RESUMO

UMA ANÁLISE EM *DATA MINING*: ÁRVORES DE DECISÃO, REDES NEURAIS E *SUPPORT VECTOR MACHINES*

Autoria: Luiz Carlos Jacob Perera, Herbert Kimura, Rui Américo Mathiasi Horta, Fabiano Guasti Lima

Este trabalho se alinha às exigências do *Basel Committee on Banking Supervision* (BCBS) na revisão do *Basel Capital Accord* e do *Basel II* os quais, ao vincularem a administração de risco ao capital das instituições bancárias, contribuíram de forma fundamental para o desenvolvimento dos modelos *Internal Risk Based* (IRB) mostrando o que seriam boas práticas de administração de risco. Lembra ainda a crise financeira iniciada em 2007 e que se estendeu até 2010, e que também incitou a reação do BCBS (2010) culminando na divulgação do chamado *Basel III*, o qual entre suas medidas mais severas prevê o aumento do capital mínimo mais buffer de conservação já em 2013 para 8% chegando a 10,5% em 2019. Estas medidas austeras aumentam a importância dos IRB e consequentemente da modelagem de crédito e outros elementos que influenciam o risco das instituições financeiras. No Brasil o Banco Central de Brasil (BCB, 2011) está em fase de consulta pública sobre a aplicação dos IRB para cálculo do Patrimônio de Referência Exigido (PRE) que, em outras palavras, delimitará a alavancagem dos bancos. Neste contexto os procedimentos de *data mining* ganham relevância e o mercado tem se debruçado sobre o aperfeiçoamento de modelos que possam ser incluídos em procedimentos de IRB, que, se aprovados pelo elemento regulador, podem ajudar as instituições financeiras a reduzirem seus PRE, permitindo-lhes uma alavancagem financeira adequada a sua capacidade ótima de administrar riscos. Neste trabalho buscou-se comparar três metodologias distintas de *data mining* com a característica comum de serem não paramétricas. O processo de árvores de decisão foi explorado com a metodologia desenvolvida por Leo Breimam et al. (1984), conhecida como CART. Para redes neurais Utilizou-se o software Statistica @ 6.1 sendo testadas as redes lineares de propagação em múltiplas camadas (*multiplayer propragation* ou MLP), rede neural polinomial (*polynomial neural network* ou PNN) e duas redes com função radial (*radial basis function* ou RBF). Já a técnica de *data mining* aplicada para extração de padrões foi a da classificação, processada através de máquinas de vetor suporte. No tratamento dos dados foi utilizado o software de *data mining* WEKA 3.5.6 (Witten, et al., 2011). Uma característica relevante deste trabalho foi o tratamento amostral. A amostra constou de uma base de dados voltada para o crédito ao consumidor de um grande magazine, com um ano de acompanhamento e cerca de 20.070 processos de crédito. Esta amostra foi dividida aleatoriamente em duas subamostras de 10.035 observações cada uma serviu para treinamento e a outra para validação. O percentual de acertos dos três processos aplicados oscilaram entre 68,08% e 76,77% na operação de treinamento e 67,70% e 74,65% na validação. O procedimento através de máquinas de vetor suporte mostrou-se superior na análise. Deve ficar claro que os procedimentos adotados com uma base de dados voltada para crédito ao consumidor, podem ser aplicados em outros contextos, como crédito empresarial e bancário. O contexto de crédito é bastante amplo e muitas empresas comerciais e de varejo já admitem a atividade financeira como parte relevante de seu negócio.

UMA ANÁLISE EM *DATA MINING*: ÁRVORES DE DECISÃO, REDES NEURAIS E *SUPPORT VECTOR MACHINES*

1. INTRODUÇÃO

Lando (2004) analisando o desenvolvimento da pesquisa em modelagem de crédito atribui relevante papel à indústria financeira. Lembra a importante participação do *Basel Committtee on Banking Supervision* (BCBS) na revisão do *Basel Capital Accord* e no *Basel II* os quais, ao vincularem a administração de risco ao capital das instituições bancárias, contribuíram de forma fundamental para o desenvolvimento dos modelos *Internal Risk Based* (IRB) mostrando o que seriam as boas práticas de administração de risco.

Complementando o pensamento de Lando (2004) torna-se importante lembrar que a crise financeira iniciada em 2007 e que se estendeu até 2010, também incitou a reação do BCBS (2010) culminando na divulgação do chamado *Basel III*, o qual entre suas medidas mais severas prevê o aumento do capital mínimo mais buffer de conservação já em 2013 para 8% chegando a 10,5% em 2019.

Estas medidas austeras aumentam a importância dos IRB e consequentemente da modelagem de crédito e outros elementos que influenciam o risco das instituições financeiras. No Brasil o Banco Central de Brasil (BCB, 2011) está em fase de consulta pública sobre a aplicação dos IRB para cálculo do Patrimônio de Referência Exigido (PRE) que, em outras palavras, delimitará a alavancagem dos bancos.

Neste contexto os procedimentos de *data mining* ganham relevância e o mercado tem se debruçado sobre o aperfeiçoamento de modelos que possam ser incluídos em procedimentos de IRB, que, se aprovados pelo elemento regulador, podem ajudar as instituições financeiras a reduzirem seus PRE, permitindo-lhes uma alavancagem financeira adequada a sua capacidade ótima de administrar riscos.

Com foco em crédito ao consumidor, neste trabalho buscou-se comparar três metodologias distintas de *data mining* com a característica comum de serem não paramétricas. O processo de árvores de decisão foi explorado com a metodologia desenvolvida por Leo Breimam et al. (1984), conhecida como CART. Para redes neurais Utilizou-se o software Statistica @ 6.1 sendo testadas as redes lineares de propagação em múltiplas camadas (*multiplayer propragation* ou MLP), rede neural polinomial (*polynomial neural network* ou PNN) e duas redes com função radial (*radial basis function* ou RBF). Já a técnica de *data mining* aplicada para extração de padrões foi a da classificação, processada através de máquinas de vetor suporte. No tratamento dos dados foi utilizado o software de *data mining* WEKA 3.5.6 (Witten, et al., 2011).

Uma característica relevante deste trabalho foi o tratamento amostral A amostra constou de uma base de dados e um grande magazine, com um ano de acompanhamento, com cerca de 20.070 processos de crédito. Esta amostra foi dividida aleatoriamente em duas subamostras de 10.035 observações, uma das quais serviu para treinamento e a outra para validação.

O fato de ter sido usada uma base de dados de um grande magazine, quando mais adequado seria uma base voltada para crédito bancário, deve-se à grande dificuldade de acesso a dados reais para aplicação em pesquisas. Hand e Henley (1997) discutem o preocupante aspecto de

que os modelos desenvolvidos e testados no meio acadêmico estejam limitados pelo acesso aos bancos de dados das empresas que, por alegadas questões de confiabilidade, geralmente só tornam disponível bases defasadas ou incompletas. Concomitantemente, as atividades creditícias das indústrias não parecem refletir aquilo que os trabalhos acadêmicos reputam como melhores práticas.

No entanto, deve ficar claro que os procedimentos adotados com uma base e dados voltada para crédito ao consumidor, pode ser aplicado em outros contextos, como o crédito empresarial e bancário. O contexto de crédito é bastante amplo e muitas empresas comerciais e de varejo já admitem a atividade financeira como parte relevante de seu negócio.

O trabalho está estruturado da seguinte forma: introdução que remete ao contexto motivador o trabalho; referencial teórico que fundamenta as variáveis utilizadas e os processos de análise; descrição da metodologia empregada; análise dos principais resultados e considerações finais.

2. REFERENCIAL TEÓRICO

2.1. MODELANDO O COMPORTAMENTO DO CONSUMIDOR

Milton Friedman (1957) desenvolveu a Teoria de Renda Permanente (TRP) na qual afirma que o padrão de gastos não é determinado pela renda corrente do consumidor e sim pela expectativa que tem com relação à sua renda permanente. Em outras palavras, a renda do indivíduo é parcialmente consumida na sua sobrevivência e seus excedentes irão para uma poupança, cujo investimento irá compor seu patrimônio e organizar sua renda futura, ou sua renda permanente. Esta renda possibilitará ao indivíduo manter um padrão de consumo. Uma das principais conclusões desta teoria é que mudanças de curto-prazo na renda terão pouco efeito na expansão do consumo.

Ainda de acordo com Friedman (1957), o elemento chave de consumo é a riqueza real do indivíduo e não sua renda corrente disponível. A renda permanente do indivíduo é determinada por seus ativos físicos (ações, títulos de renda, propriedades) e intangíveis (educação, experiência, etc.). A capacidade do consumidor em gerar renda, possibilita que ele antecipe a duração da renda, e desta forma organize um fluxo compatível de despesas. Consumidores de baixa renda terão uma propensão elevada de consumo e baixa capacidade de poupança. Consumidores de rendas mais elevadas são influenciados pela transitoriedade da sua renda e sua propensão marginal de consumo é abaixo da média.

Os principais princípios da TRP continuam em evidência e resistindo ao tempo. Carrol (2001) afirma que, apesar de esta ter sido bastante contestada nos anos setenta e oitenta, devido a ter sido apresentada de forma descritiva e empírica – sem o rigor de uma demonstração matemática – os princípios básicos da TRP resistem de forma robusta e o grande desafio dos matemáticos hoje seria quantificar em seus modelos a relação entre poupar e consumir.

O crédito é o combustível do consumo, em outras palavras, sem a energia do crédito o consumo sofreria variações que implicariam o fluxo produtivo e o próprio sistema econômico. Daí a necessidade de ser entendido o comportamento do consumidor num contexto de crédito. Attanasio (1999) e Bertola, Disney e Grant (2006), sendo que estes últimos deram uma roupagem atual ao trabalho do primeiro, apresentam um modelo econômico moderno do comportamento do consumidor, o qual será apresentado de forma breve para justificar a escolha das principais variáveis da modelagem proposta na seção de metodologia.

De acordo com a TRP, a diferença entre renda e consumo (portanto poupar ou tomar emprestado) é determinada pela expectativa em relação à incerteza da renda futura, Isto é, as famílias escolhem seu nível de consumo em cada período, sujeito a uma restrição orçamentária intertemporal, de forma que possam controlar sua volatilidade – mantendo um consumo relativamente estável em nível de despesas.

Bertola, Disney e Grant (2006) explicam que o problema das famílias é escolher o consumo c em cada período maximizando sua utilidade sujeita a uma restrição orçamentária intertemporal. O fluxo de consumo é escolhido de forma a maximizar sua vida útil, a soma de um fluxo descontado de períodos de funções utilidade u ( ) na forma:

$$max\ E_t \sum_{j=0}^{T} \beta^j u\left(c_{t+j}\right), \tag{1}$$

onde T é o horizonte de planejamento individual (pode ser infinito), E_t denota as expectativas das famílias condicionadas às informações disponíveis em t, e $\beta = 1/(1+\delta)$ é o fator de desconto das famílias, sendo δ a taxa subjetiva de desconto.

A maximização de (1) é sujeita a

$$A_{t+1} = (1 + r_{t+1})(A_t + y_t - c_t) \tag{2}$$

onde A^1 é o nível de ativos (ou passivos), renda do trabalho no tempo t é denotada y_t, e a taxa de juros determinada no mercado de crédito é a mesma r_t para ativos ou passivos. Esta é uma equação de evolução do ativo, considerando que o valor do ativo em qualquer período deve ser igual ao ativo do período anterior mais a renda (renda do trabalho mais rendimento dos ativos) menos o consumo naquele período.

A solução ótima deste problema satisfaz a equação de Euler da seguinte forma

$$u'(c_t) = E_t\, u'\,(c_{t+1})[(1 + r_{t+1})/(1 + \delta)] \tag{3}$$

na qual a utilidade marginal é uma função decrescente do consumo, se a flutuação do consumo é bem-estar decrescente. Dessa forma a otimização implica que a função marginal no tempo 't+1' é exclusivamente determinada pela satisfação e pela taxa de juros, e é não correlacionada com qualquer coisa que seja previsível (e não afeta a satisfação) no tempo 't' ou anteriormene, tal como a renda corrente e passada.

Se a utilidade marginal é (aproximadamente) linear em consumo, o crescimento do consumo depende das magnitudes relativas de r e de δ, mas mudanças no consumo de período para período são independentes das mudanças previsíveis na renda, as quais são suavizadas pelo acesso ao mercado de crédito. A linearidade da utilidade marginal torna possível combinar a condição ótima e a restrição orçamentária intertemporal para obter uma relação equilibrada entre poupança, renda e consumo e entre poupanças e a evolução da renda no tempo, conforme equações (4) e (5),

$$S_t = \frac{rA_t}{1 + r_t} + y_t - c_t \tag{4}$$

$$S_t = -\sum_{j=0}^{\infty} (1 + r)^{-j}\, E_t\left(y_{t+j} - y_{t+j-1}\right) \tag{5}$$

Quando a expectativa é de que a renda futura aumente, a poupança tende a se tornar negativa: as famílias irão antecipar despesas se desfazendo de seus ativos, ou tomando empréstimos,

caso não haja ativos disponíveis. Isto ocorre, por exemplo, quando o chefe da família perde o emprego, mas espera encontrar outro melhor rapidamente. Ao contrário, o chefe da família irá poupar se prevê um decréscimo de renda no futuro, por exemplo, no caso de se aposentar.

Emprestar e tomar emprestado torna possível equilibrar a despesa ao longo dos ciclos da vida, economizando quando a renda é alta para redistribuir nos períodos em que ela se torna mais baixa. Os salários tipicamente possuem a forma de corcova: baixos no início da vida profissional e também na maturidade, quando total ou parcialmente as pessoas se retiram do mercado de trabalho. Desta forma, o modelo prevê que os empréstimos serão mais elevados nas famílias jovens, e que na meia idade as famílias estarão poupando para a sua aposentadoria. Além disto, espera-se que a renda das famílias com maior nível educacional cresça mais rápido do que as de trabalhadores braçais, logo os primeiros tomarão empréstimos mais elevados quando jovens (BERTOLA, DISNEY E GRANT, 2006).

2.2. RECURSIVE PARTITIONING ALGORITHM (RPA)

Novak e LaDue (1999) apresentam o RPA como uma técnica computadorizada, baseada num método não-paramétrico de classificação, que não impõe a adoção de nenhuma distribuição de probabilidades *a priori.* A essência do RPA é desenvolver uma árvore de classificação particionando as observações baseada em divisões binárias de variáveis características. O processo de seleção e partição ocorre repetidas vezes até que mais nenhuma seleção ou divisão das variáveis características seja possível, ou o processo seja interrompido por algum critério pré-determinado. Finalmente, às observações dos nós terminais da árvore de classificação, são atribuídos a grupos classificatórios.

De uma forma resumida os elementos necessários para o crescimento de uma árvore, na visão de Breiman et al. (1984) são: (i) um conjunto de questões binárias aplicadas ao vetor de variáveis $\boldsymbol{x}$; (ii) um critério de partição dos vetores que possa ser aplicado a qualquer nó; (iii) uma regra de parada para os nós terminais; (iv) uma regra para indicar cada nó terminal para uma determinada classe.

Breiman et al. (1984) assume que a construção de um classificador é baseada numa amostra de aprendizado com n elementos, que consiste de uma matriz X de dados (n por $k+1$) sendo ($x_{1,1}$, $x_{2,1}$, ... $x_{k,1}$, $x_{C,1}$) o primeiro elemento e ($x_{1,n}$, $x_{2,n}$, ..., $x_{k,n}$, $x_{C,n}$) o último elemento. Os vetores x_k são vetores de medidas das variáveis do hiperplano e o vetor x_C o vetor de classificação. Considernado um sistema de classificação de crédito ao consumidor, os grupos a serem classificados podem ser designados como uma variável classificatória binária com dois estados: *1 ou 0* e poderiam ser *paga / não paga* ou *bom pagador / mau pagador*. Neste caso estaria sendo considerada uma classificação binária, porém poderiam ser mais de dois grupos.

No processo de classificação, deve-se considerar a população representada pela amostra $\boldsymbol{N}$ com k variáveis características mais uma variável classificatória. Todos os n elementos da amostra N estão contidos no nó Pai, o qual irá constituir a primeira sub-árvore T_1 referida como a classificação inicial da árvore. Todas as observações da amostra original N são alocadas para os grupos *1* ou *0* de acordo com uma determinada regra. A alocação dos nós resultantes para um dos grupos *1* ou *0* depende também das probabilidades estabelecidas *a priori* e do custo das classificações erradas.

O RPA, como explicado anteriormente, tem a forma de uma árvore binária de classificação que assinala os objetos em grupos selecionados a *priori*. As observações dos nós terminais da

árvore de classificação, são atribuídas a grupos classificatórios de modo a minimizar o custo observado de classificação errônea. Considerando um nó terminal t, o risco de atribuir o nó t ao grupo i e ao grupo j é dado, respectivamente, por:

$$R_i = C_{ji} \cdot \pi_j \cdot n_j / N_j \quad (4)$$

Assim, por exemplo, C_{12} pode representar o custo da classificação errada da empresa falida, como não falida, C_{21} pode representar o custo da classificação errada da empresa não falida como falida, π_1 e π_2 são as probabilidades a priori de a empresa falir ou não falir; n_1 e n_2 são os valores das empresas falidas e não falidas num nó terminal; N_1 e N_2 são os valores globais das empresas falidas e não falidas constantes da árvore.

O RPA tem sido aplicado, de forma crescente, devido à sua facilidade de aplicação e recursos disponíveis. Por exemplo, recentemente Min e Jeong (2009) utilizando o RPA em uma pesquisa sobre predição e falência. Ahn et al. (2011) utilizaram o RPA num sistema sofisticado de seleção de clientes em telefonia móvel. Finalmentme, Li, Sun e Wu (2010) também em pesquisa sobre falência de empresas compararam o RPA com outros procedimentos paramétricos e não paramétricos: análise discriminante multivariada (ADM), regressão logística (Logit), uuporte de vetor de máquina (SVM), k Vizinho mais Próximo (kNN).

2.3. REDES NEURAIS

De acordo com Haykin (2001), as redes neurais artificiais representam modelos de processamento paralelos distribuídos, formados por unidades simples de ajustes, que possibilitam capturar um determinado conhecimento ou relacionamento complexo experimental, tornando-o disponível para algumas aplicações como, por exemplo, previsão de novas características ou situações. Desta forma, redes neurais artificiais têm especial aplicabilidade na análise de crédito, pois possibilita, a partir de dados reais ou experimentais de atuais tomadores, a realização de estimativas sobre capacidade de pagamento de clientes propsectivos.

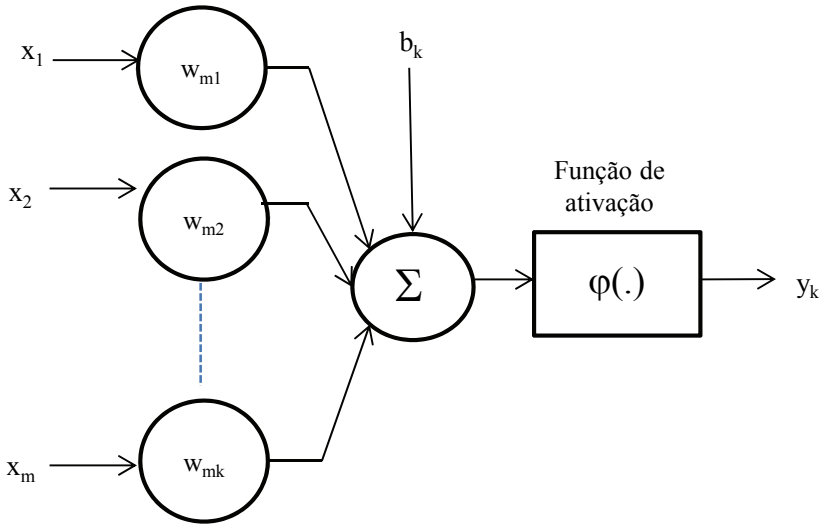

Figura 1 : Modelo de rede neural
Fonte: HAYKIN (2001, p. 36)

Nas redes neurais, o conhecimento experimental é capturado através de um processo de aprendizagem e armazenado em neurônios que, por sua vez, podem ser conectados (Haykin, 2001) a outras camadas de neurônios, permitindo que relacionamentos complexos possam ser

aprendidos. As redes neurais associam pesos sinápticos às conexões entre neurônios. Estes pesos são alterados, através de algoritmos de aprendizagem, à medida que novas informações ou novas observações são incorporadas na rede. O modelo de rede neural reflete os pesos sinápticos que permitem, após ajustes decorrentes de uma função de ativação, o estabelecimento de uma variável de saída ou de resposta a partir de dados de entrada, conforme Figura 1.

No caso de aplicações em análise de crédito, os dados de entrada podem corresponder a características do cliente e o dado de saída ou resultado do modelo pode ser a característica de bom ou mau pagador. Cada observação que entra na rede tem um valor de variável de entrada, por exemplo, características do tomador, que é multiplicada por pesos sinápticos que, por sua vez, pode passar por outras camadas de neurônios. O processamento conjunto dos dados de entrada nos neurônios induz um sinal que passa por uma função de ativação e gera a variável de saída, por exemplo, um valor que indica a capacidade de pagamento do tomador.

Considerando-se que x_1, x_2, ..., x_m são os sinais de entrada da rede e que w_{1k}, w_{2k}, ..., w_{mk} são os pesos sinápticos do k-ésimo neurônio que associam as entradas ao neurônio, a saída y_k é dada por:

$$y_k = \varphi(u_k + b_k)$$

onde

$u_k = \sum_{j=1}^{m} w_{jk} x_j$ e b_k é um viés que pode ser introduzido à função de integração das variáveis de entrada

Na amostra de aprendizagem, os valores das variáveis de entrada x_i e da variável de saída y_i são conhecidas a priori e permitem que a rede neural calibre pesos sinápticos nas interfaces com camadas de neurônios, usando uma função φ que incorpora uma soma ponderada dos sinais de entrada acrescida de um viés. Desta forma, segundo Zhang, Patuwo e Hu (1998), a predição em uma rede neural envolve a ponderação realizada nos neurônios de entrada e em camadas intermediárias de neurônios, caso existam.

2.4. ABORDAGEM *WRAPPER* USANDO *SUPPORT VECTOR MACHINES*

Em problemas de análise de concessão de crédito, há um cadastro inicial que permite o levantamento de um conjunto de variáveis ou atributos de um potencial cliente. No entanto, apenas parte desse conjunto de variáveis pode ser relevante na explicação do potencial de pagamento do cliente. Desta forma, a seleção de variáveis para a análise constitui um importante passo para a definição de um modelo de crédito.

Segundo Piramuthu (2006), a seleção de atributos em mecanismos de *data mining* para análise de crédito possibilitam: (i) o estabelecimento de modelos compactos, (ii) o refinamento do modelo de classificação ou de predição e (iii) a identificação das variáveis relevantes. Algoritmos para escolha de atributos podem envolver pelo menos dois objetivos principais: (i) busca do sub-conjunto de atributos e (ii) avaliação dos sub-conjuntos encontrados (LIU E MOTODA, 1998). Assim, após a identificação de um sub-conjunto de atributos relevantes, avalia-se sua adequação para classificação usando-se algum critério de adequação baseado, por exemplo, em distância, dependência ou precisão. De acordo com Kohavi e John (1997), o processo de interação entre definição de sub-conjuntos de atributos relevantes e avaliação da adequação dos sub-conjuntos pode ser subdividida, basicamente, em duas abordagens principais: (i) filtro, no qual a escolha dos sub-conjuntos ocorre previmente ao algoritmo de aprendizagem e (ii) *wrapper*, no qual, através de um processo iterativo,

baseado em um algoritmo externo de aprendizagem, analisa-se um sub-conjunto de atributos potencialmente relevantes.

Conforme Freitas (1998), a abordagem filtro leva em consideração um sub-conjunto de atributos ou carcaterísticas que mantém a informação pertinente de todo o conjunto de atributos. A partir de uma amostra de treinamento, a abordagem *wrapper* avalia sub-conjuntos candidatos de atributos que sirvam para subsidiar técnicas tradicionais de regressão ou classificação como, por exemplo, regressão linear, regressão robusta, regressão logística, redes neurais, *suport vector machines*, etc, para avaliar a adequação do sub-conjunto de atributos do estudo.

Em particular, pode-se usar, como mecanismo de aprendizagem, técnicas de máquinas de vetor de suporte *Support Vector Machines* (SVM), baseadas em avanços em teoria de aprendizagem estatística (VAPNIK, 1998). Assim com diversas aplicações computacionais e estatística, o SVM constitui um método não-linear que possibilita a resolução de porblemas de classificação e regressão. Segundo Haykin (2001), para o caso de problemas de classificação, a idéia que fundamenta as máquinas de vetor suporte é a construção de um hiperplano como superfície de decisão na qual a margem de separação entre observações de dois grupos difernetes seja máxima.

A máquina de vetor suporte baseia-se no método de minimização estrutural de risco que, utilizando um mecanismo indutivo, considera uma taxa de erro de uma máquina de aprendizagem sobre dados de teste limitada pelor da soma da taxa de erro de treinamento e de um termo que depende da dimensão de Vapnik-Chervonenkis (Haykin, 2001). No caso específico de padrões separáveis, nos quais encaixa-se, por exemplo, a análise de bons ou maus pagadores, uma máquina de vetor suporte busca zerar o primeiro termo da soma descrita anteriormente e minimizar o segundo termo (HAYKIN, 2001, p. 349).

Tendo em vista sua flexibilidade, as SVMs têm se popularizado. Dente suas principais características podem ser citadas (TAN *et al.,* 2006): (i) o problema de aprendizagem da SVMs pode ter uma formulação cuja solução é baseada em otimização convexa, para a qual existem diversos algoritmos efeicientes e rápidos para a obtenção de m~inimos globais, (ii) as SVMs possibilitam um controle de capacidade, através da maximização da margem do limite de decisão, (iii) as SVMs podem ser utiizados na análise de dados categóricos. Assim, Ravi *et al.* (2008) sugerem que as SVMs combina vantagens de métodos estatísticos com maior embasamento teórico com mecanismos de aprendizagem de máquina, voltados à análise de dados, sem necessidade de definição de distribuições específicas e, com isso, comumente robustos.

3. METODOLOGIA

A generalização da performance de um método de aprendizagem relaciona-se com a sua capacidade preditiva, verificada através de testes com dados independentes. Na prática, a avaliação da performance guia a escolha do método de aprendizagem e, em última análise, fornece a medida adequada para escolha do modelo a ser utilizado (HASTIE, TIBSHIRANI e FRIEDMAN, 2009, p.219). Pode-se afirmar que a avaliação da performance vai depender da adequação da amostra que está sendo utilizada e do método escolhido para sua avaliação. Na sequência discute-se sucintamente estes dois importantes temas.

Greene (2008) relata a necessidade de utilização de métodos automatizados para analisar um grande volume de processos de concessões de cartões de crédito, que libera cartões para um número massivo de usuários sem condições de fazer análises individuais. O método de *credit scoring* é utilizado tendo um pequeno número de variáveis características ou atributos como elementos de decisão.

Os modelos acompanham a farta literatura e os meios computacionais disponíveis, mas os principais problemas encontram-se na amostragem. Inicialmente, os dados de default e despesas usados para desenvolver os modelos de análise ficam restritos à seleção da amostra. Por exemplo, os modelos são desenvolvidos com dados passados, e são utilizados para analisar comportamentos futuros, que podem não estar representados anteriormente. Segundo, a amostra usada para analisar a decisão de aprovação é sistematicamente diferente da população da qual foi retirada. Exemplificando: a amostra analisada já passou por uma triagem inicial, ou seja, foi expurgada daqueles cujo demanda por crédito foi rejeitada, logo a população não é a mesma. Esta falta de representatividade dos dados é remediada através do uso de alguma forma de correção desses dados (GREENE, 2008, p. 15).

Neste trabalho, a amostra constou de operações de concessão de crédito, ocorridas durante um determinado ano, de uma grande loja de varejo com matriz o Estado de São Paulo. Esta amostra foi expurgada dos *missing values* e dividida de forma aleatória em duas amostras com 10.035 observações cada uma. A despeito de vieses de seleção decorrentes da exclusão de observações com dados faltantes, considera-se que a amostra resultante possibilita analisar as características que permitem diferenciar bons de maus pagadores. A amostra foi dividida em duas sub-amostras: (i) uma das sub-amostras foi usada para o desenvolvimento do modelo e a outra para o teste o modelo. A sub-amostra de desenvolvimento também é denominada de amostra de aprendizagem (learing) ou calibragem e a sub-amostra de teste (test) também é denominada de amostra de validação. A Tabela 1 mostra a composição das duas amostras:

Tabela 1 Composição das amostras utilizadas na análise

Amostras	Bons Pagadores	Maus Pagadores	Soma
Learning (modelo)	9.803	232	10.035
Test (validação)	9.832	203	10.035
Total da amostra	19.635	435	20.070

Com relação às variáveis explicativas do estudo, são consideradas (i) data de venda (DATA_VENDA), (ii) valor da prestação (PRESTACAO_VALOR), (iii) quantidade de prestações (PRESTACOES_QTD) que variou de 1 a 24 meses, (iv) valor financiado (VALOR_FINANCIADO), (v) estado civil (ESTADO_CIVIL), cuja classificação é solteiro/1, casado/2, viúvo/3, divorciado/4 e outros/5, (vi) gênero (SEXO) do cliente, cuja classificação é masculino/0 e feminino/1, (vii) rendimento (RENDIMENTO_CLIENTE), (viii) data de nascimento (NASCIMENTO_DATA) para análise da influência da idade na característica de bom ou mau pagador, (ix) o código de endereçamento postal do indivíudo (CEP_RESIDENCIA). A variável dependente STATUS reflete uma variável classificatória, com dois estados, mau pagador/0 e bom pagador/1. Na definição deste trabalho, um indivíduo é classificado como mau pagador bastando atrasar um pagamento de prestação.

4. ANÁLISE DE RESULTADOS

A Tabela 2 apresenta ainda as principais estatísticas descritivas das variáveis referentes aos clientes/consumidores, permitindo comparar as duas amostras utilizadas no trabalho: treinamento (learning) e validação (test). Deve-se levar em consideração que as estatísticas foram calculadas para todas as variáveis, independentemente de seu significado lóico, como é o caso de média e desvio padrão da variável CEP, ou mesmo STATUS.

Tabela 2 Estatísticas descritivas: Learning x Test

	Variáveis	N	Mínimo	Máximo	Soma	Média	Variância	Assimetria		Curtose	
			Estatística	Estatística	Estatística	Estatística	Estatística	Estatística	Erro Pd.	Estatística	Erro Pd.
AMOSTRA LEARNING	DATA_VENDA	10035	36.893,0	37.256,0	371.979.322	37.068,19	11.779	0,175	0,024	-1,319	0,049
	PRESTACAO_VALOR	10035	0,6	2.168,0	566.559	56,46	10.376	9,476	0,024	130,136	0,049
	PRESTACAO_QTD	10035	1,0	24,0	67.880	6,76	18	0,613	0,024	-0,240	0,049
	VALOR_FINANCIADO	10035	1,9	3.840,0	2.443.688	243,52	59.836	2,779	0,024	16,307	0,049
	ESTADO_CIVIL	10035	1,0	5,0	15.992	1,59	1	2,053	0,024	4,287	0,049
	SEXO	10035	0,0	1,0	3.970	0,40	0	0,427	0,024	-1,818	0,049
	RENDIMENTO_CLIENTE	10035	0,0	6.500,0	8.815.075	878,43	644.785	2,679	0,024	8,954	0,049
	NASCIMENTO_DATA	10035	307,0	31.249,0	228.470.473	22.767,36	26.153.900	-0,751	0,024	0,356	0,049
	CEP_RESIDENCIA	10035	1,0	98,0	343.263	34,21	775	1,035	0,024	-0,576	0,049
	STATUS	10035	0,0	1,0	9.803	0,98	0	-6,347	0,024	38,298	0,049
AMOSTRA TEST	DATA_VENDA	10035	36.893,0	37.256,0	371.997.201	37.069,98	11.876	0,141	0,024	-1,329	0,049
	PRESTACAO_VALOR	10035	0,4	4.690,0	579.048	57,70	13.506	15,447	0,024	422,656	0,049
	PRESTACAO_QTD	10035	1,0	24,0	68.409	6,82	18	0,583	0,024	-0,308	0,049
	VALOR_FINANCIADO	10035	3,3	4.690,0	2.553.091	254,42	69.436	3,014	0,024	20,707	0,049
	ESTADO_CIVIL	10035	1,0	5,0	15.912	1,59	1	2,112	0,024	4,721	0,049
	SEXO	10035	0,0	1,0	3.943	0,39	0	0,439	0,024	-1,808	0,049
	RENDIMENTO_CLIENTE	10035	100,0	5.000,0	8.730.208	869,98	614.637	2,676	0,024	8,811	0,049
	NASCIMENTO_DATA	10035	367,0	30.678,0	229.651.927	22.885,09	26.066.847	-0,795	0,024	0,484	0,049
	CEP_RESIDENCIA	10035	2,0	87,0	342.196	34,10	765	1,045	0,024	-0,545	0,049
	STATUS	10035	0,0	1,0	9.832	0,98	0	-6,817	0,024	44,477	0,049

Fonte: os autores

4.1. RESULTADOS DA APLICAÇÃO DO RPA

De uma maneira simplista pode-se afirmar que classificando toda a amostra de teste como bons pagadores, teríamos um nível geral de acertos de 98%, no entanto isto poderia ter um custo financeiro muito alto, pois a perda com maus pagadores no crédito ao consumidor geralmente equivale ao valor do bem adquirido e o ganho em termos lucro com bons ou maus pagadores pode ser estimado em torno de 10 a 30%. A Tabela 3 mostra os resultados da aplicação do RPA com a utilização do programa de árvores de decisão CART® 6.0, desenvolvido a partir do trabalho de Breiman et al. (1984). A versão foi usada fixando *Priors* = 1,2 para o estado 1 (bom pagador). *Priors* é um comando que atribui valores para as classes, no módulo default trata as classes côo se fossem do mesmo tamanho, independente de seu estado real. Atribuindo 1,2 à classe 1 (bom pagador) equivale a atribuir um peso maior a esta classe no cálculo dos custos/impureza. A regra de construção utilizada foi o índice de GINI com os custos alterados (priors).

Tabela 3a : Matriz de confusão para a amostra de treinamento RPA

Status	**Previsão do modelo**		**Total**	
valor real	Pagam	Não Pagam	Acertos	Erros
Pagam	73,03%	26,97%	72,94%	27,06%
Não Pagam	31,60%	69,40%		

Tabela 3b : Matriz de confusão para a amostra de validação - RPA

Status	**Previsão do modelo**		**Total**	
valor real	Pagam	Não Pagam	Acertos	Erros
Pagam	71,85%	28,15%	71,62%	28,38%
Não Pagam	29,41%	60,59%		

A árvore escolhida foi a que revelou o melhor ROC no módulo de validação (0,6932) apresentando um percentual de acertos de 71,62%. Outro aspecto que deve ser considerado é que através dos diversos comandos (costs, priors, etc) o pesquisador pode escolher a árvore mais conveniente para o seu objetivo, priorizando por exemplo o número de acertos de bons

ou maus pagadores. É importante ressaltar que uma análise mais pormenorizada do modelo de classificação deve envolver o custo das classificações erradas.

Para efeitos de classificação são relacionadas a seguir variáveis que influíram o processo (entre parêntesis uma medida relativa da importância das variáveis): quantidade de prestações (100), data de nascimento (88,6), valor financiado (71,2), valor da prestação (36,5), estado civil (26,5), rendimento do cliente (22,2), data da venda (8,4), CEP da residência (5,0) e sexo (5,0).

4.2. RESULTADOS DA APLICAÇÃO DE REDES NEURAIS

No aspecto metodológico, a construção da rede neural para a determinação do *status* do cliente como os que "não pagam" e os que "pagam" em dia suas prestações foi baseada nas redes de *perceptron*. Para a etapa de treinamento foi considerada 50% da amostra, e o restante para teste. Utilizou-se o software Statistica @ 6.1 onde se testou as seguintes redes: lineares, as de propagação em múltiplas camadas (*multiplayer propragation* ou MLP), rede neural polinomial (*polynomial neural network* ou PNN) e duas redes com função radial (*radial basis function* ou RBF).

A função de ativação da rede foi a sigmoide, com taxa de aprendizagem de 0,09. O objetivo aqui era encontrar as redes que conseguissem julgar a melhor classificação de crédito. Foram testadas 10 redes neurais com diferentes perfis de configuração de neurônios e selecionaram-se as 5 mais eficientes pelos menores erros de treinamento.

Tabela 4 : Comparação das redes aplicadas em treinamento e validação

Rede	Desempenho			Erro Treinamento	Nº de neurônios nas camadas		
	Treinamento	Validação	Teste		Entrada	Inter.(1)	Inter(2)
PNN	97,85%	97,47%	97,94%	96,89%	9	6835	2
Linear	67,23%	67,69%	67,53%	34,40%	9	0	0
RBF	65,46%	64,16%	64,80%	34,28%	8	97	0
Linear	68,31%	67,66%	68,45%	34,41%	8	0	0
MLP	66,99%	66,47%	66,89%	31,18%	9	10	0

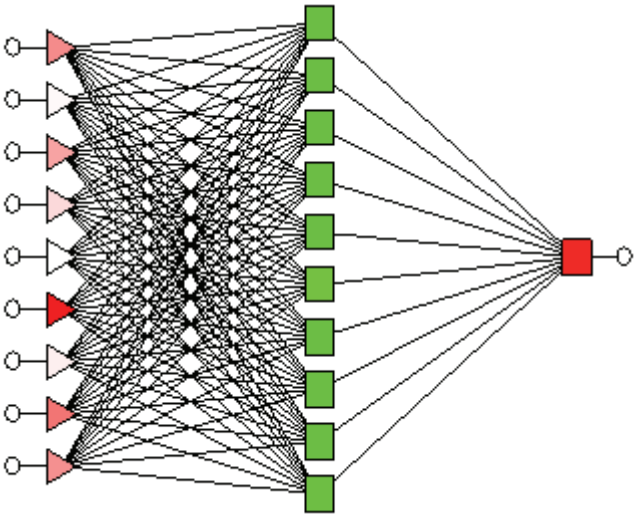

Figura 2 : Arquitetura da melhor rede encontrada na pesquisa

Considerando os resultados obtidos, a melhor arquitetura de rede encontrada foi a MLP com 65,46% de acertos na fase de treinamento, e erro de 14,28%, mostrada na Figura 2. A rede

que apresentou o desempenho menos aceitável foi a PNN, com 96,89% de erro na fase de treinamento. É importante destacar que as redes lineares tiveram desempenho bem próximos ao da MLP. A melhor rede teve-se uma arquitetura de nove neurônios na camada de entrada, sendo um para cada variável, uma única camada intermediária com dez neurônios na ligação e um neurônio na camada de saída.

A amostra de testes, isto é, validação do modelo desenvolvido a partir da amostra de aprendizagem apresentou 10.035 clientes, com 97,07% com classificação "pagam", e 2,93% com classificação "não pagam". Nas Tabela 3 e 4 a seguir são apresentadas a matriz de confusão que denotam os resultados das classificações feitas pela rede MLP tanto na sub-amostra de treinamento e na sub-amostra de validação.

Tabela 5a : Matriz de confusão para a amostra de treinamento usando redes neurais

Status	**Previsão do modelo**		**Total**	
valor real	Pagam	Não Pagam	Acertos	Erros
Pagam	65,91%	34,09%	68,08%	31,92%
Não Pagam	29,74%	70,26%		

Tabela 5b : Matriz de confusão para a amostra de validação usando redes neurais

Status	**Previsão do modelo**		**Total**	
Valor real	Pagam	Não Pagam	Acertos	Erros
Pagam	67,69%	32,31%	67,70%	32,3%
Não Pagam	30,05%	69,95%		

Embora a maior parte da base de dados seja de clientes com *status* 'pagam', a rede classifica corretamente 67,7% desses clientes e, também acerta 69,95% quando um cliente está classificado com 'não pagam' e a rede o enxerga como 'não pagam'. Analisando com mais cuidado os dados, 35,31% dos clientes sofrem uma classificação onde 'não pagam' quando, na verdade, pagam. Essa classificação equivocada pelo modelo pode gerar para a empresa um ônus caro, podendo chegar até a perda do cliente, pois um eventual pedido de crédito pode ser negado, embora o cliente seja um bom pagador. Do mesmo modo, em 30,05% dos casos da amostra de validação, seriam aprovadas operações de crédito a clientes que na realidade não pagariam suas prestações. Nesse caso, o impacto pode ser significativo para a empresa, dado o potencial de perda em função da inadimplência.

Os resultados da tabela a seguir mostram os problemas ocasionados por classificações erradas. O modelo baseado no processamento neural evidencia que, em termos médios, o modelo de rede neural classifica bons pagadores como maus em empréstimos de valor maior (R$ 418,96) e classifica maus pagadores como bons pagadores em empréstimos de valor menor (R$ 205,98).. Desta forma, o modelo de análise de crédito preserva a empresa, a classificação errada de inadimplências está associada a empréstimos de menor montante.

Tabela 6 : Problemas ocasionados por classificações erradas

Perfil Amostra/ Resposta da Rede	Valor da prestação		Valor Total Financiado		Salário familiar	
	Média	DP	Média	DP	Média	DP
Paga / Paga	R$ 55,96	R$ 101,88	R$ 173,11	R$ 170,85	R$ 959,22	R$ 864,03
Paga / Não Paga	R$ 61,49	R$ 140,73	R$ 418,96	R$ 332,33	R$ 689,40	R$ 546,92
Não Paga / Paga	R$ 48,53	R$ 44,11	R$ 205,98	R$ 208,95	R$ 811,69	R$ 625,37
Não Paga / Não Paga	R$ 56,54	R$ 33,49	R$ 453,20	R$ 285,26	R$ 656,61	R$ 419,98

4.3. RESULTADOS DA APLICAÇÃO DE WRAPPER COM SVM

Tabela 7a: Matriz de confusão para a amostra de treinamento, sem seleção de atributos

Status	Previsão do modelo		Total	
valor real	Pagam	Não Pagam	Acertos	Erros
Pagam	72,07%	27,93%	72,08%	27,92%
Não Pagam	27,59%	72,41%		

Tabela 7b : Matriz de confusão para a amostra de treinamento, com seleção de atributos

Status	Previsão do modelo		Total	
valor real	Pagam	Não Pagam	Acertos	Erros
Pagam	76,77%	23,23%	76,77%	23,23%
Não Pagam	23,28%	76,72%		

Tabela 8a : Matriz de confusão para a amostra de validação, sem seleção de atributos

Status	Previsão do modelo		Total	
valor real	Pagam	Não Pagam	Acertos	Erros
Pagam	70,63%	29,37%	70,63%	29,37%
Não Pagam	29,61%	70,39%		

Tabela 8b : Matriz de confusão para a amostra de validação, com seleção de atributos

Status	Previsão do modelo		Total	
valor real	Pagam	Não Pagam	Acertos	Erros
Pagam	74,65%	25,35%	74,65%	25,35%
Não Pagam	25,43%	74,57%		

5. COMENTÁRIOS FINAIS

Os resultados encontrados mostram que de uma forma geral os métodos de *data mining* têm evoluído de acordo com o aumento e barateamento da capacidade computacional e os avanços das ciências matemáticas e técnicas estatísticas. Um dos grandes problemas continua sendo a falta de bases de dados para pesquisas acadêmicas.

Com relação aos procedimentos adotados verifica-se que o procedimento de *wrapper* com *support vector machines* alcançou um resultado de 76,77% de acertos em treinamento e 74,65% na validação. O CART atingiu 72,94% de acertos em treinamento e 71,62% na validação. Já as redes neurais tiveram um resultado um pouco abaixo 68,08% de acertos em treinamento e 67,70% na validação.

O aspecto de acertos e validação ainda deve passar por um crivo muito importante que é o custo das classificações erradas como a Tabela 6 mostra, classificações erradas têm custos diferenciados que podem influir na decisão gerencial de optar por um determinado balanceamento entre acertos de bons e maus pagadores. Este é um aspecto que muitos trabalhos têm negligenciado. Por exemplo, no caso estudado a base de validação tem apenas 2% de maus pagadores. Classificando todos como bons pagadores o percentual geral de acertos será de 98%. Tudo bem se a mercadoria negociada for de baixo custo, mas se o bem for de valor elevado, a classificação errada pode causar sérios prejuízos materiais.

Fica aqui uma sugestão para futuros estudos de que levem em maior consideração o custo das classificações erradas e a possibilidade de ser interesse gerencial um balanceamento entre bons e maus pagadores, de acordo com o valor do bem negociado.

BIBLIOGRAFIA

- AHN, Hyunchul; HAN, Jae Joon; OH, Kyong Joo; KIM, Dong Ha. Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques. USA: *Expert Systems with Applications*, 38, p. 5005-5012, 2011.
- ATTANASIO, Orazio p. Consumption. In TAYLOR, J.B.; WOODFORD, M. (Edit.). *Handbook of macroeconomics.* Vol. 1. Amsterdan: Elsevier Science, 1999, p. 741-812.
- Banco Central do Brasil (BCB). Edital de Audiência Pública nº 37. *Trata da utilização de sistemas internos de risco de crédito (abordagens IRB) para cálculo do Patrimônio de Referência Exigido (PRE)*. Brasília, 18 de fevereiro de 2011.
- Basel Committee on Banking Supervision (BCBS). *The Basel Committee's response to the financial crisis: report to the G20*. Bank For International Settlements, October, 2010.
- BERTOLA, G.; DISNEY, R.; GRANT, C. *The economics of consumer credit demand and supply*. In BERTOLA, G.; DISNEY, R.; GRANT, C. (Edit.). The economics of consumer credit. USA: MIT Press, 2006, p. 1-26.
- BREIMAN, Leo; FRIEDMAN, J.H.; OLSHEN R.A; STONE, C.J. *Classification and regression trees.* Wadsworth International Group, Belmont, California. 1984.
- CARROLL, Christopher D. A Theory of the Consumption Function, with and without liquidity constraints. *Journal of Economic Perspectives*, American Economic Association, vol. 15(3), pages 23-45, Summer, 2001.
- FREITAS A. A. Data mining and knowlwdge discovery with evolutionary algorithms. Springer-Verlag Berlin Heidelberg New York, 1998.
- FRIEDMAN, Milton. *A theory of the consumption function*. USA: Princeton Press, 1957
- GREENE, W.H. *A statistical model for credit scoring*. In JONES S.; HENSHER, D.A. Advances in credit risk modeling and corporate bankruptcy prediction. UK: Cambridge University Press, 2008.
- HAND D. J. e HENLEY, W. E. Statiscal classification methods in consumer credit scoring: a review. *Journal of the Royal Statisical Society*. Series A (Statistics in Society), V. (160) nº 3, p. 523-41, 1997
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning – Data mining, inference and prediction*. USA: Springer, 2009.
- HAYKIN, Simon. Redes neurais: Princípios e prática. Trad. Paulo Martins Engel. 2ª. ed. Porto Alegre: Bookman, 2001.
- KOHAVI, R..; JOHN, G. H. "Wrappers for feature subset selection". Artif. Intell., v.97, 1997. p.273-324.
- LANDO, David. *Credit risk modeling. Theory and applications*. USA: Princeton Press, 2004.
- LI, Hui; SUN, Jie; WU, Jian. Predicting business failure classification and regression tree: an empirical comparison with popular classical methods and top classification mining methods. USA: *Expert Systems with Applications*, 37, p. 5895-5904, 2010
- LIU, H.; MOTODA, H. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Massachusetts, 1998.
- MIN, Jae H; JEONG, Chulwoo. A binary classification metho for bankruptcy prediction. USA: *Expert Systems with Applications*, 36, p. 5256-5263, 2009.

- NOVAK, Michael P. e LaDUE, Eddy. Application of recursive partitioning to agricultural credit scoring. *Journal of Applied Economics*, 31, 1 (April), p. 109-122, 1999.
- PIRAMUTHU S. "On preprocessing data for financial credit risk evaluation". Expert Systems with Applications, v. 30, 2006, p.489-497.
- RAVI, V.; Kurniawan, H.; THAI, Peter Nwee Kok,; KUMAR, P. Ravi. "Soft computing system for bank performance predition". Applied Soft Computing, v. 8, jan. 2008, p.305-315.
- TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. 2006. Introduction to Data Mining. Pearson Education, Inc.Boston USA.
- Vapnik,V.N. Statistical Learning Theory, John Wiley, New York, 1998.
- Witten, I.H.; Frank, E.; HALL, Mark .Data Mining: Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems, 3ª ed. 2011.
- ZHANG, G.; PATUWO, B. E.; HU, M. Y. Forecasting with artificial neural networks: the state of the art. **International Journal of Forecasting**, Kent(Ohio) 14, p. 35–62, 1998.