Revista de Finanças Aplicadas

www.financasaplicadas.net
ISSN 2176-8854

APRENDIZAGEM ESTATÍSTICA APLICADA À PREVISÃO DE DEFAULT DE CRÉDITO

STATISTICAL LEARNING APPLIED TO FORECAST OF CREDIT DEFAULT

Rodrigo Alves Silva
Doutor em Administração de Organizações com foco em Gestão e Análise de Portfólio de Crédito pela FEARP-USP.
Professor da Pontifícia Universidade Católica do Paraná, PUC/PR.
ralves08@usp.br

Evandro Marcos Saidel Ribeiro
Doutor em Física pela UFSCAR.
Professor da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto da Universidade de São Paulo.
esaidel@usp.br

Alberto Borges Matias
Doutor em Administração pela USP.
Professor da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto da Universidade de São Paulo.
matias@usp.br

Recebido em 07/04/2015
Aprovado em 05/10/2015
Disponibilizado 19/08/2016
Avaliado pelo sistema
double blind review

APRENDIZAGEM ESTATÍSTICA APLICADA À PREVISÃO DE DEFAULT DE CRÉDITO

OBJETIVO
Esta pesquisa tem por objetivo comparar a capacidade preditiva das principais técnicas de aprendizagem estatística atualmente utilizadas nacional e internacionalmente para tratamento do problema de risco de crédito, analisando-as à luz de sua eficiência preditiva.

METODOLOGIA
No presente estudo foram utilizados os dados da base German Credit Dataset. As estimações foram obtidas à partir do pacote e1071 aplicado ao software estatístico R®. Foram testadas, comparativamente, as técnicas: Análise Discriminante, Regressão Logística, Redes Bayesianas Naive Bayes, kdB-1, kdB-2, SVC e SVM. Como ponto de corte para Análise Discriminante, Regressão Logística, Redes Bayesianas Naive Bayes, kdB-1, kdB-2 foi empregada a curva ROC - Receiver Operating Characteristic. Os resultados foram comparados com base na acurácia (medida pelo método ACC) e pelo critério de custo pelo erro (medido com base nos critérios trazidos pela base de dados) e os resultados foram apresentados sob o formato de uma matriz de confusão e validados pelo método k-Fold Cross-Validation.

RESULTADOS E CONCLUSÕES
Analisando os resultados das técnicas o SVM apresentou o maior número médio de acertos, obtendo índice de acurácia 86.6%, bem como as maiores médias de acerto em termos de previsão dentre todos os modelos, com aproximadamente 85,5% de acertos na previsão de default.
Pela análise do custo total é possível verificar que o custo do erro do SVM foi de 50.46, sendo que 41.44% desse custo foi produzido por erros de falso positivo e 58.56% do custo foi produzido por erro de falso negativo. Este foi o menor custo dentre os custos produzidos pelas técnicas estudadas, sendo aproximadamente 42% menor que o segundo colocado na análise. Para ambas as métricas (ACC e custos do erro) observou-se a superioridade do método SVM – Support Vector Machine, sugerindo que este método apresenta características que potencializam o seu poder preditivo para o caso estudado. Tais resultados corroboram com indicações de autores que mostram que a superioridade do SVM se deve ao fato de que métodos ensemble, como é o caso do SVM, em dados desbalanceados.

IMPLICAÇÕES PRÁTICAS
Com base nos resultados é possível verificar que a evolução das técnicas de aprendizagem estatística tem contribuído para a maximização dos resultados de classificação, sendo este o problema central da análise de crédito. Pela quantidade de estudos realizados internacionalmente e destacados na bibliografia, este se mostra um campo fértil para o desenvolvimento de pesquisas futuras, especialmente para comparação dos resultados alcançados por este estudo. Especificamente sobre o SVM, os resultados corroboram com estudos que indicam boa aderência de tal modelo para situações reais nas quais o banco de dados a ser classificado é desbalanceado.

PALAVRAS-CHAVE
Risco de crédito, Aprendizagem estatística, previsão de default.

STATISTICAL LEARNING APPLIED TO CREDIT DEFAULT PREDICTION

OBJECTIVE

This research compares the predictive potential of the main statistical learning techniques currently used in both nationally and internationally extent for the treatment of credit risk problem, by analyzing their predictive efficiency.

METHODOLOGY

In this study we have used the data base nominated German Credit Dataset. The estimates were obtained from e1071 package applied in R®. The comparatively seven techniques were tested: linear discriminant analysis, logistic regression, Naive Bayes, KDB-1, KDB-2, SVC and SVM. To discriminant analysis, logistic regression, Naive Bayes, KDB-1, KDB-2 employed the ROC curve - Receiver Operating Characteristic as a cut-off point. The results were compared based on accuracy (measured by the ACC technique) and on error cost criterion (measured by criteria observed in the database studied) and the results were presented in the confusion matrix format and validated by the k-fold Cross Validation technique.

RESULTS AND CONCLUSIONS

The results showed the highest average number of SVM technique hits, getting the highest accuracy (86.6%), as well as the biggest hits in terms of average forecast among all the other models with approximately 85.5% of correct default prediction.

By the total cost results, we can verify that the SVM error cost was 50.46, with 41.44 percent of this cost produced by false positive errors and 58.56% by false negative error. It was the lowest cost among the costs produced by the studied techniques, approximately 42% lower than the second one.

For both metrics (ACC and error costs) this study observed superiority of SVM method suggesting this method has characteristics that enhance the predictive power for the case studied. These results corroborate with results of authors that shows the SVM's superiority is due to the ensemble methods such as SVM, by their nature, are more effective for unbalanced data.

PRACTICAL IMPLICATIONS

Based on the results is possible to check that the evolution of the statistical learning techniques have contributed to the maximization of classification results, being this the central problem of credit analysis. The amount of studies conducted internationally and showed in the bibliographical study, this is a fertile field to the future research development, especially to compare the results obtained by this study. About the SVM method, the results corroborate studies that indicate good adhesion of such model to real situations in which the dataset's unbalanced.

KEYWORDS

Credit Risk, Statistical Learning, Default Prediction.

INTRODUÇÃO

A gestão da carteira de crédito é uma das mais importantes atividades das finanças bancárias e corporativas, tendo em vista o volume de operações presentes em qualquer economia, o volume de recursos movimentados e o peso relativo dos eventos de *default* sobre as empresas. No Brasil, o mercado de crédito, considerando todas as fontes e destinos, representava perto de 31% do PIB – Produto Interno Bruto – em 2007. Este percentual representava cerca de 56,5% em 2013, perfazendo um crescimento de mais de 82% no período segundo dados do Banco Central do Brasil [BCB] (2014).

Autores como Lu *et al* (2013) destacam a importância da otimização das carteiras de crédito. Esta otimização passa, segundo os autores, pela redução do risco de perdas associadas à mudança da qualidade dos tomadores de crédito presentes na carteira, seja por eventos isolados ou mesmo por mudanças em variáveis macroambientais.

Duan & Shrestha (2011) destacam que um dos principais componentes da gestão da carteira de crédito é a análise e previsão de insolvência, também conhecida como default de crédito. Tal trabalho de previsão tem sua importância relacionada às consequências provocadas pelo default sobre a saúde financeira da empresa, seus clientes, fornecedores, funcionários e outros *stakeholders*. Os autores ressaltam ainda que a extensão dos danos provocados por um evento de default sobre a economia e sobre a sociedade como um todo está diretamente relacionado ao tamanho e às características sistemáticas do tomador de crédito em situação de default, mas que, contudo, tanto em pequenas empresas sem qualquer implicação sistemática quanto em grandes empresas, as perdas oriundas destes eventos são potencialmente danosas.

Além de tais questões, especialmente para os bancos, regulações nacionais e internacionais têm tornado cada vez mais importante o desenvolvimento de modelos de tratamento e gestão de risco de crédito, tais como os acordos da Basiléia (Jacks, 2008).

É neste contexto que emerge a aprendizagem estatística como uma das áreas que têm fornecido técnicas úteis à predição de eventos de *default*. Autores como Hastie, Tibshirani & Friedman (2008) entende a aprendizagem estatística como uma das áreas que têm papel chave no trabalho de predição de dados econômicos, com grande aplicabilidade em risco de crédito, volatilidade de ativos, dentre outras.

Os métodos de aprendizagem estatística tradicionalmente empregados para tratamento deste tipo de problema atuam de forma a modelar um conjunto de variáveis preditoras de forma tal que essas ofertem um *output* - como uma previsão - de uma dada variável predita. As técnicas simulam, desta forma, relações entre as variáveis preditoras (independentes) e as variáveis preditas (dependentes) em uma abordagem conhecida como aprendizagem supervisionada (Hastie, Tibshirani & Friedman, 2008).

Aplicadas ao crédito, tais técnicas criam modelos de previsão comumente conhecidos como "modelos de classificação", os quais buscam predizer, com base nos dados do solicitante de crédito, se o mesmo pertenceria ao grupo de solicitantes solventes (que tradicionalmente honram seus compromissos) ou insolventes (que tendem a não honrar seus compromissos). Segundo Lu *et al* (2013) as decisões de crédito, bem como as decisões relacionadas à composição do portfólio de crédito, são fortemente influenciadas por modelos quantitativos de classificação para previsão e otimização. Estes modelos são mecanismos importantes para a gestão das carteiras de crédito, servindo de bases para a gestão e desenvolvimento de meios para minimização de dados provenientes de eventos de *default*, auxiliando na alocação de capital econômico e na formação de *ratings* da carteira de crédito.

Ainda segundo Lu *et al* (2013) grande parte destes modelos estão dedicados à minimização das perdas esperadas, contudo, a otimização do crédito vem de um equilíbrio entre as perdas derivadas dos eventos de *default* e das perdas derivada de políticas de restrição de crédito. Segundo Duan & Shrestha (2011) novas tecnologias e pesquisas voltadas para o desenvolvimento de modelos, métodos e técnicas de análise crédito, especialmente para a previsão de *default*, têm sido desenvolvidas no intento de otimizar carteiras de crédito.

Tendo como perspectiva estes avanços a presente pesquisa tem por objetivo **comparar a capacidade preditiva de modelos das principais técnicas de aprendizagem estatística atualmente utilizadas** nacional e internacionalmente, analisando-as à luz de sua eficiência preditiva.

2 REFERENCIAL TEÓRICO

A aprendizagem estatística vem sendo amplamente empregada em análise de risco de crédito, demonstrando grande capacidade preditiva e flexibilidade para incorporação dos vários riscos associados à atividade de crédito. A evidenciação dos conceitos de análise e concessão de crédito, da importância da previsão de eventos de *default* e das propriedades que tornam a aprendizagem estatística interessante para o tratamento da problemática é foco do presente referencial teórico.

2.1 Risco de crédito e previsão de *default*

A concessão de crédito é um tipo de operação financeira na qual incide vários tipos diferentes de riscos. Autores como García, Giménez & Guijarro (2013) destacam cinco tipos de riscos associados à concessão de crédito, quais sejam:

- Risco de mercado: proveniente da observação de mudanças nos preços ou taxas praticadas no mercado;
- Risco de crédito: proveniente de mudanças na qualidade do tomador de crédito ou na carteira de crédito que provocam a redução do valor desta carteira;
- Risco de liquidez: proveniente do aumento do custo financeiro da liquidez da carteira de crédito ou mesmo de dificuldades de acesso ao financiamento de suas atividades via carteira de crédito;
- Risco operacional: proveniente de fatores humanos na execução da operação, tais como: fraudes, erros humanos ou mesmo falhas nos sistemas de informações; e
- Risco sistemático: proveniente de fatores externos que afetam todo o mercado de crédito, tais como crises, guerras, dentre outros.

Especificamente sobre risco de crédito, como abordado na definição anterior, o mesmo pode ser entendido como a probabilidade de o tomador não cumprir com os compromissos financeiros firmados. A esse descumprimento dá-se o nome de evento de *default* (Duan & Shrestha, 2011, Mathias, 2007). Este conceito está associado a dois elementos principais:

a) Associada à mudança da qualidade pelo não cumprimento dos compromissos que provoca uma mudança no *status* do tomador de adimplente para inadimplente e que poderá, posteriormente, alterar-se novamente de inadimplente para insolvente.

b) Associado à mudança na expectativa de que este tomador honre seus compromissos, podendo, num futuro próximo, entrar em *default*.

Em uma carteira de crédito, estes fatores combinados alteram a qualidade da carteira, alterando assim o risco associado à mesma. Como abordado por Saita (2007) o risco de crédito associa uma série de tipos de perdas, as quais variam segundo a metodologia utilizada pela firma que concede o crédito. Para o presente artigo, se limita à previsão do evento de *default* como elemento do risco de crédito. Segundo Duan & Shrestha (2011) a previsão do evento de *default* é um dos principais elementos da análise de risco de crédito, servindo de base para a gestão do portfólio em suas expectativas e valores.

Bluhm, Overbeck & Wagner (2003) indicam que, em teoria de probabilidade, a expectativa é a definição de esperança, isto é, uma média ponderada de resultados das probabilidades associadas aos possíveis estados da natureza, refletindo um valor esperado. Segundo os autores, este é um conceito também aplicado à gestão de riscos de crédito, na qual a ideia é atribuir uma expectativa da probabilidade de *default* para cada cliente que compõe a carteira - DP (*Default Probability*), bem como de uma porção de perda incorrida pela firma em caso de *default* dos clientes – LGD (*Loss given default*), descrevendo a porção de empréstimos que será perdida em caso de *default*, bem como a exposição total ao *default* - EAD (*Exposure at default*) no período.

A expectativa de perda é a esperança da perda proveniente das variações nas probabilidades de *default*, nas perdas dado *default* e na exposição ao *default* da firma:

$$EL = E\left[L\right] \; = \; EAD \; x \; LGD \; X \; L \;\; \text{Com: L=1D, P(D)=DP.}$$

Em que: D representa o evento onde o cliente entra em default em algum momento do tempo, normalmente em anos. P(D) é a probabilidade de ocorrência do default (D). EAD é o montante exposto ao risco de default e LGD é a porção de perda esperada em caso de default.

EAD e LGD assumem valores constantes e L assume valores de uma variável de Bernoulli. As probabilidades de *default* podem ser obtidas à partir de processos de calibração, seja por modelos internos ou mesmo por agências de Ratings, tais como: Mood's Investors Services, Standard & Poor's e Fitch ou por dados de mercado, através de modelos como KMV, Credit Metrics, Credit risk+, dentre outros (Bluhm, Overbeck & Wagner, 2003).

Como referenciado por Bluhm, Overbeck & Wagner (2003) as estimativas de expectativa de *default* são importantes pela capacidade de maximizar a eficiência dos processos de gestão de carteiras de crédito, auxiliando na decisão de concessão, estimativas de volume de cobertura a expectativas de perdas (alocação de capital econômico), análise e definição de prêmios pelo risco presentes nas taxas de retorno exigidas ao tomador pelo fornecedor de crédito, dentre outras decisões relevantes no contexto.

Segundo Annibal & Koyama (2011) os modelos internos são amplamente empregados no Brasil, especialmente por idiossincrasias existentes no mercado brasileiro que dificultam o emprego integral destes modelos. Tais elementos, associados ao desenvolvimento tecnológico e a marcos regulamentais (tais como os Acordos da Basileia), produziram um grande número de modelos baseados em métodos quantitativos, que objetivaram a otimização do *trade-off* risco e retorno em carteiras de crédito, especialmente aqueles voltados classificação e previsão de risco de *default* de crédito.

2.2 Aprendizagem estatística em análise de risco de crédito

A aprendizagem estatística é uma área dentro das ciências estatísticas dedicada à produção de ferramentas e métodos voltados pra a compreensão de dados, sendo de grande valia dentro de finanças corporativas. Segundo James *et al* (2013) a aprendizagem estatística envolve ferramentas supervisionadas (modelos que buscam prever o estimar uma saída com base em uma ou mais entradas) e não supervisionada (onde existem entradas, mas não saídas do sistema). Ainda segundo os autores, dadas as propriedades da aprendizagem supervisionada e as características dos problemas, esta é a mais empregada para problemas nas áreas de negócios.

Para os autores, ainda que o termo aprendizagem estatística seja recente, as ferramentas que a compõem datam desde o século XIX, incluindo ferramentas amplamente conhecidas dentro das áreas de economia e negócios, tais como: regressões lineares, regressões discriminantes, regressões logísticas, modelos generalizados, dentre outros. De acordo com Duan & Shrestha (2011), os métodos baseados em regressões discriminantes (tais como *Altman Z-score*) e as regressões logísticas são os mais difundidos dentro da área em âmbito mundial.

A análise discriminante, ou regressão discriminante é um método voltado para a classificação de entidades em um número finito de classes (g), com base em um conjunto de variáveis. Estas entidades podem indivíduos ou objetos, sendo que as variáveis denotam as características destas entidades (McLachlan, 2001).

Considerando a sua forma funcional a regressão discriminante (linear no caso estudado) pode ser representada da seguinte forma:

$$Z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n$$

Em que: Z representa a variável predita que se converte em um dos estados da natureza que a entidade pode assumir, α representa a constante e $\beta_1 ... \beta_n$, representam os coeficientes das variáveis (ou características) e $x_1 ... x_n$ são as variáveis discriminantes (ou características).

Os métodos de regressão discriminante são amplamente empregados pela sua semelhança aos métodos de regressão múltipla, facilitando a interpretação dos resultados de sua estimação, bem como de sua inferência. Autores como Duan & Shrestha (2011) indicam o uso de tal técnica em modelos como o conhecido Altman's Z-score, dentre outros.

A função linear discriminante define linhas limites de classificação, sendo esta, fruto de uma combinação linear entre as variáveis preditoras. O poder de discriminação do modelo está, então, limitado à linearidade das relações entre as variáveis preditoras e a predita.

A regressão logística é uma das técnicas mais utilizadas para a área de análise de risco de crédito. Apresenta como característica que a difere da regressão linear discriminante a possibilidade de identificação de crescimento não linear do *default* sob um formato de uma função sigmoide, com crescimento acelerado, gerando maior acurácia preditiva em muitos casos (McLachlan, 2001). Segundo McLachlan (2001) este método é bastante utilizado em situações onde a variável dependente assume valores dicotômicos, como é o caso dos problemas de decisão de concessão de crédito. Sua execução consiste em estimar a probabilidade de ocorrência de um evento com base em um conjunto de variáveis. Em sua forma funcional a regressão logística pode ser representada por:

$$P_a = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}}$$

Em que Pa representa a probabilidade apurada de uma entidade assumir um dado valor (normalmente expresso em termos de uma variável dicotômica), $\beta_1 ... \beta_n$, representam os coeficientes das variáveis (ou características) e $x_1 ... x_n$ são as variáveis ou características.

A principal diferença entre os modelos lineares e a regressão logística é que, na regressão logística não se busca modelar diretamente uma resposta para a variável a ser estimada (James, *et al*, 2013). Ao invés disto, neste método busca-se estimar a probabilidade de Y pertencer a uma determinada categoria.

Mais recentemente os modelos baseados em teoria de grafos probabilísticos, tais como árvores probabilísticas e redes bayesianas têm ganhado destaque na área. Estudos como o de Pascual, Martinéz & Alamillos (2013), Brown & Mues (2012), Danenas, Garsva & Gudas (2011) dentre outros, têm se dedicado a demonstrar a eficiência das redes bayesianas na análise e tratamento do risco de crédito. Segundo Pascual, Martinéz & Alamillos (2013) estas técnicas têm se mostrado uma poderosa ferramenta para tratamento deste tipo de problema.

Nagarajan, Scutari & Lèbre (2013) definem uma rede bayesiana como uma classe de modelos gráficos probabilísticos que têm por objetivo construir uma representação de dependência probabilística entre um conjunto de variáveis aleatórias X=[X1, X2, ..., Xp] representadas em um grafo sob a forma de nós e arcos organizados em função da relação de precedência dos mesmos, sendo que este grafo pode ser direto ou indireto.

Modelos gráficos probabilísticos são baseados em uma estrutura de dependência de uma tabela de probabilidade condicional – CBT – que expressa a probabilidade de ocorrência de um evento de interesse dada a ocorrência de outro evento, isto é:

$$p((x_1, x_2, \dots x_n) = \prod_{j=1}^{n} p\left[x_1 | PAR_g(x_j)\right]$$

Em que *PARg* representa as variáveis que precedem x_i (conhecidas como pais de xi).

Estes nós (nodes) são ligados a outros nós por meio de arcos. Uma vez ligado um nó a um arco, estabelece-se uma relação de interdependência, onde o nó de origem (no caso de grafos diretos e acíclicos) e nomeado de pai e o nó de destino é nomeado como filho.

Como referenciado à priori, o estudo empregará os modelos de redes bayesianas Naive Bayes e kdB *(k-dependence Bayesian Network)*.

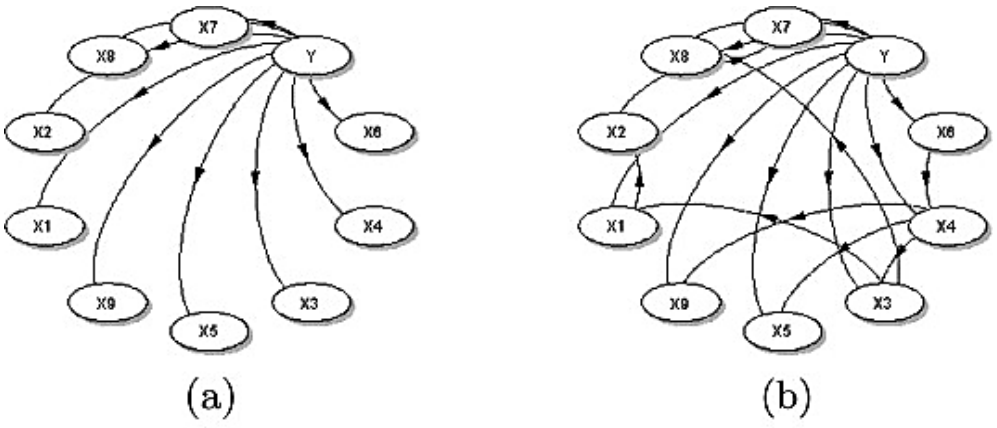
Segundo Pascual, Martinéz & Alamillos (2013) o Naive Bayes tem como premissa fundamental a independência entre as variáveis, isto é, um nó filho não poderá ser pai de outro nó pai. Na rede bayesiana kdB *(k-dependence Bayesian Network)*, ao contrário do Naive Bayes, existe uma relação de precedência onde, pelo menos um filho é pai de outro filho. Ambas as proposições podem ser observadas na figura 1.

Como é possível observar na figura 1 (a), não existe relação de dependência entre os nós filhos, isto é, um filho de "y" não é pai de outro filho de "y". Segundo Pascual, Martinéz e Alamillos (2013) esta é um pressuposto que não encontra sustentação na realidade, contudo, como expressam os autores, têm surtido bons resultados em muitos casos.

Na figura a (b), é representado o kdB-1. Nessa estrutura existe, pelo menos um filho, que é pai de outro. Nesta estrutura o número 1 (kdB-1) significa que existe pelo menos um filho que é pai de um outro filho. A estrutura kdB pode assumir então n relações de dependência, o

que implica em dizer a possibilidade de existência de um kdB-n. Já no caso do kdB-0, a ideia é a de inexistência de dependência condicional entre as variáveis preditoras (ou classes), o que implica dizer que o kdB-0 tem suas premissas idênticas às do Naive Bayes.

Figura 1. Estrutura Naive Bayes



Fonte: Louzada & Ara (2012, p.11585).

Considerando os elementos que compõem um problema de *credit scoring* Pascual, Martinéz & Alamillos (2013) entendem que técnicas de estimação de estruturas, tais como as redes bayesianas, são estatisticamente interessantes para o tratamento desta problemática. Adicionalmente Duan & Shrestha (2011) indicam a evolução recente nos métodos computacionais trouxeram novos métodos, quais sejam: redes neurais artificiais, método *SVM – Suport Vector Machine e PIM – Poisson Intensity Model*. Para o presente estudo, como indicado anteriormente, serão abordados os modelos SVC – *Support Vector Classifier* e SVM – *Support Vector Machine*. Considerando o conceito de hiperplano, a problemática de classificação, como é o caso da previsão de default, existem muitos casos nos quais não se é possível separar perfeitamente um conjunto de dados em duas (ou mais) classes pela observância de um máximo hiperplano.

O SVC, também conhecido como *Soft Margin Classifier* é um método que baseia a classificação no entendimento de que o máximo hiperplano não necessariamente é o melhor vetor de separação ou classificação das entidades. O método abre mão do máximo hiperplano e determina um hiperplano que seja menor quanto necessário para conseguir o máximo em termos de acerto nas classificações, o que é entendido como o ótimo hiperplano. Tal classificação é obtida à partir de um conjunto de dados de treinamento onde o algoritmo busca fornecer a melhor separação possível dentro do conjunto de dados (James, *et al*, 2013).

O *Support Vector Machine* – SVM é uma extensão do SVC que tem com principal diferença para o SVC a possibilidade de incorporação de estimações não lineares para as classes das quais as entidades são pertencentes (James, *et al*, 2013). Desta forma, o SVM abre mão do conceito de hiperplano como sendo linear.

Segundo James, *et al* (2013) os principais benefícios do emprego do SVM para a problemática de risco de crédito são três: i) requer menos concepções sobre os dados, tais como linearidade e continuidade; ii) possibilidade de realização de um mapeamento não linear da estrutura dos dados; e iii) capacidade de implementação de uma estrutura de minimização de risco (SRM – *Structure Risk Minimization*) à partir de algoritmos que buscam aprender o hiperplano que maximiza as margens.

3 PROCEDIMENTOS METODOLÓGICOS

No presente estudo foram utilizados os dados da base *German Credit Dataset*, disponibilizada por Bache & Lichman (2013). As estimações foram obtidas à partir do pacote e1071 aplicado ao software estatístico R (R Team, 2008).

Esta base de dados é composta de mil observações, das quais 30% são classificadas à priori como em *default* e o restante como não *default*. Como variáveis preditoras a base traz um conjunto de 20 (vinte variáveis) e uma variável predita. Deste conjunto um total de sete variáveis são contínuas (as quais foram discretizadas dada a necessidade para inclusão no modelo Naive Bayes), cinco variáveis ordinais, seis variáveis nominais e três variáveis binárias (incluindo a variável predita). Seu tamanho e o pequeno número de *missing values* fazem desta uma das mais utilizadas atualmente para esta finalidade (Bache & Lichman, 2013), sendo utilizada em estudos como os de Duan & Shrestha (2011), Pascual, Martinéz e Alamillos (2013) e Louzada e Ara (2012). A composição desta base pode ser observada no quadro 1.

Quadro 1. Variáveis presentes na base de dados

Variável	Descrição	Tipo de Variável	nº Categorias
Checking	Status da conta do solicitante em relação ao salário declarado	Ordinal	4,00
Duration	Duração do empréstimo requerido	Contínua	-
history	histórico de créditos anteriores	Ordinal	4,00
purpose	Finalidade do crédito requerido	Nominal	11,00
amount	Montante do crédito requerido	Contínua	-
savings	Economias do cliente	Ordinal	5,00
employ	Tempo no emprego atual	Ordinal	5,00
rates	Juros praticados (%) no empréstimo requerido	Contínua	-
status	Sexo e estado civil	Nominal	5,00
debtors	Outras dívidas	Nominal	3,00
residence	Tempo na residência atual	Contínua	-
property	Propriedades	Nominal	4,00
Age	Idade	Contínua	-
other_inst	Outros planos de parcelamento	Nominal	3,00
housing	Tipo de moradia	Nominal	3,00
exist_cr	Nº de créditos concedidos no banco	Contínua	-
Job	Profissão	Ordinal	4,00
provider	Número de dependentes	Contínua	-
phone	Telefone próprio	Binária	2,00
foreign	Estrangeiro	Binária	2,00
goodbad	Qualificação do cliente como bom ou mal pagador	Binária	2,00

Fonte: Bache e Lichman (2013)

Foram testadas, comparativamente, as técnicas: Análise Discriminante, Regressão Logística, Redes Bayesianas Naive Bayes, kdB-1, kdB-2, SVC e SVM. Como ponto de corte para Análise Discriminante, Regressão Logística, Redes Bayesianas Naive Bayes, kdB-1, kdB-2 foi empregada a curva ROC - *Receiver Operating Characteristic*. Para os modelos SVC e SVM foi empregada a ferramenta *"best model"* que maximiza a acurácia do modelo frente aos dados de treinamento. A acurácia, empregada como métrica estatística é uma medida amplamente empregada dentro de estudos de classificação.

No presente estudo a acurácia segue os procedimentos utilizados por Iscoe *et al* (2012), sendo apresentada sob a forma de uma matriz de confusão de estimação e observação. A acurácia, pelo critério ACC é dada por:

$$ACC = \frac{Tp + Tn}{Tp + Fp + Tn + Fn}$$

Em que *Tp* é o número de classificações a posteriori verdadeiras positivas, *Tn* é o número de classificações a posteriori verdadeiras negativas, *Fp* é o número de classificações a posteriori falsas positivas e *Fn* é o número de classificações a posteriori falsas negativas.

Além das estatísticas de acurácia a pesquisa apresentada, sob a forma de matriz de confusão, o custo do erro de classificação. Este custo é definido pelo próprio conjunto de dados e indica que o custo do erro de classificação de falso negativo é cinco vezes maior do que o custo do erro de classificação de falso positivo (Bache & Lichman, 2013).

A comparação por custos é bastante utilizada, tendo em vista o objetivo de otimização dos modelos. Desta forma, a comparação fica mais eficiente, pois sabe-se qual é o tipo de erro de classificação que apresenta maior custo para a organização. Para validação foi empregado o método *k-Fold Cross-Validation*. Este método consiste em dividir em grupos aleatórios e uniformes a base de dados (em k grupos), sendo todos de mesmo tamanho aproximadamente. Neste método o primeiro grupo fica como grupo de teste ou validação e os demais para treinamento. Em um segundo momento o segundo grupo é requerido para validação enquanto os demais são utilizados para treinamento e assim sucessivamente de forma que se obtém uma estimativa do erro para o conjunto de amostras. A estimação por *–Fold Cross Validation* pela média do valor dos erros como segue:

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

Em que MSEi representa a média do erro quadrático e k representa o número de grupos utilizados para validação.

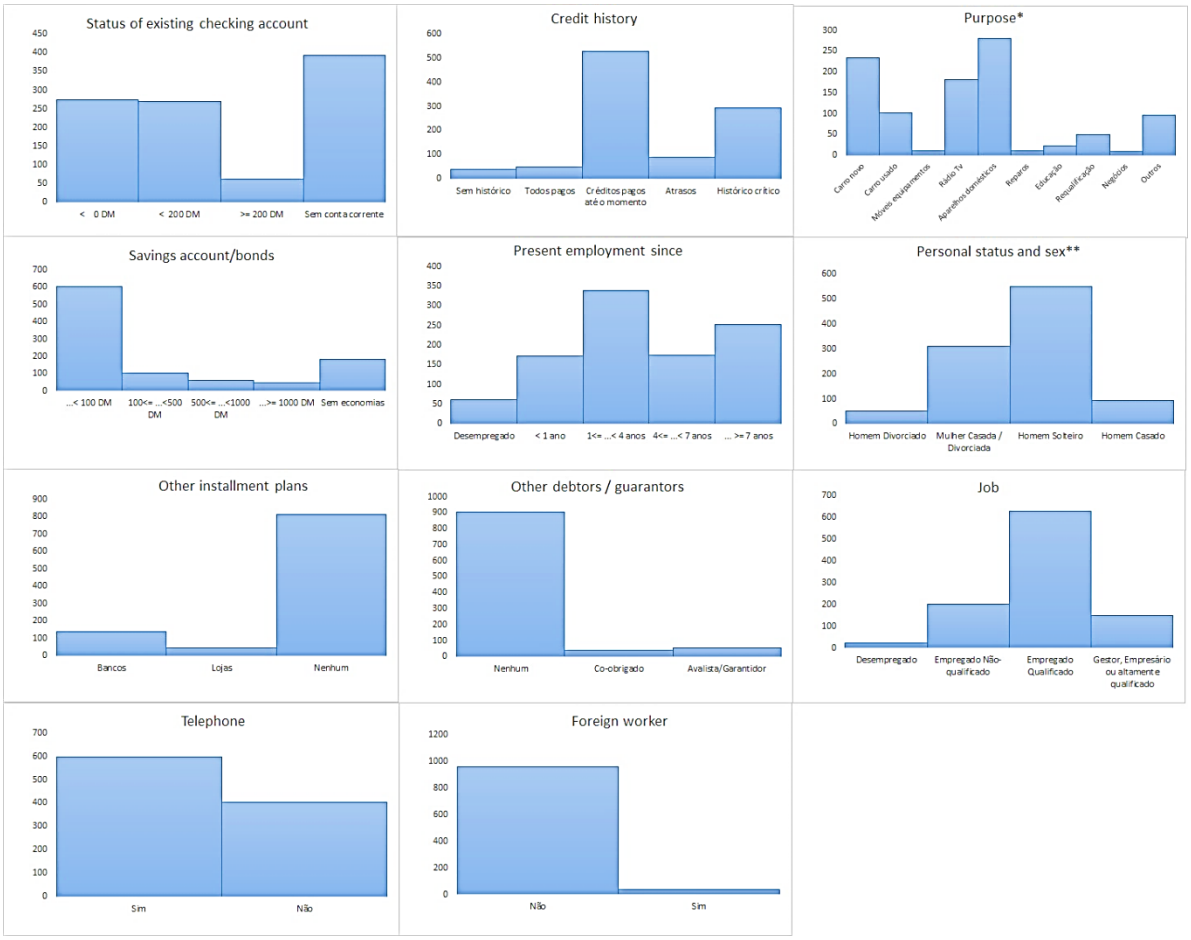
No presente estudo, este procedimento separou amostras de 20% do total de observações do conjunto de dados para teste, ficando 80% para treinamento para todos os métodos testados. Para entrada dos dados de forma idêntica em todas as técnicas optou-se, por limitação do Naive Bayes e kdB, pela discretização dos dados. Não foram realizados tratamentos dos dados para balanceamento das amostras seguindo procedimento utilizado por Chung, Ho & Hsu (2011) que sugerem que amostras artificialmente balanceadas não são fiéis à realidade dos dados ori-

ginais. Os procedimentos de estimação e validação foram repedidos 100 (cem) vezes para cada uma das técnicas analisadas.

4 RESULTADOS E DISCUSSÃO

Para iniciar a análise dos dados foram apuradas as estatísticas descritivas dos dados.

Figura 2. Distribuição da frequência de variáveis categóricas



Fonte: Elaboração Própria.

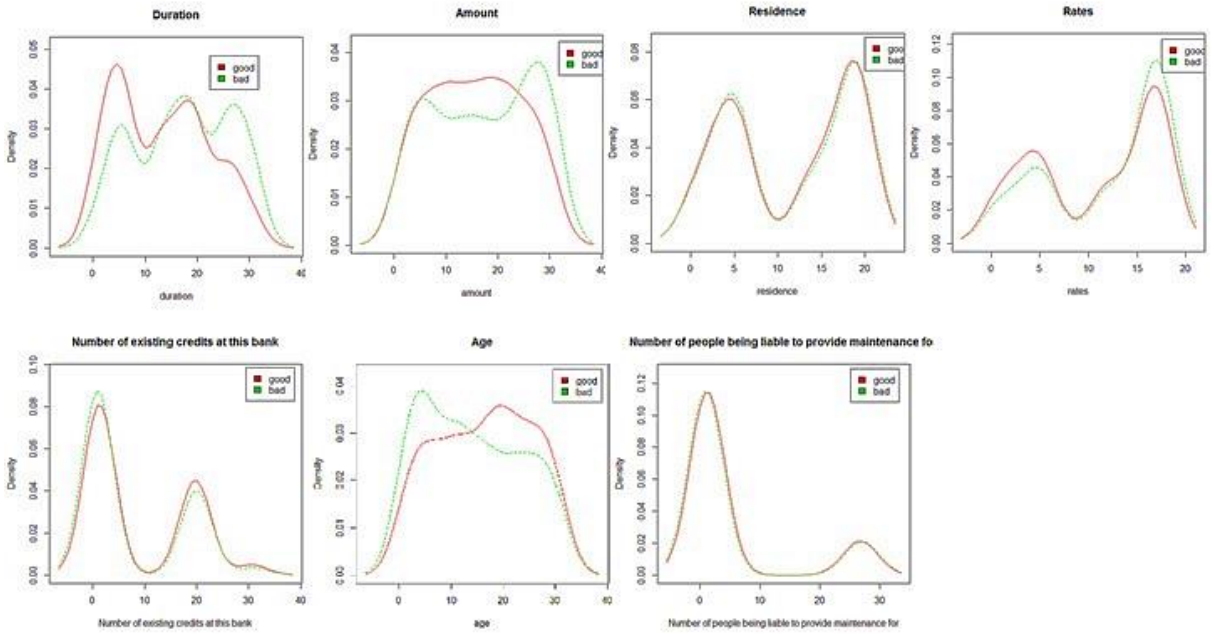
Notas: * Purpouse: A classe A47 – férias – não é observada na base de dados; ** Personal status and sex: A classe A95 – Mulher Solteira – não é observada na base de dados.

Na figura 2 são apresentados os gráficos das estatísticas descritivas das variáveis cuja base de dados classificara à priori como ordinais ou nominais. Como é possível observar as variáveis apresentam grande volatilidade no que tange a concentração, havendo parte das variáveis com grande concentração em uma das classes (ou estados da natureza) e outras com distribuição bastante equilibrada. Este comportamento privilegia o desenvolvimento de ferramentas não enviesadas, já que suas observações têm natureza aleatória.

Na figura 3 são apresentados os comportamentos das variáveis contínuas, quais sejam: a duração do empréstimo (prazo), montante emprestado, tempo de residência do proponente, ta-

xas praticadas, números de créditos anteriores, idade do proponente e o número de pessoas sustentadas pelo proponente. As curvas contínuas em vermelho representam os clientes que não entraram em *default*, enquanto as tracejadas em verde denotam clientes em *default*.

Figura 3. Comportamento do *default* em relação às variáveis contínuas da base



Fonte: Elaboração Própria.

Estes gráficos foram obtidos após discretização dos dados, sendo esta uma necessidade para algumas das técnicas empregadas como tratado nos procedimentos metodológicos.

Pelo comportamento do gráfico é possível observar que créditos em curto prazo apresentam menor probabilidade de *default* do que em longo prazo para este banco de dados, Assim como clientes com montante concentrados na classe 30 têm maior probabilidade de entrar em default, ainda que exista certa uniformidade de dados de bons pagadores. Nota-se ainda, pela análise da variável "age", uma concentração de bons pagadores (clientes que não entraram em default) na faixa 20 a 30. Estas faixas representam clientes com idades entre 35 e 55 anos, indicando que ocorre um maior risco para clientes de menor idade.

Como apresentado anteriormente, a pesquisa aplicou as seguintes técnicas de previsão de *default*: Análise Discriminante, Regressão Logística, Redes Bayesianas Naive Bayes, kdB-1, kdB-2, SVC e SVM. A classificação de acertos e erros foi realizada pela comparação entre a classificação obtida à partir da previsão de cada modelo de forma separada e a classificação trazida à priori pelo próprio banco de dados.

Na tabela 1 as classificações do grupo "observado" são aquelas trazidas pelo banco de dados à priori, enquanto as nomeadas como "predito" são aquelas trazidas pelo modelo de previsão. ND representa os clientes classificados como "Não *Default*" enquanto D representa clientes classificados como "*Default*" e, sob a sigla ACC, a tabela apresenta a acurácia. Adicionalmente a tabela traz o percentual de acerto para cada possível estado da natureza.

Tabela 1. Desempenho De Acurácia

Análise Discriminante

Observado		Predito ND	Predito D	%
		1	2	
ND	1	87.5	52.5	62.5
D	2	15.9	44.1	73.5
ACC			65.8	

Regressão Logística

Observado		Predito ND	Predito D	%
		1	2	
ND	1	104.02	35.98	74.3
D	2	17.22	42.78	71.3
ACC			73.4	

Naive Bayes

Observado		Predito ND	Predito D	%
		1	2	
ND	1	87.26	52.74	62.3
D	2	6.79	53.21	88.7
ACC			70.235	

KDB-1

Observado		Predito ND	Predito D	%
		1	2	
ND	1	94.75	33.75	73.74
D	2	28.00	43.50	60.84
ACC			69.125	

KDB-2

Observado		Predito ND	Predito D	%
		1	2	
ND	1	82.75	54.10	60.47
D	2	35.90	27.25	43.15
ACC			55	

SVC

Observado		Predito ND	Predito D	%
		1	2	
ND	1	124.27	32.64	79.2
D	2	15.36	27.73	64.4
ACC			76	

SVM

Observado		Predito ND	Predito D	%
		1	2	
ND	1	138.45	20.91	86.88
D	2	5.91	34.73	85.46
ACC			86.59	

Fonte: Elaboração Própria.

Desta forma, analisando os resultados das técnicas o SVM apresenta o maior número médio de acertos, obtendo o maior índice de acurácia, bem como as maiores médias de acerto em termos de previsão dentre todos os modelos.

Pela tabela, o valor observado de positivos à priori, no caso das amostragens do SVM, foi de mais de 159 classificações em média para ND e pouco mais de 40 para D. Como é possível observar, o método classificou corretamente 86.88% das observações de ND e 85.46% das observações de D, isto é, o método teve um desempenho muito próximo na classificação de cada um dos estados da natureza.

Este comportamento não é observado para todos os métodos, como é o caso do Naive Bayes que classificou corretamente 88.7% de D. Desta forma, este é o método que apresentou o melhor resultado na previsão do indivíduo em *Default*. Em parte este resultado se deve ao direcionamento do algoritmo para previsão de verdadeiro negativo, com emprego da curva ROC

como mecanismo de classificação, maximizando a especificidade e a sensibilidade do método ao conjunto de dados.

Para melhorar o entendimento e fundamentar as considerações acerca dos métodos e seu potencial de predição, se mostra importante analisar os custos acerca dos erros de classificação. Como informado nos procedimentos metodológicos estes custos são trazidos pelo próprio conjunto de dados e apresentam uma razão de cinco para um em favor do falso negativo, isto é, o custo do erro de falso negativo é cinco vezes maior do que o custo do erro de falso positivo.

Tabela 2. Desempenho de custos

Análise Discriminante

Observado		Predito ND	Predito D	%
		1	2	
ND	1	0	52.5	39.77
D	2	79.5	0	60.23
Custo Total			132	

Regressão Logística

Observado		Predito ND	Predito D	%
		1	2	
ND	1	0	35.98	29.47
D	2	86.1	0	70.53
Custo Total			122.08	

Naive Bayes

Observado		Predito ND	Predito D	%
		1	2	
ND	1	0	52.74	60.84
D	2	33.95	0	39.16
Custo Total			86.69	

KDB-1

Observado		Predito ND	Predito D	%
		1	2	
ND	1	0	33.75	19.42
D	2	140	0	80.58
Custo Total			173.75	

KDB-2

Observado		Predito ND	Predito D	%
		1	2	
ND	1	0	54.10	23.16
D	2	179.5	0	76.84
Custo Total			233.6	

SVC

Observado		Predito ND	Predito D	%
		1	2	
ND	1	0	32.64	29.82
D	2	76.8	0	70.18
Custo Total			109.44	

SVM

Observado		Predito ND	Predito D	%
		1	2	
ND	1	0	20.91	41.44
D	2	29.55	0	58.56
Custo Total			50.46	

Fonte: Elaboração Própria.

Na tabela 2 os percentuais referem-se à proporção de cada erro para o custo total do erro gerado pela aplicação do modelo aos dados. Este custo é obtido à partir da multiplicação no número médio de erros de cada classificação pelo seu respectivo custo de erro de classificação declarado no conjunto de dados.

Pela análise do custo total é possível verificar que o custo do erro do SVM foi de 50.46, sendo que 41.44% desse custo foi produzido por erros de falso positivo e 58.56% do custo foi produzido por erro de falso negativo. Este foi o menor custo dentre os custos produzidos pelas técnicas estudadas. O segundo menor custo produzido foi aferido pelo Naive Bayes, contrariando o que se observou na análise do ACC, no qual o resultado obtido pelo Naive Bayes ficou inferior ao resultado da Regressão Logística e do SVC. Isto ocorreu pela adequação apurada pelo Naive Bayes na predição de *defaults*, o qual apresenta maior custo de erro, reduzindo assim a penalização por erros totais deste modelo em relação aos demais.

Em seu resultado Brown e Mues (2012), encontraram indícios de que métodos *ensemble* (métodos que misturam outros classificadores com o objetivo de maximizar o resultado preditivo), tais como *boosting* gradiente e *randon forests*, são melhores para dados desbalanceados. Os autores sugerem que, na medida em que são desbalanceadas as classes, métodos baseados em árvore de decisão (como é o caso do Naive Bayes) se tornam mais ineficientes para predição. Os estudos de Chung, Ho e Hsu (2011) e Louzada-Neto, Ferreira-Silva e Diniz (2012) realizaram o desbalanceamento gradativo de amostras e, para cada desbalanceamento, foram aplicadas as técnicas, encontrando indícios que sugerem que conjuntos de dados desbalanceados desfavorecem a acurácia de modelos bayesianos. Este conjunto de estudos pode explicar desempenho superior para acurácia da técnica logística ao Naive Bayes no caso da métrica ACC.

Já a inclusão de sinalizadores para determinação da aprendizagem (indicação do evento a ser predito), que fazem com que o modelo deve se focar na predição dos indivíduos classificados à priori como *default*, somado ao maior custo do erro de falso negativo, fazem com que os modelos bayesianos, em especial o Naive Bayes tenham apresentado desempenho superior em custos. Para ambas as métricas (ACC e custos do erro) observou-se a superioridade do método SVM – Support Vector Machine, sugerindo que este método apresenta características que potencializam o poder seu preditivo para o caso estudado.

5 CONSIDERAÇÕES FINAIS E RECOMENDAÇÕES

Com o objetivo de analisar o desempenho de classificação de técnicas de aprendizagem estatísticas para um problema de previsão de *default* de crédito, o presente artigo utilizou as ferramentas de aprendizagem estatística de classificação de dados Naive Bayes, KDB-1, KDB-2, SVC e SVM em conjunto com técnicas clássicas para tal finalidade, tais como análise linear discriminante, regressão logística, comparando-as através de uma matriz de confusão construída com base na métrica ACC e uma matriz construída com base na métrica de penalização por erros (custo).

O estudo foi realizado junto a um conjunto de dados pessoa física de domínio público, qual seja o *German Credit Data Set*, e seus resultados foram comparados em termos da acurácia de cada modelo e de custos do erro de classificação. Os testes foram realizados através do software estatístico R, utilizando o pacote e1071.

Em termos de acurácia de classificação, a técnica SVM se mostra a mais eficiente tendo classificado 86,59% das amostras de forma correta. Este desempenho é aproximadamente 14% maior do que o segundo colocado em termos de assertividade na classificação, qual seja o método SVC.

Dada a característica de otimização financeira, o custo dos erros de classificação se mostra uma importante métrica para o estudo (penalização pelo erro). Neste aspecto a técnica SVM

se mostrou a mais eficiente com custo 46,79% menor que o custo apurado pelo segundo colocado, qual seja o classificador bayesiano Naive Bayes.

Estudos futuros que repliquem os procedimentos apresentados no presente estudo em outras bases de dados, fazendo uso das técnicas aqui empregadas, bem como incorporando novas técnicas, seriam desejáveis para que se analise o desempenho destes métodos para que, desta forma, seja possível confrontar os resultados obtidos. Este confronto auxilia na construção de modelos mais eficientes para a previsão de default de crédito.

O presente estudo cumpre seus objetivos ao demonstrar comparativamente a capacidade preditiva de sete técnicas de aprendizagem estatística para tratamento do problema de risco de crédito, demonstrando características e peculiaridades de cada uma dessas técnicas na composição de modelos de análise e concessão de crédito.

REFERÊNCIAS BIBLIOGRÁFICAS

ANNIBAL; C. A., & KOYAMA, S. M. (2011). Pesquisa trimestral de condições de crédito no Brasil. Trabalhos para discussão, Brasília, n. 245, p. 1-62, Jul. 2011.

BACHE, K., & LICHMAN, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, Disponível em: http://archive.ics.uci.edu/ml.

BANCO CENTRAL DO BRASIL. Série histórica do sistema financeiro nacional: Operações de crédito do sistema financeiro. 2014. Disponível em: http://www.bcb.gov.br/?SERIESFN. Acesso em: 10/02/2014.

BLUHM, C., OVERBECK, L., & WAGNER, C. (2003). An introduction to credit risk modeling. London: Chapman & Hall.

BROWN, I., & MUES, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications. Vol. 39, pages 3446–3453.

CHUNG, H. Y., HO, C. H., & HSU, C. C. (2011). Support vector machines using Bayesian-based approach in the issue of unbalanced classifications. Expert Systems with Applications. Num. 9, Vol. 38, pages 11447–11452.

DANENAS, P., GARSVA, G., & SAULIUS, G. (2011). Credit Risk Evaluation Model Development Using Support Vector Based Classifiers. Procedia Computer Science. Vol. 4, pages 1699–1707.

DUAN, J. C., & SHRESTHA, K. (2011). Statistical Credit Rating Metholds. Global Credit Review, N. 1. Vol. 1, pages 43-64.

GARCÍA, F., GIMÉNEZ, V., & GUIJARRO, F. (2013). Credit risk management: A multicriteria approach to assess creditworthiness. Mathematical and Computer Modelling, v. 57, p. 2009-2015.

HASTIE, T., TIBSHIRANI, R., &FRIEDMAN, J. (2008). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer. Disponível em: <http://www.stanford. edu/~hastie/local.ftp/Springer/ESLII_print5.pdf>. Acesso em: 21/10/2013

ISCOE, I., KREININ, A., MAUSSER, H., & ROMANKO, O. (2012). Portfolio credit-risk optimization. Journal of Banking & Finance. Vol. 36, nº 6, pages 1604–1615.

JACKS, K. (2008). Capital shocks, bank ass*et al*location, and the revised Basel Accord. Review of Financial Economics. Vol. 17, Issue 2, pages 79-91.

JAMES, G., WITTEN, D., HASTIE, T., & TIBSHIRANI, R. (2013). An introduction to statistical learning: with application in R. New York: Springer.

LOUZADA-NETO, F., & ARA, A. (2012). Bagging k-dependence probabilistic networks: An alternative powerful fraud detection tool. Expert Systems with Applications, v. 39, n. 14, p. 11583-11592.

LOUZADA-NETO, F., FERREIRA-SILVA, P. H., & DINIZ, C. A. R. (2012). On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data Expert Systems with Applications, v. 39, p. 8071–8078.

LU, F. Q., HUANG, M., CHING, W. K., & SIU, T. K. (2013). Credit portfolio management using two-level particle swarm optimization. Information Sciences, Volume 237, Num. 10, pages 162-175.

MATIAS, A. B. (2007). Finanças corporativas de curto prazo. Vol. 1. São Paulo: Atlas.

MCLACHLAN, G. (2001). Multivariate Analysis: Classification and Discrimination. International Encyclopedia of the Social & Behavioral Sciences, pages 10214-10218.

NAGARAJAN, R., SCUTARI, M., & LÈBRE, S. (2013). Bayesian Networks in R with Applications in Systems Biology. Nova York: Springer.

PASCUAL, M. B., MARTÍNEZ, A. M. & ALAMILLOS, A. M. (2013). Redes bayesianas aplicadas a problemas de credit scoring. Una aplicación práctica. Cuadernos de Economía, In Press.

R DEVELOPMENT CORE TEAM. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Disponível em: www.R-project.org

SAITA, F. (2007). Value at Risk and Bank Capital Management: Risk adjusted performance, capital management and capital allocation decision making. San Diego: Elsevier.