de regressão logística se caracteriza por ter como resposta uma variável dicotômica, sendo essa a diferença entre este e o modelo de regressão linear (HOSMER e LEMESHOW, 1989). O objetivo da técnica é modelar a proporção de uma das duas categorias da variável resposta, onde a presença (1) ou ausência (0) de uma característica é modelada em função das variáveis explicativas (DINIZ e LOUZADA, 2012). A regressão logística possui muitas vantagens frente à análise discriminante quando se tem um modelo com resposta dicotômica, pois não depende de suposições tão rígidas, tais como a normalidade das variáveis explicativas e a igualdade de matrizes de variâncias e covariâncias dos grupos (HAIR *et al.,*2005).

A regressão logística surgiu nos anos 80 e, segundo Hosmer e Lemeshow (1989), se tornou o método mais utilizado quando se tem uma resposta dicotômica. De acordo com Hair *et al*. (2005), ela possui muitas similaridades com a regressão linear, como, por exemplo, os testes estatísticos e a capacidade de inserir efeitos não lineares no modelo (com a criação de variáveis *dummies*), mas difere no sentido de ter como objetivo prever a probabilidade de um evento ocorrer. No caso específico de *cross-selling*, o modelo define a relação entre a probabilidade de um cliente adquirir mais de

um produto ou serviço e um conjunto de fatores ou atributos que o caracterizam. Segundo Diniz e Louzada (2012), esta relação é definida pela transformação *logito*.

A transformação *logito* consiste em aplicar a função log no *odds*. Da razão entre os *odds* resulta o *odds ratio*, ou razão de chances, que é a probabilidade do indivíduo assumir o evento de interesse quando a característica está presente (1), comparado com a ausência da mesma (0) (DINIZ e LOUZADA, 2012). A transformação *logito* é definida pela expressão apresentada na Equação 1.

$$log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\,, \tag{1}$$

em que $\pi(x)$ é definido como

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}, \tag{2}$$

onde $x_1, \ldots, x_k$ são as covariáveis, ou variáveis explicativas, e $\beta_0, \beta_1, \ldots, \beta_k$ são os coeficientes estimados. O efeito das covariáveis sobre a variável resposta é medido através dos seus respectivos coeficientes. Quando positivo indica um aumento na probabilidade prevista de ocorrer o evento, e negativo o efeito é contrário. No estudo de *cross-selling*, isso significa dizer que a variável que possui coeficiente positivo aumenta a probabilidade prevista de o cliente adquirir mais de um produto ou serviço, enquanto a variável que possui coeficiente negativo diminui a probabilidade prevista de que a compra aconteça.

A transformação *logito* possui natureza não linear e, então, os parâmetros do modelo são estimados através do método da máxima verossimilhança, diferente do modelo de Regressão linear, que utiliza o método dos mínimos quadrados. (HOSMER e LEMESHOW, 1989; HAIR *et al.*, 2005). De acordo com Hosmer e Lemeshow (1989), para testar a significância dos coeficientes, uma alternativa é a razão de verossimilhança, e Hair *et al.* (2005) sugerem a estatística de Wald.

Apesar da facilidade na explicação dos resultados e flexibilidade de utilização da técnica de regressão logística, ela está sujeita, assim como os outros modelos de regressão, à multicolinearidade. De acordo com Hair *et al.* (2005), o ideal é que se tenha alta correlação das variáveis explicativas com a variável resposta, mas com baixa

correlação entre elas próprias. Utilizar variáveis explicativas altamente correlacionadas no modelo pode resultar em estimativas errôneas dos coeficientes (HOSMER e LEMESHOW, 1989). Para avaliar a colinearidade, duas medidas são frequentemente utilizadas: o valor de tolerância e o seu inverso, o VIF (fator de inflação de variância), que possui como referência um valor VIF acima de 10 (HAIR *et al.,*2005). Uma alternativa é o uso do método *stepwise* para seleção das variáveis, muito utilizado em regressão linear e encontrado na maioria dos pacotes estatísticos, que consiste basicamente em incluir ou excluir variáveis do modelo de acordo com a sua importância.

2.2.2 *Redes Neurais*

O cérebro humano possui muitas habilidades, dentre elas, a capacidade de relacionar situações novas com experiências passadas e de identificar, entender e interpretar características com grande eficiência. De acordo com Haykin (2001), desejando entender e reproduzir uma máquina que se aproximasse do cérebro humano, Mc Culloch e Pitts, Hebb e Rosenblat iniciaram entre os anos de 1940 e 1950, os primeiros trabalhos sobre redes neurais (RN) que foram originados de estudos de inteligência artificial (IA). Uma RN busca explicar, com base na forma funcional das observações, a relação entre a variável resposta e as variáveis explicativas, possuindo o processo de aprendizagem como um grande diferencial em relação às demais técnicas de *data mining*.

A rede neural se assemelha ao cérebro humano no sentido de que o conhecimento é adquirido através de um processo de aprendizagem e armazenado por pesos sinápticos, onde o elemento fundamental é o neurônio, também chamado de nó. O uso de redes neurais possui benefícios como, por exemplo, a capacidade de lidar com a não linearidade das relações, realizar o processo de treinamento até que não haja mais mudanças significativas nos pesos sinápticos e a habilidade de se adaptar a modificações do ambiente na qual foi treinada, podendo alterar seus pesos sinápticos em tempo real. Os neurônios de uma rede neural são uniformes, o que permite que as mesmas teorias e algoritmos de aprendizagem sejam utilizados em diferentes aplicações (HAYKIN, 2001).

A Figura 1 apresenta o modelo não linear de um neurônio, comumente chamado de *perceptron*. Ele possui três elementos que processam a informação de entrada

gerando as informações de saída: (*i*) um conjunto de sinapses; (*ii*) um somador; e (*iii*) uma função de ativação. Esse processamento é feito multiplicando cada variável de entrada por seu respectivo peso sináptico e então o valor resultante é processado por uma função de ativação, que restringe a amplitude de saída em um intervalo finito, resultando na informação que será a entrada para o nó seguinte. Este modelo também inclui um bias, ou viés, que tem o efeito de aumentar (viés positivo) ou diminuir (viés negativo) os valores de entrada da função de ativação (HAYKIN, 2001).

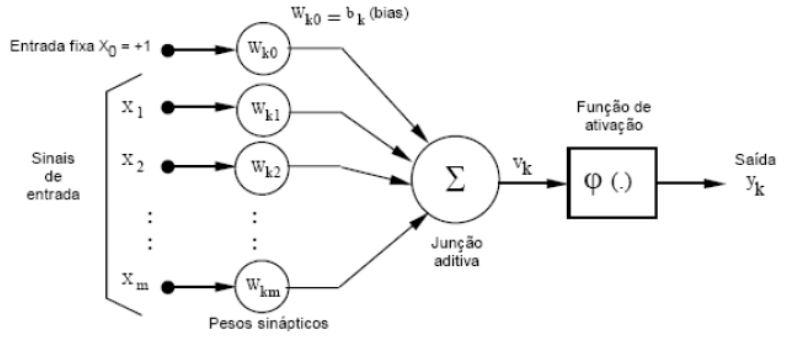


Figura 1. Modelo não linear de um neurônio [Haykin, 2001]

O modelo da Figura 1 pode ser descrito matematicamente através da Equação 3, que possui semelhança com o modelo de regressão não linear múltipla, onde as entradas do neurônio $x_1, \dots, x_m$ são as variáveis explicativas, a saída y_k é a variável resposta, os pesos sinápticos $w_{k1}, \dots, w_{km}$ são os coeficientes de regressão, w_{k0} o intercepto e $\varphi(.)$ é a função de ativação não linear.

$$y_k = \varphi(u_k + b_k) \tag{3}$$

onde $u_k = \sum_{j=1}^{m} w_{kj}\, x_j$ (4)

A primeira camada de uma rede neural é a camada de entrada, as intermediárias são as camadas escondidas e a última é a camada de saída. O número de camadas e a quantidade de neurônios na rede devem ser determinados conforme a natureza do problema. Geralmente, o aumento do número de neurônios na rede é utilizado quando uma característica específica é importante, para assegurar um grau de precisão maior na

tomada de decisão (HAYKIN, 2001). De acordo com Hair *et al.* (2005), as variáveis métricas necessitam de um neurônio para cada variável, enquanto que as não métricas precisam ser codificadas, criando categorias representadas por uma variável binária, que serão representadas, cada uma, por um neurônio. A Figura 2 apresenta uma rede neural na sua forma mais simples, uma rede de camada intermediária única.

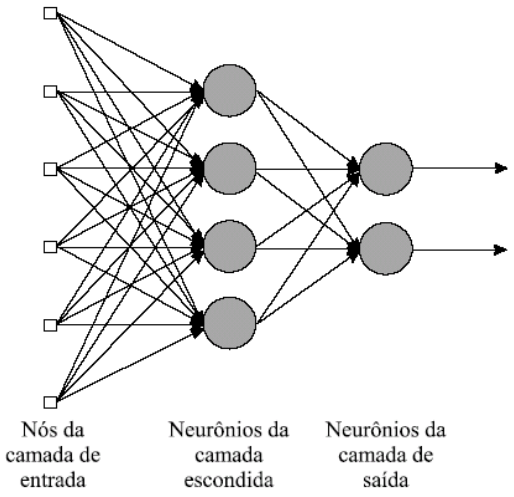


Figura 2. Modelo estrutural de camada única [Haykin, 2001]

Os fatores principais que caracterizam uma RN são: (i) arquitetura da rede: divididas em três classes, em que a primeira delas, redes alimentadas adiante com camada única, é a mais simples e consiste na projeção da camada de entrada sobre a de saída, mas não o inverso; a segunda, redes alimentadas diretamente com múltiplas camadas, se diferencia da primeira pela presença de neurônios na camada escondida e; a terceira, redes recorrentes, se diferencia por permitir a realimentação; (ii) tipo de treinamento: os parâmetros do modelo podem ser estimados de duas formas; pelo treinamento supervisionado, que se caracteriza por ter uma saída para cada vetor de entrada; ou através do não supervisionado, que não possui uma saída para cada entrada. O algoritmo mais utilizado no treinamento supervisionado é o *backpropagation*, que busca encontrar os pesos sinápticos que minimizam o erro e (*iii*) função de ativação: as fundamentais são a função de limiar, função linear por partes e função sigmóide. Dado que a entrada de um neurônio na camada escondida é a combinação linear das saídas anteriores, as funções lineares não seriam adequadas. Portanto, no caso de modelos com neurônios na camada escondida, normalmente se usa a função logística. Já na camada de saída a função varia de acordo com a natureza do problema. (HAYKIN, 2001).

De acordo com Selau (2012), como resultado da combinação não linear dos pesos sinápticos que ocorre na camada escondida, o modelo não informa a influência direta das variáveis explicativas na variável resposta, dificultando na interpretação dessas variáveis. Nesse sentido, Hair *et al.* (2005) sugerem que se utilize a técnica de redes neurais para previsão e classificação quando o interesse está somente na precisão do modelo.

2.3 Métodos de Combinação

2.3.1 *Combinação em série (Hybrid)*

Conhecida na literatura como modelo híbrido, a combinação em série é um método que vem sendo foco de estudos recentes e que tem mostrado resultados promissores, sendo utilizada principalmente com o intuito de minimizar os inconvenientes das técnicas de inteligência artificial e melhorar a classificação, previsão e desempenho dos modelos (LEE *et al.*, 2002; HSIEH, 2005; LEE e CHEN, 2005; CHEN *et al.*, 2009; SELAU, 2012; TSAI e CHEN, 2010). Segundo Tsai e Chen (2010), o modelo híbrido em dois estágios consiste basicamente na combinação em sequência de duas técnicas diferentes de agrupamento ou de classificação. De acordo com Ghodselahi (2011), para desenvolver um modelo híbrido, a primeira técnica é utilizada para orientar o processamento da segunda.

Existem diversas maneiras de se combinar técnicas de modelagem individuais para o desenvolvimento de modelos híbridos. Ao se utilizar a abordagem híbrida em dois estágios, os tipos de métodos de combinação possíveis são: combinação de duas técnicas de classificação; combinação de duas técnicas de agrupamento; uma técnica de agrupamento combinada com uma de classificação; e uma técnica de classificação combinada com uma de agrupamento (TSAI e CHEN, 2010).

Lee *et al.* (2002) combinaram duas técnicas de classificação com o objetivo de melhorar a solução inicial e aumentar a precisão de classificação de um modelo de pontuação de crédito, fazendo uso da análise discriminante para selecionar as variáveis preditoras significativas que são então utilizadas como as variáveis de entrada do modelo de redes neurais, utilizando ainda, o resultado da previsão obtida na análise discriminante como informação extra na camada de entrada da rede neural. Comparando o desempenho dos modelos desenvolvidos, o modelo híbrido convergiu mais rápido e

obteve uma maior acurácia que os modelos obtidos através das duas técnicas de classificação.

Com vista à necessidade de se comparar a eficiência entre as diferentes formas de combinações de modelos híbridos, no estudo pioneiro de Tsai e Chen (2010) foram desenvolvidos os quatro tipos de combinação de modelos híbridos em dois estágios. Como resultado, o melhor método de combinação indicado é entre duas técnicas de classificação, para o qual foram utilizadas a regressão logística e redes neurais, onde a regressão logística seleciona as variáveis significativas que serão utilizadas como nós de entrada na rede neural.

Não existe na literatura um método para selecionar as variáveis de entrada da rede neural (LEE *et al*., 2002). Desta forma, a modelagem híbrida torna-se uma alternativa eficiente, onde essas variáveis são selecionadas através de outra técnica que é utilizada na modelagem em um estágio anterior a rede neural.

2.3.2 *Combinação em paralelo (Ensemble)*

O primeiro estudo a cerca da combinação em paralelo, conhecida como combinação de previsões ou simplesmente *ensemble*, surgiu com Bates e Granger (1969), sendo consagrado em diversos estudos posteriores (CLEMEN, 1989; MAKRIDAKIS e HIBON, 2000; HIBON e EVGENIOU, 2005; WERNER, 2005; CONSTANTINO e PAPALARDO, 2010; MARTINS, 2011) como um método eficaz para se reduzir os erros gerados com a previsão. Segundo Clemen (1989), ao invés de escolher a melhor técnica de previsão a ser utilizada, são definidas, de acordo com o objetivo do estudo, quais técnicas poderiam aumentar a acurácia da previsão, de modo que cada técnica pode contribuir capturando algum tipo de informação intrínseca aos dados.

O método de combinação em paralelo consiste em combinar as decisões de diferentes técnicas de previsão individuais aplicadas a um mesmo conjunto de dados (KITTLER *et al*., 1998). Em *data mining*, esse método é bastante utilizado também na combinação de diferentes algoritmos de aprendizagem, como por exemplo, em redes neurais. De acordo com Haykin (2001), estudos apontam que modelos obtidos a partir da combinação de diferentes redes resultam em melhor desempenho quando comparados aos modelos obtidos através da rede que apresenta melhor desempenho individual.

Segundo Werner e Ribeiro (2006), é preciso saber quais técnicas de previsão utilizar e de que maneira combiná-las. Selau (2012) elenca três etapas para o desenvolvimento de classificadores baseados na combinação em paralelo, onde, primeiramente são escolhidas as técnicas de classificação que geram os modelos individuais, posteriormente os modelos que obtiveram melhor desempenho são então combinados gerando por fim uma única previsão. Dado que o principal objetivo ao se combinar técnicas individuais é aumentar a acurácia da previsão, não há vantagem em combinar técnicas com características de previsão similares. Segundo Polikar (2006), existem diversas maneiras para realizar a combinação, entre elas, os métodos algébricos, como média aritmética ou ponderada, mediana, soma, produto, mínimo e máximo, e também os métodos baseados em votação, como por exemplo, a votação majoritária simples, onde mais de duas técnicas são comparadas quanto a sua decisão. De acordo com Clemen (1989), métodos de combinação simples funcionam razoavelmente bem quando comparados a métodos mais complexos.

3. MÉTODO

A metodologia proposta para este estudo é uma adaptação dos métodos propostos por Berry e Linoff (2004), SAS Institute Inc. (2003) e da sistemática proposta por Selau (2012). É composta por quatro etapas, em que a primeira, chamada pré-processamento, consiste na definição da população alvo, seleção da amostra e na escolha e categorização das candidatas a variáveis explicativas. Na segunda etapa são desenvolvidos os modelos de *cross-selling* através das técnicas de modelagem individuais: regressão logística e redes neurais. Na terceira etapa são aplicados os métodos de combinação em série e em paralelo. A quarta e última etapa tem por objetivo realizar um comparativo entre os modelos desenvolvidos nas duas etapas anteriores. Estas etapas estão ilustradas na Figura 3 e são descritas na sequência.

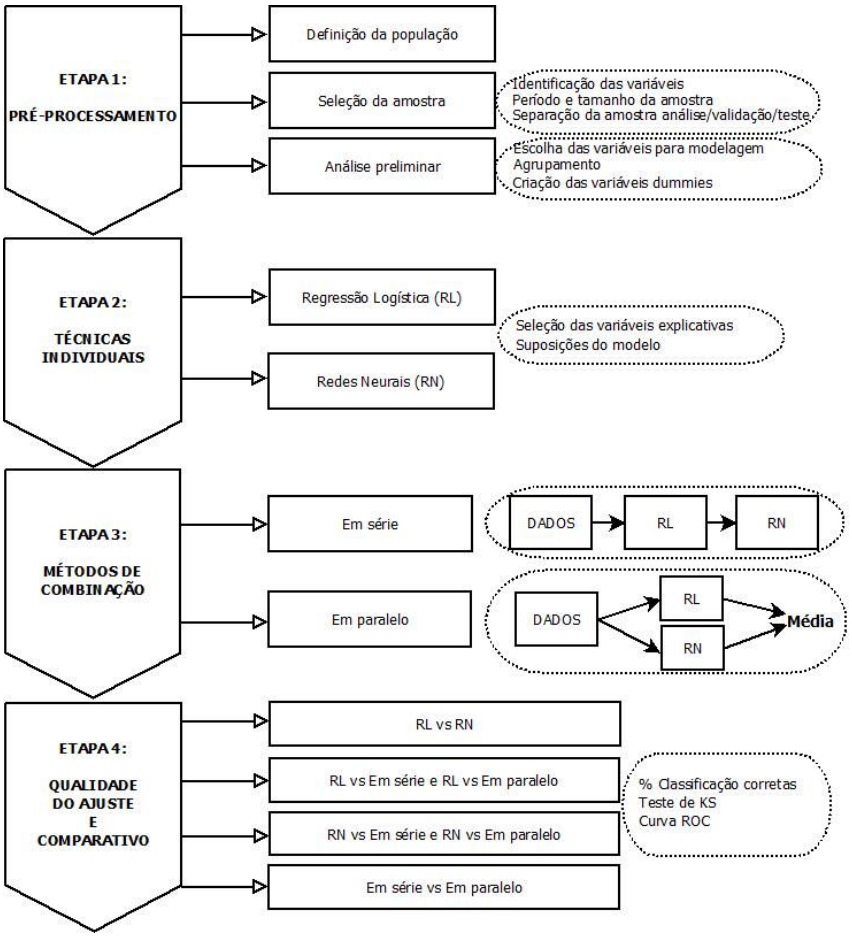


Figura 3. Método proposto

3.1 Pré-processamento

A etapa de pré-processamento engloba desde a definição da população alvo e seleção dos dados, período de tempo e tamanho da amostra até a validação dos dados, agrupamento e seleção das variáveis que serão utilizadas no desenvolvimento dos modelos. Esta etapa é dividida em três passos: *(i)* Definição da população; *(ii)* Seleção da amostra; e *(iii)* Análise preliminar.

3.1.1 *Definição da população*

A escolha da população alvo e dos dados a serem utilizados devem contemplar os objetivos do estudo. Muitas vezes a seleção das variáveis relevantes ao problema de pesquisa e o volume e período de tempo a serem analisados é realizada de acordo com a disponibilidade dos dados. Primeiramente torna-se necessário avaliar o volume e a qualidade dos dados e definir quais variáveis podem descrever o cliente, como por exemplo, variáveis de perfil e demográficas. A escolha dessas variáveis pode ser

baseada na literatura, no conhecimento de especialistas ou de acordo com as variáveis disponíveis no banco de dados da empresa. Considerando que o modelo de *cross-selling* proposto é baseado no comportamento de compra do cliente após adquirir o primeiro produto, ou seja, o produto pelo qual o cliente se direciona até o estabelecimento para efetuar a compra, todos os dados utilizados são oriundos das operações deste produto.

3.1.2 *Seleção da amostra*

A seleção da amostra tem impacto direto na qualidade do modelo final, tanto em relação a sua representatividade perante a população quanto à sua dimensão, visto que amostras maiores tendem a obter resultados melhores. Um ponto crucial a ser avaliado é a consistência e preenchimento dos dados, de modo que dados incorretos devem ser eliminados para que não sejam realizadas análises errôneas e inadequadas. É fundamental então que seja feita uma exploração prévia dos dados, avaliando todos os campos quanto ao seu conteúdo, qualidade de preenchimento, consistência e presença de observações faltantes (*missing*).

Com o intuito de se obter uma amostra mais representativa, o conjunto de dados é dividido de acordo com a natureza da variável de interesse, formando-se então dois grupos distintos: o grupo de indivíduos com resposta positiva (evento de interesse): clientes que adquiriram o segundo produto; e o grupo com resposta negativa: clientes que não adquiriram o segundo produto. Segundo Berry e Linoff (2004), em grande parte das aplicações de *data mining* é observado um desbalanceamento entre o primeiro e o segundo grupo, dado que o número de ocorrências do evento de interesse é consideravelmente menor. Uma alternativa à amostragem aleatória simples, que pode não ser eficiente quando se observa eventos raros (ou até mesmo nem tão raros), é aumentar a proporção do evento menos frequentes de modo a se obter o equilíbrio na amostra e minimizar os problemas de discriminação. Para Berry e Linoff (2004), quando se tem uma variável resposta binária, valores entre 20% e 30% para a categoria mais rara na amostra produzem bons resultados, já Thomas (2002) sugere uma proporção 1:1.

É necessário ainda, que mais de uma amostra seja retirada da mesma população, pois ao se utilizar a mesma amostra para desenvolver e testar o modelo pode-se obter informações equivocadas quanto ao seu desempenho, como por exemplo, concluir que o modelo está bom quando, na verdade, ele se ajusta bem apenas para aquelas observações. O conjunto de dados é usualmente dividido em três amostras

independentes: *(i)* análise (ou treinamento): é a amostra sobre a qual o modelo será construído; *(ii)* validação: é a amostra utilizada para medir a capacidade de generalização do modelo; e *(iii)* teste: é a amostra que servirá para avaliar o desempenho do modelo perante novos dados, considerando o desbalanceamento inicial entre os grupos. De acordo com Berry e Linoff (2004) a partição 60%-30%-10% (análise-validação-teste) tem bom desempenho em casos práticos.

3.1.3 *Análise preliminar*

A análise preliminar consiste em explorar os dados com o objetivo de adquirir algum conhecimento sobre o comportamento das variáveis e qualidade dos dados. Esta etapa contempla tarefas importantes como a escolha das variáveis para entrar na modelagem, o agrupamento dos atributos de variáveis e a criação de variáveis *dummies*. Para escolher quais variáveis irão seguir para o processo de modelagem, é analisada a associação entre cada variável explicativa e a variável resposta através do risco relativo (RR), calculado a partir de tabelas de contingência. O cálculo do RR consiste em dividir o percentual de clientes do grupo 1 (adquiriram o segundo produto) pelo percentual de clientes do grupo 2 (não adquiriram o segundo produto) para cada atributo de cada variável explicativa. Quanto maior a diferença entre os percentuais dos que adquiriram o segundo produto e dos que não adquiriram, maior será a utilidade dessa variável no modelo.

Selau (2012) sugere a escala proposta por Lewis (1992) e Hand e Henley (1997) como método para agrupar os atributos de acordo com o risco que o cliente possui de adquirir o segundo produto após ter adquirido o primeiro, definida como: péssimo – RR<0,50; muito mau – RR entre 0,50 e 0,67; mau – RR entre 0,67 e 0,90; neutro – RR entre 0,90 e 1,10; bom – RR entre 1,10 e 1,50; muito bom – RR entre 1,50 e 2,00; e excelente – RR maior que 2,00, de modo que esse risco seja homogêneo dentro de cada categoria da variável e heterogêneo entre elas. Após esse agrupamento são então criadas as variáveis *dummies*, que assumem os valores 1 (o cliente está neste grupo de risco) ou 0 (o cliente não está nesse grupo de risco).

3.2 Técnicas de Modelagem Individuais

Para construção dos modelos individuais, é necessário primeiramente se avaliar as suposições das técnicas de regressão logística e redes neurais, escolher as variáveis

explicativas e, por fim, avaliar o seu ajuste aos dados. Como resultado, é obtido um *score* para cada cliente que classifica a sua propensão a adquirir o segundo produto após adquirir o primeiro.

3.2.1 *Regressão Logística*

A técnica de regressão logística possui o pressuposto de ausência de multicolinearidade. Uma forma de contornar este pressuposto é utilizar o método *stepwise*, que seleciona automaticamente a combinação de variáveis que melhor explicam o modelo. Esse método é geralmente utilizado porque, na presença de duas ou mais variáveis explicativas altamente correlacionadas, o método seleciona dentre elas a que possui maior influência na variável resposta e as demais não entram no modelo.

3.2.2 *Redes Neurais*

A técnica de redes neurais não possui pressupostos a serem analisados, pois o modelo é gerado através do processo de aprendizagem. Pelo mesmo motivo, não existe um número de nós, de camadas escondidas ou um algoritmo fixo para que sejam aplicados a qualquer conjunto de dados. Desta maneira, é necessário realizar diversos arranjos possíveis destas características e das variáveis em estudo para avaliar qual o modelo que obtém melhor desempenho.

3.3 Métodos de Combinação

Para aplicar os métodos de combinação em série e em paralelo serão utilizados os modelos a serem construídos conforme descrito na Etapa 2. Ao final, são avaliados os ajustes destes métodos ao conjunto de dados.

3.3.1 *Combinação em série (Hybrid)*

No método de combinação em série serão informados ao nó de entrada da rede neural apenas as variáveis explicativas que foram significativas no modelo obtido através da técnica de regressão logística. Esta é a única diferença no processo em relação ao modelo de redes neurais tradicional, uma vez que, após a entrada dos dados a modelagem segue da mesma forma.

3.3.2 *Combinação em paralelo (Ensemble)*

No método de combinação em paralelo, é combinado, para cada cliente, o *score* que determina a sua propensão a adquirir o segundo produto, obtido pelo modelo de regressão logística e pelo modelo de redes neurais, originando um novo *score* para o cliente. Essa combinação é realizada através da média aritmética. Outras formas de combinação estão disponíveis, mas não serão abordadas neste estudo.

3.4 Qualidade do ajuste e comparativo

Diversas medidas estatísticas são utilizadas para testar o desempenho e medir a qualidade dos modelos com o objetivo de comparar os métodos empregados e escolher o melhor entre eles. As medidas abordadas neste trabalhão são: *(i)* percentual de classificação corretas; *(ii)* índice Kolmogorov-Smirnov (KS); e a *(iii)* curva ROC.

O percentual de acerto do modelo nas classificações dos clientes propensos ou não a adquirir o segundo produto é avaliado pela divisão da quantidade de clientes corretamente classificados pelo total de clientes da amostra.

O teste não paramétrico de Kolmogorov-Smirnov (KS) para 2 amostras independentes tem por objetivo determinar se as duas amostras provém da mesma população. Consiste basicamente em obter a máxima diferença entre as distribuições acumuladas das duas amostras. De acordo com Picinini *et. al.* (2003) uma diferença entre as distribuições acumuladas maiores que 30% e taxas de acerto superior a 65% são consideradas satisfatórias na classificação dos dois grupos.

A curva ROC (*Receiver Operating Characteristic*) é uma representação gráfica da sensibilidade versus especificidade e tem por objetivo primário mensurar a capacidade do modelo em reduzir os erros tipo I e tipo II (SILVA, 2009). Para este estudo, entende-se por sensibilidade a probabilidade de identificar os clientes propensos a adquirir o segundo produto, já especificidade é a probabilidade de identificar os clientes não propensos, sendo que o valor de corte para estes valores é a probabilidade a partir da qual o especialista considera o cliente propenso ou não. No entanto, segundo Kisahleitner (2008), a medida que é analisada de fato é a área sob a Curva ROC (AUROC – *Area Under ROC*), também conhecida por discriminação. Uma discriminação perfeita vale 1, e uma discriminação de 0,5 significa que o modelo não agrega valor, pois estimular os clientes aleatoriamente teria o mesmo resultado.

4. RESULTADOS

Os resultados da aplicação do método proposto em uma base de dados real são expostos nesta seção e seguem a mesma ordenação da metodologia descrita na seção 3. O *software* utilizado para as análises é o *SAS Enterprise Miner* versão 4.3.

4.1 Pré-processamento

4.1.1 *Definição da população*

A população de interesse são os clientes de uma instituição financeira brasileira que, após contratarem o produto crédito pessoal (CP), podem contratar também o produto seguro de vida. Com o intuito de classificar os clientes da instituição quanto à sua propensão a adquirir o segundo produto, a base de dados contempla todos os clientes que adquiriram o crédito pessoal entre 01 de julho de 2011 e 31 de julho de 2012, totalizando 61.365 clientes. A base é composta por todas as variáveis de perfil e demográficas dos clientes disponíveis na base de dados da empresa (preenchidas na proposta em que é feita a solicitação de crédito pessoal) e também informações referentes à venda do primeiro produto; estas variáveis estão descritas na Tabela 1. A variável resposta é dicotômica e define se o cliente adquiriu ou não o seguro de vida depois de ter contratado o crédito pessoal.

Tabela 1. Variáveis disponíveis para o modelo

Variável	Descrição
Sexo	Feminino ou masculino
Idade	Idade, em anos, no dia da contratação do CP
Estado Civil	Casado, solteiro, divorciado, viúvo e outros
Renda	Valor da renda (R$)
Órgão de trabalho	Órgão em que o cliente trabalha
Grupo do órgão	Critérios estabelecidos pela instituição
CEP Residencial	CEP do local de residência do cliente
Banco	Banco que o cliente é correntista
Origem da venda	Venda interna ou externa
Pessoa captação	Vendedor pessoa física ou jurídica
Valor Liberado CP	Valor liberado no CP (R$)

4.1.2 *Seleção da amostra*

Dados inconsistentes, incorretos ou faltantes representaram menos de 0,01% da base total e, portanto, foram excluídos da análise. Os clientes foram divididos aleatoriamente em amostras de análise, validação e teste, na proporção de 60%, 30% e 10%, respectivamente. Considerando o desbalanceamento inicial entre os grupos, em que 30% dos clientes adquiriram o seguro de vida (evento de interesse) e 70% não adquiriram, optou-se então, com o intuito de se obter o equilíbrio na amostra e minimizar os problemas de discriminação, dividir os clientes aleatoriamente na proporção 1:1, tanto na amostra de análise como na amostra de validação. Para a amostra de teste, a verdadeira proporção foi mantida. A amostra de análise é então constituída por 27.352 clientes, a amostra de validação por 6.838 e a amostra de teste por 6.136 clientes, resultado em uma proporção de 68%, 17% e 15%, respectivamente. Esta proporção entre as amostras ficou diferente do proposto inicialmente, pois nas amostras de análise e validação foram desconsiderados grande parte dos clientes do segundo grupo (não compradores do segundo produto) para se alcançar o equilíbrio entre os dois grupos.

4.1.3 *Análise Preliminar*

As 11 variáveis selecionadas na seção 4.1.1 como candidatas a compor o modelo foram analisadas individualmente e a variável "estado civil" foi excluída da análise. Isso porque, conclui-se que esta informação não é preenchida corretamente no momento da proposta. No caso da regressão logística em particular, faz-se necessário a criação de variáveis *dummies* para as variáveis qualitativas, para que estas sejam melhor interpretáveis na equação resultante. As variáveis *dummies* foram criadas para as 10 variáveis candidatas restantes, sendo que, para aquelas variáveis que apresentam um grande número de atributos, as *dummies* foram criadas com base nas categorias de risco relativo apresentadas na seção 3.1.3, de modo que, para cada variável nominal (órgão de trabalho, grupo de órgão, CEP e banco) se obteve sete agrupamentos de atributos de acordo com o risco de adquirir o segundo produto. Para a variável CEP foram consideradas as 2 primeiras posições para fazer o agrupamento de regiões com risco relativo próximos, nos casos em que o agrupamento das 3 primeiras posições foi significativo, este foi utilizado. Também com base no risco relativo, foram criadas classes para as variáveis de natureza numérica (idade, renda e valor liberado) de modo a

se evitar problemas decorrentes da não linearidade. No total, foram geradas 58 variáveis *dummies* que serão as variáveis explicativas para a construção do modelo desenvolvido através da regressão logística. A descrição dessas variáveis se encontra no Apêndice.

4.2 Técnicas de Modelagem Individuais

4.2.1 *Regressão Logística*

Para a construção do modelo logístico utilizou-se o método *stepwise*, com níveis de significância para a entrada e saída de variáveis de 5%. Com o uso deste método garantiu-se o atendimento da suposição de ausência de multicolinearidade entre as variáveis explicativas. Para que algumas variáveis tivessem significância foi necessário agrupar *dummies* com risco próximo, como por exemplo, agrupar uma com risco péssimo e outra com risco muito mau. Ao final, 19 variáveis *dummies* foram significativas para compor o modelo final. A equação que retorna a probabilidade de um cliente vir a comprar o seguro de vida (segundo produto) após a compra do produto crédito pessoal (primeiro produto) é apresentada na Equação 5, a legenda com a descrição das variáveis pode ser encontrada no Apêndice.

$$P(Y=1) = \frac{1}{1 + \exp(-1{,}1586 - 0{,}5632\ \text{DGCEPR12} + 0{,}2610\ \text{DGCEPR56} + 0{,}4516\ \text{DGCEPRE2} + 0{,}1055\ \text{DGCEPRE4} + 0{,}6773\ \text{DGCEPRE7} + 0{,}1590\ \text{DGGORGAO6} - 0{,}0531\ \text{DGORGAO2} + 0{,}1610\ \text{DGORGAO7} - 0{,}2996\ \text{DIDAD1} + 0{,}0893\ \text{DIDAD10} - 0{,}1178\ \text{DIDAD3} + 0{,}0442\ \text{DIDAD7} + 0{,}0538\ \text{DIDAD8} + 0{,}1136\ \text{DPSCAPTF} + 0{,}0322\ \text{DSEXOF} - 0{,}7072\ \text{DVENDE} + 0{,}0964\ \text{DVLIB4} + 0{,}0725\ \text{DVLIB7} + 0{,}0932\ \text{DVLIB8})} \quad (5)$$

Para a interpretação do modelo, é necessário se observar o sinal dos coeficientes de cada uma das variáveis; coeficientes positivos indicam que clientes com aquela característica têm maior probabilidade de adquirir o segundo produto, enquanto coeficientes negativos representam diminuição na probabilidade de adquirir o segundo produto. Observando a Equação 5, pode-se concluir, por exemplo, que o cliente que possui residência no grupo de CEP de risco excelente (DGCEPRE7) ou que trabalhe em um dos órgãos do grupo de risco excelente (DGORGAO7) tem um aumento na probabilidade de adquirir o segundo produto. Já um cliente que contratou o primeiro produto com a venda externa (DVENDE) ou que trabalha em um dos órgãos do grupo

de risco muito mau (DGORGAO2) tem uma diminuição na probabilidade de adquirir o segundo produto.

4.2.2 *Redes Neurais*

O modelo neural foi construído através do treinamento supervisionado utilizando o algoritmo *backpropagation* e a função de ativação logística. Foram desenvolvidos diversos modelos variando a quantidade de neurônios (de 1 a 58 neurônios) na camada escondida. A partir dos modelos neurais desenvolvidos, se optou pelo modelo com 31 neurônios na camada escondida, pois este possui o maior percentual de classificações corretas, maior valor de KS a de área abaixo da curva ROC. A Tabela 2 mostra os resultados das três melhores redes construídas.

Tabela 2. Melhores redes construídas para o modelo neural – amostra de teste

N de neurônios	% Classificação	KS	ROC
18	72,39%	0,4850	0,8000
31	72,80%	0,4860	0,8000
32	71,91%	0,4840	0,7960

4.3 Métodos de Combinação

4.3.1 *Combinação em série (Hybrid)*

O modelo de combinação em série foi construído utilizando o resultado da regressão logística como entrada na rede neural. As 19 variáveis *dummies* que foram significativas para compor o modelo logístico descrito na seção 4.2.1 foram informadas ao modelo neural como variáveis explicativas. As demais variáveis não são informadas com o intuito de minimizar o tempo de aprendizado da rede e facilitar a interpretação dos resultados. Analogamente ao processo realizado para a construção do modelo neural na seção 4.2.2, o modelo de combinação em série foi construído através do treinamento supervisionado utilizando o algoritmo *backpropagation* e a função de ativação logística. A partir dos modelos neurais desenvolvidos (de 1 a 58 neurônios na camada escondida), se optou pelo modelo com 30 neurônios na camada escondida, por possuir o maior percentual de classificações corretas, maior valor de KS e de área abaixo da curva ROC, dentre as redes testadas. A Tabela 3 mostra os resultados dos três melhores modelos construídos.

Tabela 3. Melhores redes construídas para a combinação em série – amostra de teste

N de neurônios	% Classificação	KS	ROC
29	72,83%	0,4870	0,8030
30	72,85%	0,4870	0,8030
33	72,67%	0,4850	0,8000

4.3.2 *Combinação em paralelo (Ensemble)*

A combinação em paralelo foi construída utilizando o *score* obtido através do modelo logístico construído em *4.2.1* e o *score* obtido através do modelo neural construído em *4.2.2*. Primeiramente são obtidos os *scores* através das técnicas de modelagem individuais para cada cliente e então estes valores são combinados por meio de média aritmética. Assim, um cliente que possui, por exemplo, um *score* de 0,80 via regressão logística e 0,86 via redes neurais, através da combinação em paralelo esse cliente passa a ter um *score* de 0,83.

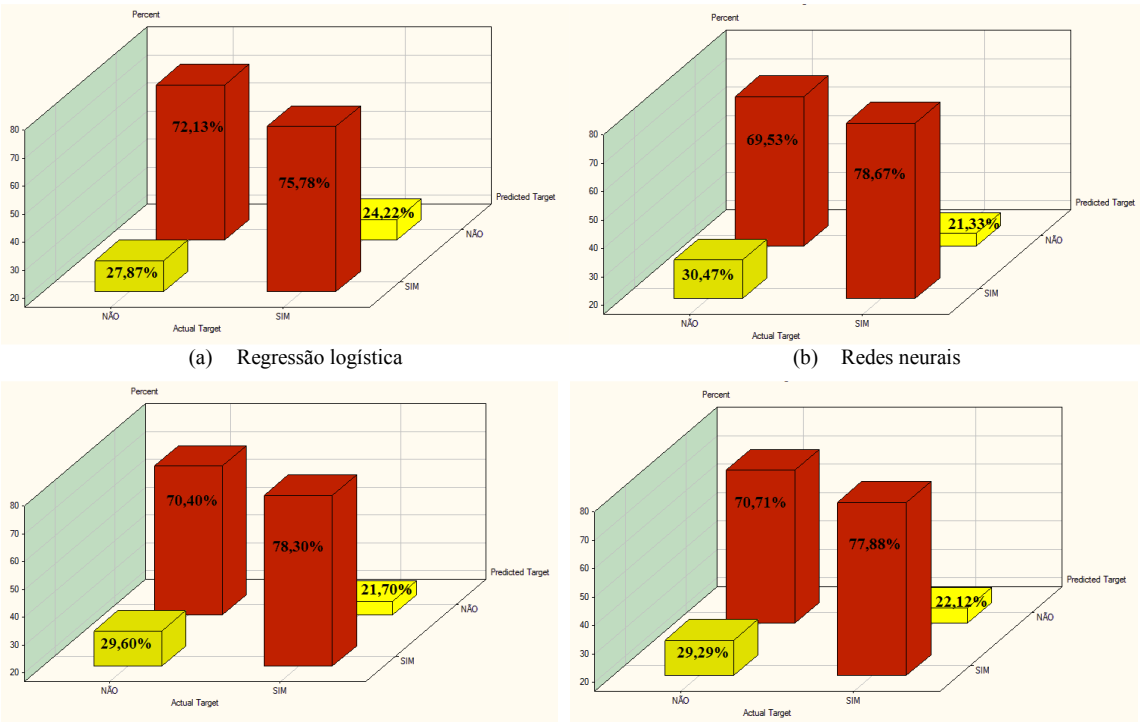
4.4 Qualidade do ajuste e comparativo

A avaliação de cada um dos modelos construídos se dá mediante o emprego das medidas já expostas na seção 3.4; são elas: percentual de classificação corretas, teste de Kolmogorov-Smirnov (KS) e curva ROC. Essas medidas são avaliadas nas três amostras (análise, validação e teste) e são expostas na Tabela 4, com o intuito de realizar um comparativo entre as técnicas e métodos utilizados no desenvolvimento do modelo de *cross-selling*.

Tabela 4. Medidas de desempenho para as três amostras

Amostra	Medida de Desempenho	RL	RN	Em série	Em paralelo
Análise	% Classificações corretas	75,61%	75,08%	75,07%	75,03%
	KS	0,4860	0,5033	0,5029	0,5023
	ROC	0,8030	0,8091	0,8116	0,8078
Validação	% Classificações corretas	74,14%	74,82%	74,92%	74,80%
	KS	0,4829	0,4975	0,4993	0,4970
	ROC	0,8029	0,8069	0,8086	0,8090
Teste	% Classificações corretas	73,26%	72,80%	72,85%	72,93%
	KS	0,4790	0,4861	0,4873	0,4882
	ROC	0,7965	0,7999	0,8033	0,8054

A primeira das medidas apresentada na Tabela 4 mostra o percentual de classificações corretas para cada uma das amostras, sendo que a técnica ou método que melhor classifica os clientes na amostra de teste é o modelo logístico, com 73,26% de classificações corretas. Tanto na amostra de análise quanto na amostra de validação e de teste, os percentuais de acerto total encontrados para todas as técnicas e métodos são superiores a 65%, indicando que todos possuem boa capacidade de classificação. Os percentuais de classificações corretas para cada técnica/método em cada modelo na amostra de teste podem ser encontradas na Figura 4.



(a) Regressão logística (b) Redes neurais

(c) Combinação em série (d) Combinação em paralelo

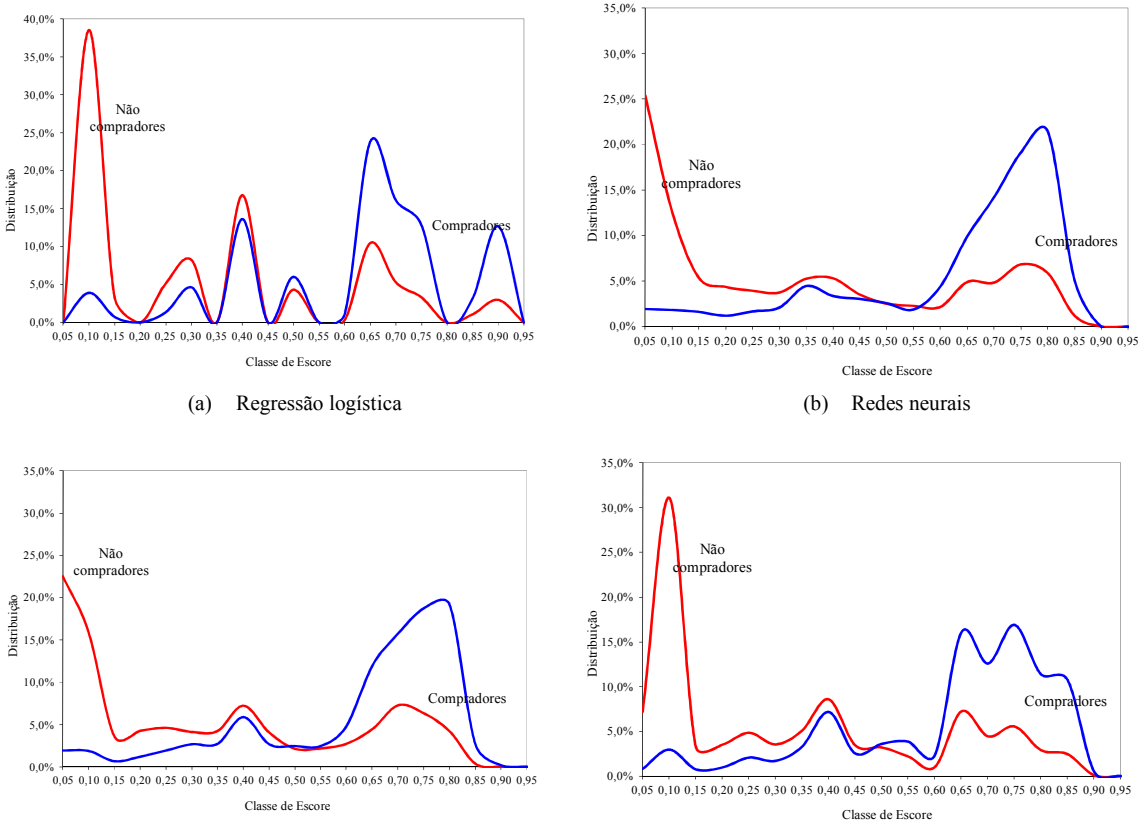
Figura 4. Percentual de classificação corretas

A segunda medida observada é o valor do teste de KS, que tem por objetivo determinar se duas amostras provêm de uma mesma população. Espera-se que as duas amostras (clientes propensos e clientes não propensos a adquirir o segundo produto) sejam oriundas de populações distintas, pois assim o modelo é eficiente em separar os dois grupos de clientes. Os quatro modelos construídos apresentaram diferença entre as distribuições acumuladas dos clientes propensos e dos não propensos maior de 0,30 de modo que estes podem ser considerados eficientes. O maior valor de KS para a amostra de teste é o obtido através do método de combinação em paralelo, o que significa dizer

que este método alcançou os maiores níveis de diferença entre as distribuições acumuladas dos clientes propensos e dos não propensos.

A terceira e última medida a ser analisada é a área abaixo da curva ROC, que mostra uma boa capacidade de discriminação para todos os modelos, indicando que a capacidade de identificar corretamente os propensos (sensibilidade), assim como a capacidade de identificar os não propensos (especificidade) está bem ajustada, sendo que o método de combinação em paralelo apresentou um valor ligeiramente maior.

Uma avaliação dos modelos construídos, na amostra de teste, é apresentada na Figura 5 com a distribuição dos dois grupos de clientes: compradores e não compradores do segundo produto. Analisando o comportamento das curvas de distribuição dos clientes compradores e não compradores, verifica-se que os quatro modelos conseguem separar os dois grupos de clientes, já que é possível observar a tendência de que os clientes não compradores se concentram à esquerda da escala e os compradores à direita.



(a) Regressão logística

(b) Redes neurais

(c) Combinação em série

(d) Combinação em paralelo

Figura 5. Distribuição de compradores e não compradores por classe de *score*

A Tabela 5 apresenta uma comparação entre os modelos construídos e a escolha, dentre estes, daquele que melhor se ajustou aos dados, segundo os resultados das medidas apresentadas para verificar a qualidade de ajuste.

Tabela 5. Comparativo dos modelos construídos

Técnica/Método	Melhor ajuste
Regressão logística *vs* Redes neurais	Redes neurais
Regressão logística *vs* Em série	Em série
Redes neurais *vs* Em série	Em série
Regressão logística *vs* Em paralelo	Em paralelo
Redes neurais *vs* Em paralelo	Em paralelo
Em série *vs* Em paralelo	Em paralelo

Ao analisar as três medidas empregadas para comparação das técnicas e métodos abordados, observa-se, na comparação entre as técnicas de modelagens individuais, que os resultados obtidos com o modelo neural foram ligeiramente superiores aos obtidos com a regressão logística. Ao analisar os dois métodos de combinação percebe-se que o método de combinação em paralelo obteve resultados ligeiramente superiores em relação ao método de combinação em série. As comparações entre as técnicas individuais e os métodos de combinação indicam que os dois métodos de combinação são ligeiramente superiores às duas técnicas de modelagem individuais. Cabe ressaltar ainda, que os métodos de combinação possuem um objetivo em comum, que é o de melhorar a acuracidade da previsão, porém, o método de combinação em série possui a vantagem da facilidade na interpretação das variáveis, o que não acontece no método de combinação em paralelo.

5. CONSIDERAÇÕES FINAIS

Este artigo desenvolveu um método para resolução de um problema real de *cross-selling*, buscando identificar que clientes, após terem adquirido um primeiro produto (crédito pessoal), têm maior probabilidade de adquirir também um segundo produto (seguro de vida). Para tanto, foram apresentadas e comparadas duas das técnicas mais utilizadas em *data mining* para previsão e classificação de clientes, regressão logística e redes neurais, e dois métodos de combinação dessas técnicas, combinação em série e em paralelo, com o intuito de se chegar ao modelo que melhor

classifica os clientes quanto à sua propensão a adquirir um segundo produto. Foram apresentados todos os passos necessários para a obtenção dos modelos utilizando as técnicas e métodos abordados, detalhando desde a obtenção das variáveis e das amostras, a modelagem e aplicação das técnicas e também dos métodos de combinação para se chegar ao melhor modelo. Esse trabalho traz uma abordagem bastante inovadora principalmente ao que se refere aos modelos de *cross-selling* e às comparações das técnicas individuais de modelagem com os dois métodos de combinação; técnicas e métodos estes que podem ser utilizados nas mais diversas aplicações de *data mining*.

Os dois objetivos principais foram alcançados: resolver um problema real de *cross-selling* de uma instituição financeira e contribuir na pesquisa de técnicas de *data mining*. A efetividade desta proposta para a empresa se dá pelo fato de que a utilização do modelo de *cross-selling* elimina a subjetividade da análise tradicional, aproveitando as informações expostas para direcionar as campanhas de marketing. Uma vez que foi possível identificar o perfil dos clientes mais propensos a adquirir o produto, atribuindo a estes clientes um *score* que determina a sua propensão a adquiri-lo, esta análise se faz de grande utilidade também para possíveis aprimoramentos do produto. Além disso, a padronização do procedimento de decisão e o direcionamento correto das campanhas diminuem os gastos e aumentam o retorno das mesmas, aumentando a rentabilidade da empresa e possibilitando uma maior eficiência no atendimento aos clientes.

Os métodos de combinação apresentaram desempenho ligeiramente superior às técnicas individuais, sendo que o método de combinação em paralelo obteve maior ajuste aos dados em relação ao método de combinação em série, e, consequentemente, aos modelos individuais. Diferenças entre os modelos logístico e neural e os métodos de combinação em série e em paralelo empregados parecem pouco relevantes, pois produzem praticamente o mesmo resultado. Apesar disso, cabe ressaltar ainda que, algumas operações possuem um nível de criticidade alto para a empresa e, mesmo diferenças singelas de acuracidade das previsões podem resultar em ganhos ou perdas significantes para a empresa. A escolha do melhor modelo a ser utilizado deve ir ao encontro das necessidades da empresa, dado que o melhor resultado encontrado em termos de ajuste se deu com a aplicação do método de combinação em paralelo, porém, este não possui grandes diferenças quando comparado ao método de combinação em série, método este que possui a vantagem da facilidade na interpretação das variáveis, o que não acontece no método de combinação em paralelo.

No decorrer do desenvolvimento deste artigo surgiram algumas questões que não foram abordadas, mas que foram consideradas importantes como sugestões para trabalhos futuros: (i) fazer uso de algoritmos genéticos para encontrar os parâmetros ótimos das redes neurais (quantidade ótima de neurônios da camada oculta e quantidade de camadas ocultas), pois o método utilizado de tentativa e erro é limitado, no sentido que os parâmetros encontrados podem não ser os melhores, resultando em um poder de classificação menor do que poderia se obter ao otimizar a rede; (ii) utilizar outras formas de combinar os *scores* no método de combinação em paralelo, como por exemplo, a média ponderada, onde se atribui um peso maior para aquele modelo que obteve um melhor desempenho individual; (iii) combinar redes com parâmetros diferentes, como por exemplo, diferentes algoritmos de aprendizagem, número de neurônios e quantidade de camadas ocultas; e (iv) avaliar o quanto representa de ganho financeiro pequenas diferenças de acurária entre modelos.

REFERÊNCIAS BIBLIOGRÁFICAS

ADORNO, C. F.; BUENO, J. F. *Modelos de Propensão: Oferta de Crédito Pessoal,* 2011.

BATES, J. M.; GRANGER, C. W. J. The combination of forecasts. *Operational Research Quarterly*. v. 20, n. 4. p. 451-468, 1969.

BERRY, M. J. A; LINOFF, G. S. *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*. 2.ed. New York: John Wiley & Sons, 2004.

CLEMEN, R. T.; Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*. v. 5, p. 559-583, 1989.

COSTANTINE, C.; PAPPALARDO, C. A Hierarchical procedure for combination of forecasts. *International Journal of Forecasting*. v. 26, p. 725-743, 2010.

DINIZ, C.; LOUZADA, F. Modelagem Estatística para Risco de Crédito. In: Minicurso no XX SINAPE – Simpósio Nacional de Probabilidade e Estatística, João Pessoa-PB, 2012.

DYCHE, J. *The CRM handbook: a business guide to customer relationship management*. Reading, MA: Addison-Wesley, 2001.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*. v. 17 n. 3, 1996.

GHODSELAHI, A. A Hybrid support vector machine ensemble model for credit scoring. *International Journal of Computer Applications*, v.17, n.5, 2011.

MORAES, L. G. *Uma abordagem alternativa de behavioral scoring usando modelagem híbrida de dois estágios com regressão logística e redes neurais*. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

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

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: concepts and techniques*. 3.ed. San Francisco: Morgan Kaufmann, 2012.

HAYKIN, S. *redes neurais: princípios e prática*. Trad. Paulo Martins Engel. 2.ed. Porto Alegre: Bookman, 2001.

HIBON, M.; EVGENIOU, T. To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting*. v. 21, p. 15-24, 2005.

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

HSIEH, N. C.; HUNG, L. P. A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications*, v.37, n.1, p. 534-545, 2010.

KISAHLEITNER, M. *Análise de Técnicas de Data Mining na aquisição de clientes de cartão de crédito não correntistas*. 93f. Dissertação (Mestrado em Administração) – Fundação Getúlio Vargas, São Paulo, 2008.

KITTLER, J.; HATEF, M.; DUIN, R. P. W.; MATAS, J. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v.20, n.3, p.226-239, 1998.

KNOTT, A.; HAYES, A.; NESLIN, S. A. Next-product-to-buy models for cross-selling applications. *Journal of Interactive Marketing*, v. 16, 2002.

LEE, T.; CHIU, C.; LU, C.; CHEN, I. Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, v.23, n.3, p.245-254, 2002.

LEE, T.S.; CHEN, I. F. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, v.28, p.743-752, 2005.

MAKRIDAKIS, S. G.; HIBON, M. The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, v. 16, p. 451-476, 2000.

MARTINS, V. L.M *Comparação de Combinação de Previsões correlacionadas e não correlacionadas com as suas previsões individuais: um estudo com séries industriais*. 100f. Dissertação (Engenharia de Produção) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2011.

NISBET, R.; ELDER, J.; MINER, G. *Handbook of Statistical Analysis and Data Mining Applications*.Amsterdam, NL: Academic Press, 2009.

OLSON, D. L.; DELEN, D. *Advanced Data Mining Technique.* New York: Springer, 2008.

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*, Bauru, SP, 2003.

POLIKAR, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, v.6, n.3, p.21-45, 2006.

SAS Institute Inc. *Data Mining Using SAS® Enterprise Miner™: A Case Study Approach, Second Edition.*Cary, NC: SAS Institute Inc, 2003.

SAS Institute Inc. *Getting Started with SAS® Enterprise Miner™ 4.3*.Cary, NC: SAS Institute Inc, 2004.

SELAU, L. P. R. *Modelagem para Concessão de Crédito a pessoas físicas em empresas comerciais: da decisão binária para a decisão monetária*. 111f. Tese (Doutorado em Administração) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

SILVA, V. F. *Modelos de Propensão ao Consumo baseados em Redes Neuronais Artificiais, o caso particular do Crédito Pessoal*. 105f. Dissertação de Mestrado (Estatística e Gestão da Informação) – Universidade Nova de Lisboa, Lisboa, 2009.

THOMAS, L. C.; EDELMAN, D. B.; CROOK, J. N. Credit Scoring and Its Applications, Philadelphia: SIAM.

TSAI, C. F.; CHEN, M. L. Credit rating by hybrid machine learning techniques. *Applied Soft Computing*, v.10, p.374-380, 2010.

WERNER, L. *Um Modelo Composto para Realizar Previsão de Demanda Através da Integração da Combinação e de Previsões e Ajuste Baseado na Opinião*. 166f. Tese de Doutorado (Engenharia de Produção) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2005.

WERNER, L.; RIBEIRO, J. L. D. Modelo composto para prever demanda através da integração de previsões. *Produção*, v.16, n.3, p.493-509, 2006.

APÊNDICE – VARIÁVEIS DUMMIES CRIADAS

VARIÁVEL	CÓDIGO	DESCRIÇÃO
SEXO	DSEXOF	DUMMY SEXO FEMININO
	DSEXOM	DUMMY SEXO MASCULINO
IDADE	DIDAD1	DUMMY IDADE ATÉ X ANOS
	DIDAD2	DUMMY IDADE X1 a Y1 ANOS
	DIDAD3	DUMMY IDADE X2 a Y2 ANOS
	DIDAD4	DUMMY IDADE X3 a Y3 ANOS
	DIDAD5	DUMMY IDADE X4 a Y4 ANOS
	DIDAD6	DUMMY IDADE X5 a Y5 ANOS
	DIDAD7	DUMMY IDADE X6 a Y6 ANOS
	DIDAD8	DUMMY IDADE X7 a Y7 ANOS
	DIDAD9	DUMMY IDADE X8 a Y8 ANOS
	DIDAD10	DUMMY IDADE X9 a Y9 ANOS
	DIDAD11	DUMMY IDADE X10 a Y10 ANOS
	DIDAD12	DUMMY IDADE ACIMA DE X11 ANOS
RENDA	DRENDA1	DUMMY RENDA MENOR QUE Y
	DRENDA2	DUMMY RENDA X1 a Y1
	DRENDA3	DUMMY RENDA X2 a Y2
	DRENDA4	DUMMY RENDA X3 a Y3
	DRENDA5	DUMMY RENDA X4 a Y4
	DRENDA6	DUMMY RENDA X5 a Y5
	DRENDA7	DUMMY RENDA X6 OU +
ÓRGÃO DE TRABALHO	DGORGAO1	DUMMY ÓRGÃO PÉSSIMO
	DGORGAO2	DUMMY ÓRGÃO MUITO MAU
	DGORGAO3	DUMMY ÓRGÃO MAU
	DGORGAO4	DUMMY ÓRGÃO NEUTRO
	DGORGAO5	DUMMY ÓRGÃO BOM
	DGORGAO6	DUMMY ÓRGÃO MUITO BOM
	DGORGAO7	DUMMY ÓRGÃO EXCELENTE
GRUPO DO ÓRGÃO	DGGORGAO1	DUMMY GRUPO DE ÓRGÃO PÉSSIMO
	DGGORGAO2	DUMMY GRUPO DE ÓRGÃO MUITO MAU
	DGGORGAO3	DUMMY GRUPO DE ÓRGÃO MAU
	DGGORGAO4	DUMMY GRUPO DE ÓRGÃO NEUTRO
	DGGORGAO6	DUMMY GRUPO DE ÓRGÃO MUITO BOM
	DGGORGAO7	DUMMY GRUPO DE ÓRGÃO EXCELENTE
CEP RESIDENCIAL	DGCEPRE1	DUMMY CEP RESIDENCIAL PÉSSIMO
	DGCEPRE2	DUMMY CEP RESIDENCIAL MUITO MAU
	DGCEPRE3	DUMMY CEP RESIDENCIAL MAU
	DGCEPRE4	DUMMY CEP RESIDENCIAL NEUTRO
	DGCEPRE5	DUMMY CEP RESIDENCIAL BOM
	DGCEPRE6	DUMMY CEP RESIDENCIAL MUITO BOM
	DGCEPRE7	DUMMY CEP RESIDENCIAL EXCELENTE
BANCO	DGBANCO3	DUMMY BANCO MAU
	DGBANCO4	DUMMY BANCO NEUTRO
ORIGEM DA VENDA	DVENDE	DUMMY ORIGEM DE VENDA EXTERNA
	DVENDI	DUMMY ORIGEM DE VENDA INTERNA
PESSOA CAPTAÇÃO	DPSCAPTF	DUMMY PESSOA DE CAPTAÇÃO FÍSICA
	DPSCAPTJ	DUMMY PESSOA DE CAPTAÇÃO JURÍDICA
VALOR LIBERADO CP	DVLIB1	DUMMY VALOR MENOR QUE Y
	DVLIB2	DUMMY VALOR X1 a Y1
	DVLIB3	DUMMY VALOR X2 a Y2
	DVLIB4	DUMMY VALOR X3 a Y3
	DVLIB5	DUMMY VALOR X4 a Y4
	DVLIB6	DUMMY VALOR X5 a Y5
	DVLIB7	DUMMY VALOR X6 a Y6
	DVLIB8	DUMMY VALOR X7 OU +
CEP RESIDENCIAL - agrupamento de classes	DGCEPR12	DUMMY CEP RESIDENCIAL PÉSS+MM
	DGCEPR56	DUMMY CEP RESIDENCIAL BOM+MB
	DGCEPR67	DUMMY CEP RESIDENCIAL MB+EXCEL