ALGUMAS MEDIDAS DO VALOR PREDITIVO DE UM MODELO DE CLASSIFICAÇÃO

Josimara MAZUCHELI[1]
Francisco LOUZADA-NETO[1]
Lorene GUIRADO[1]
Edson Zangiacomi MARTINEZ[2]

▪ RESUMO: Neste artigo apresentamos algumas das medidas comumente utilizadas na avaliação da capacidade preditiva de um modelo de classificação, dentre as quais, a sensibilidade, a especificidade, os valores de predição positivo e negativo e a acurácia. Também descrevemos a curva ROC, que consiste em uma representação gráfica do desempenho de um modelo, de acordo com o conjunto de suas possíveis respostas. Os procedimentos são ilustrados em um conjunto de dados reais considerando um modelo de regressão logística.

▪ PALAVRAS-CHAVE: Modelo de classificação; medidas da capacidade preditiva; sensibilidade; especificidade; valores preditivos positivo e negativo; curva ROC.

1 Introdução

Considere uma situação em que o interesse está centrado na classificação de indivíduos em dois grupos. De portadores ou não de uma determinada característica de interesse, com base na probabilidade predita por um modelo de classificação. Em *credit scoring* a característica de interesse pode ser a inadimplência, sendo que o cliente pode ser classificado como bom ou mau. Em detecção de fraude uma transação efetuada pode ser legítima ou fraudulenta. Em seguros a característica de interesse pode ser a ocorrência ou não de sinistro.

[1]Departamento de Estatística, Universidade Federal de São Carlos – UFSCar, CEP: 13565-905, São Carlos, SP, Brazil. E-mail: *dfln@power.ufscar.br*
[2]Departamento de Medicina Social, Faculdade de Medicina de Ribeirão Preto – FMRP, Universidade de São Paulo – USP, CEP: 14049-900, Ribeirão Preto, SP, Brazil. E-mail: *edson@fmrp.usp.br*

Os modelos de classificação são desenvolvidos a partir de bases históricas do comportamento dos clientes, bem como a partir de bases que contenham informações pertinentes às características cadastrais dos mesmos, tais como, sexo, idade, estado civil dentre outras. A base estatística da modelagem de classificação é o planejamento amostral utilizado para a obtenção da resposta de interesse dentro de um período de desempenho, geralmente, fixado em dias ou meses, de acordo com a necessidade e rapidez, imprescindíveis para a solução do problema. O período de desempenho compreende o intervalo de tempo entre o início do relacionamento cliente-empresa e a observação do desempenho do mesmo, isto é, a sua classificação dicotômica como portador ou não de uma característica de interesse.

Com a classificação dicotômica pontual do cliente e suas características individuais disponíveis, por meio do desenvolvimento de um modelo de classificação, podemos descobrir quais características dos clientes estão relacionadas significativamente com a sua classificação e qual a intensidade e direção desse relacionamento. Existem vários modelos de classificação, dentre os quais podemos citar, análise discriminante, regressão por árvores, regressão logística, regressão logística limitada, redes neurais dentre outros. Maiores detalhes com relação ao desenvolvimento de modelos de classificação em *credit scoring* podem ser encontrados em Abreu (2004).

Atualmente, as bases de dados consideradas para o desenvolvimento de modelos de classificação são muito grandes, e estratégias tradicionais de análises estatísticas podem ficar comprometidas (Louzada-Neto e Diniz, 2002). Neste contexto, procedimentos que medem a capacidade preditiva da modelagem são úteis (Hastie, Tibshirani e Friedman, 2001). A capacidade preditiva de um modelo de classificação é caracterizada pela capacidade do mesmo em classificar corretamente sujeitos como susceptíveis ou não à ocorrência de um determinado fenômeno.

Para efeito de desenvolvimento da modelagem de classificação, um procedimento bastante difundido na prática consiste em subdividir o conjunto de dados em dois conjuntos de dados mutuamente excludentes, usualmente chamados de amostra de treinamento e amostra teste ou *holdout*, respeitando-se a proporção de 70% dos dados para a amostra de treinamento (Thomas, Edelman e Crook, 2002). Esta amostra de treinamento é utilizada para o desenvolvimento da modelagem, enquanto a amostra de teste é utilizada para verificar a capacidade preditiva da mesma. Este procedimento é possível sempre que os conjuntos de dados são muito grandes.

Neste artigo, na Seção 2 apresentamos algumas medidas úteis que são utilizadas para verificar o valor preditivo do modelo. Na Seção 3 descrevemos a curva ROC, que consiste em uma representação gráfica do desempenho de um modelo, de acordo com o conjunto de suas possíveis respostas. Na Seção 4 um modelo de regressão logística, aplicado a um conjunto de dados reais, é avaliado de acordo com as medias apresentadas nas seções anteriores. Na Seção 5 são apresentados alguns comentários finais.

2 Avaliação do modelo

Uma vez construído um modelo de classificação passamos à etapa de avaliação do mesmo, isto é, o quanto o escore produzido pelo modelo consegue distinguir entre bons e maus clientes, uma vez que o objetivo é identificar previamente esses grupos e tratá-los de forma distinta através de diferentes políticas de relacionamento.

A avaliação do modelo é direcionada pela comparação das previsões feitas por ele com a verdadeira condição do cliente, que é geralmente conhecida e está presente, como informação básica, na amostra de teste (Webb, 2002). A prática sugere que a avaliação do modelo na amostra de treinamento, utilizada para o seu desenvolvimento, apresenta resultados melhores do que se avaliado na amostra teste, uma vez que o modelo incorpora peculiaridades inerentes da amostra utilizada para a sua construção (Abreu, 2004). Desta forma, um procedimento sugerido consiste na consideração da amostra de treinamento na avaliação do modelo.

Em princípio, as métricas para avaliação de modelo podem ser vistas como uma adaptação direta da metodologia estatística direcionada para avaliação do desempenho clínico de testes diagnósticos e laboratoriais com respostas dicotomizadas (Martinez e Louzada-Neto, 2000).

Seja $c+d$ o número de clientes bons de uma determinada amostra teste e $a+b$ o número de maus clientes. A partir de um determinado modelo de classificação podemos determinar para cada individuo i, um escore s_i, tal que $0 \leq s_i \leq 1$. Suponha que um indivíduo seja classificado como bom cliente se $s_i > P_c$ e como mau se $s_i \leq P_c$, onde P_c é uma probabilidade denominada ponto de corte (*cut-off*).

Se um bom cliente for classificado como bom ou um mau cliente for classificado como mau, podemos dizer que ele foi classificado corretamente. Fixando-se um ponto de corte podemos construir a matriz de confusão dada pela Tabela 1.

Tabela 1 - Matriz de confusão

Resultado	Real		
do modelo	Positivo ($D+$)	Negativo (D-)	Total
Positivo ($T+$)	*a (VP)*	*b (FP)*	*a + b*
Negativo (T-)	*c (FN)*	*d (VN)*	*c + d*
Total	*a + c*	*b +d*	*a + b + c + d*

Na Tabela 1, temos que:

- (a) representa o número de maus clientes, classificados corretamente como maus, isto é, verdadeiros positivos (VP);
- (b) representa o número de maus clientes, classificados incorretamente como bons, portanto revela o número de resultados falso-positivos (FP) que o método detectou;
- (c) representa o número de clientes bons, classificados incorretamente como maus, ou seja, o número de resultados falso-negativos (FN);

- (d) representa o número de clientes bons, classificados corretamente como bons, isto é, o número de resultados verdadeiramente negativos (VN);
- $(a + c)$ - representa o número total de maus clientes;
- $(b + d)$ - representa o número total de bons clientes;
- $a + b$ - representa o número de clientes identificados pelo modelo como maus;
- $(c + d)$ - representa o número de clientes identificados pelo modelo como bons.

A capacidade preditiva um modelo está relacionada com suas medidas de desempenho que podem ser calculadas a partir da Tabela 1, dentre as quais podemos citar: a sensibilidade, a especificidade, os valores de predição positivo e negativo, e a acurácia. Estas medidas são definidas a seguir juntamente com a importante definição de prevalência.

Definição 1 (Prevalência): *a prevalência de uma característica é definida como a proporção de clientes em uma população que são propensos à mesma. Dessa maneira, ela expressa a probabilidade do indivíduo ser, por exemplo, um mau cliente, antes de o modelo ser ajustado. A prevalência é também denominada probabilidade pré-modelagem e é dada por,*

$$p = P(D_+) = \frac{a + c}{a + b + c + d}. \tag{1}$$

Definição 2 (Sensibilidade): *a sensibilidade é definida como a proporção de verdadeiros positivos entre todos os portadores da característica, e expressa a probabilidade do modelo sob investigação fornecer um resultado positivo dado que o indivíduo tem a característica. Isto é, a sensibilidade corresponde a proporção de maus clientes que são classificados corretamente através de um modelo qualquer por ter um escore superior a P_c. A sensibilidade do modelo é também conhecida como capacidade de acerto de positivos ou maus clientes. Considerando a Tabela 1, a sensibilidade é dada por,*

$$S = P(T_+|I_+) = \frac{VP}{VP + FN} = \frac{a}{a + c}. \tag{2}$$

Definição 3 (Especificidade): *a especificidade é definida como a proporção de verdadeiros negativos entre todos os não portadores da característica, e expressa a probabilidade do modelo sob investigação fornecer um resultado negativo dado que o indivíduo está livre da característica. Isto é, a especificidade corresponde a proporção de clientes bons, classificados corretamente através de um modelo qualquer por terem escore menor ou igual à P_c. A especificidade do modelo é também conhecida como capacidade de acerto de negativos ou bons clientes. Considerando a Tabela 1, a especificidade é dada por,*

$$E = P(T_-|D_-) = \frac{VN}{VN + FP} = \frac{d}{b + d}. \tag{3}$$

Observe que um modelo sensível raramente deixará de diagnosticar indivíduos portadores da característica e que um modelo muito específico raramente classificará como mau cliente um indivíduo sem esta característica.

Uma outra medida da capacidade preditiva do modelo é dada pela acurácia.

Definição 4 (Acurácia): *a acurácia é definida como a proporção de acertos de um modelo, tanto positivos quanto negativos, ou seja, é a proporção de verdadeiros positivos e verdadeiros negativos em relação a todos os resultados possíveis. A acurácia é também denominada capacidade total de acerto do modelo* (CTA). *Considerando a Tabela 1, acurácia é dada por,*

$$CTA = \frac{VP + VN}{VP + FP + VN + FN} = \frac{a + d}{a + b + c + d}. \tag{4}$$

Observe que a acurácia também pode ser vista como a média ponderada da sensibilidade e especificidade em relação ao número de clientes maus e clientes bons de uma população. Ressaltamos que a acurácia não é a melhor medida para orientar a escolha de um modelo, pois é influenciada pela sensibilidade, especificidade e prevalência. Além disso, dois modelos com sensibilidades e especificidades diferentes podem fornecer valores semelhantes de acurácia se forem aplicados a populações com prevalências muito diferentes. Muitas vezes a escolha de um modelo privilegia grande sensibilidade ou grande especificidade, dependendo dos objetivos de uso do mesmo.

Considerando a Tabela 1, quando do resultado de um modelo, clientes maus são classificados como bons, os respectivos resultados errôneos deste modelo são denominados de falsos-negativos (FN); quando clientes bons são classificados como maus, os resultados errôneos deste modelo são denominados falsos-positivos (FP).

Uma outra medida de interesse é a probabilidade da presença da característica após o resultado do modelo, ou seja, qual o valor preditivo positivo do modelo ou a probabilidade do cliente ser mau dado que o modelo o classificou como positivo para a característica em estudo. Por outro lado, também de interesse é a probabilidade da não presença da característica após o resultado do modelo, ou seja, qual o valor preditivo negativo do modelo ou a probabilidade do cliente ser originalmente bom dado que o modelo o classificou como negativo para a característica em estudo. Estas medidas são definidas a seguir.

Definição 5 (Valor Preditivo Positivo): *o valor preditivo positivo* (VPP) *do modelo é definido como a proporção de verdadeiros positivos entre todos os indivíduos classificados pelo modelo como positivo para a característica em estudo, ou seja, é a proporção de indivíduos maus, dado que o modelo os apontou como positivos. Considerando a Tabela 1 o* VPP *é dado por,*

$$VPP = P(D_+|T_+) = \frac{VP}{VP + FP} = \frac{a}{a + b}. \tag{5}$$

Definição 6 (Valor Preditivo Negativo): *o valor preditivo negativo* (VPN) *do modelo pode ser visto como a proporção de verdadeiros negativos entre todos*

os indivíduos classificados pelo modelo como negativo, ou seja, é definido como a proporção de indíviduos bons, dado que o modelo os apontou como negativos para a característica em estudo. Considerando a Tabela 1 o VPN é dado por,

$$VPN = P(D_-|T_-) = \frac{VN}{VN + FN} = \frac{d}{c + d}. \quad (6)$$

Quanto mais sensível o modelo, maior seu valor preditivo negativo, isto é, maior é a segurança de que um cliente com resultado negativo não seja mau, e quanto mais específico o modelo, maior é o seu valor preditivo positivo, ou seja, maior é a segurança de que um indivíduo com resultado positivo seja originalmente mau.

3 A curva ROC

O valor do ponto de corte P_c escolhido influencia na sensibilidade e na especificidade do modelo. Quanto maior o ponto de corte, maior a especificidade do modelo e menor a sua sensibilidade, e quanto menor o ponto de corte, maior a sensibilidade e menor a especificidade.

Neste contexto, a curva ROC (Zweig e Campbell, 1993) pode ser utilizada. Esta curva é construída variando os pontos de corte ao longo da amplitude dos escores fornecidos pelo modelo, a fim de se obter diferentes classificações para os clientes. Para cada ponto de corte P_c obtemos os respectivos valores para as medidas de sensibilidade e especificidade. Assim, a curva ROC é construída tendo no seu eixo horizontal os valores de $(1 - E)$, ou seja, a proporção de clientes bons que são classificados como maus pelo modelo e no eixo vertical os valores de S, isto é, a proporção de clientes maus que são classificados como maus. Uma curva ROC obtida ao longo da diagonal principal corresponde a uma classificação obtida sem a utilização de qualquer ferramenta preditiva, ou seja, sem a utilização de modelos. Conseqüentemente, a curva ROC deve ser interpretada de forma que quanto mais a curva estiver distante da diagonal principal melhor é o desempenho do modelo associado a ela. Esse fato sugere que quanto maior for a área entre a curva ROC produzida e a diagonal principal, melhor é o desempenho global do modelo, como podemos observar na Figura 1 (painel esquerdo). Uma vantagem da curva ROC está em sua simplicidade. Consiste em uma representação direta do desempenho de um modelo, de acordo com o conjunto de suas possíveis respostas.

4 Dados reais

Nesta seção descrevemos a aplicação dos procedimentos apresentados nas seções anteriores em um conjunto de dados reais. Os dados correspondem à classificação de clientes de uma carteira de um banco como bons ou maus, de acordo com seus desempenhos de crédito. As variáveis consideradas foram: tipo de cliente, tempo de emprego, sexo, idade, estado civil, limite de crédito, tempo de residência e profissão. Um modelo de regressão logística usual (Hosmer e Lemeshow, 1989), foi

ajustado a uma amostra de treinamento correspondente a 70% da amostra original. Os restantes 30% dos dados foram utilizados como amostra de teste para verificação da adequabilidade do modelo.

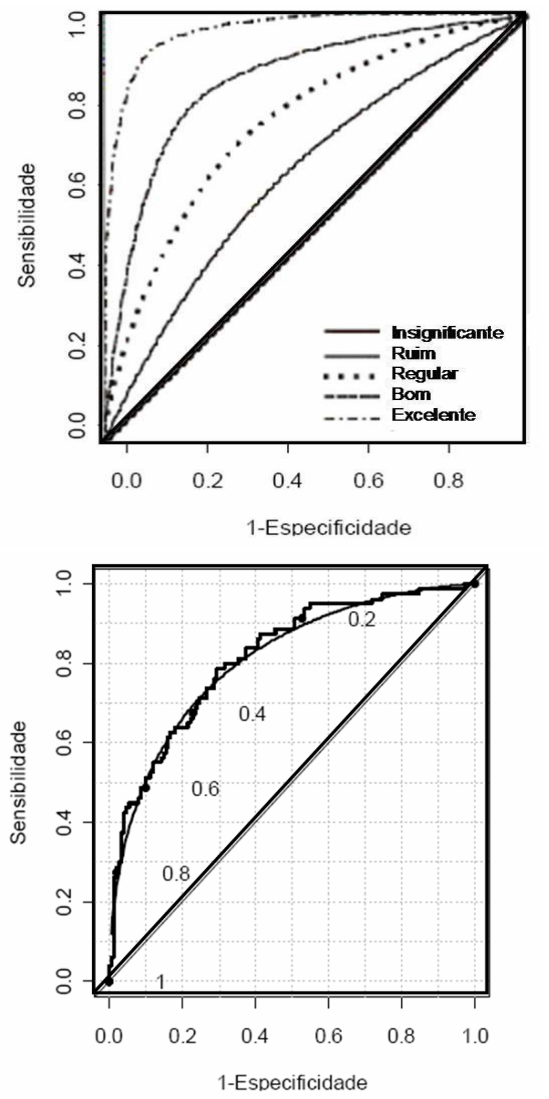

Figura 1 - Curvas ROC: Painel Superior - diferentes curvas ROC para verificação da capacidade do modelo. Painel Inferior - Curva ROC com ajuste do modelo de regressão logística.

Ajustado o modelo a partir da amostra de treinamento, sua capacidade de predição foi verificada na amostra de teste. Com auxílio da curva ROC (Figura 1

- painel direito) escolhemos um ponto de corte igual a $0,33$ e com isso as medidas relacionadas a capacidade preditiva do modelo foram: $CTA = 0,76$, $S = 0,79$, $E = 0,71$, $VPP = 0,59$ e $VPN = 0,86$, que formam um indicativo de uma boa capacidade preditiva. Esta conclusão é corroborada pela curva ROC apresentada na Figura 1 (painel direito).

5 Comentários finais

Em nossa experiência, a melhor estratégia para verificar o valor preditivo da modelagem consiste em utilizar conjuntamente todas as medidas de desempenho apresentadas neste artigo.

Uma outra maneira de avaliar o desempenho de um modelo é através da razão de verossimilhanças (Martinez e Louzada-Neto, 2000), que deverá ser considerada em estudos futuros. Esta razão é definida como a razão entre a probabilidade de um determinado resultado de um modelo em indivíduos portadores da característica estudada (maus) e a probabilidade do mesmo resultado em indivíduos sem a característica (bons). Ao contrário dos valores preditivos, a razão de verossimilhanças não sofre influência da prevalência da característica na população em estudo.

Embora os procedimentos apresentados tenham sido ilustrados quando do ajuste de um modelo de regressão logística, em princípio, tais procedimentos podem ser diretamente aplicados quando do ajuste de outros modelos de classificação.

Agradecimentos

Os autores agradecem os comentários e sugestões dos revisores. Francisco Louzada-Neto é parcialmente financiado pelo CNPq por meio de uma Bolsa de Pesquisa.

MAZUCHELI, J.; LOUZADA-NETO, F.; GUIRADO, L.; MARTINEZ, E. Z. Some measures for the evaluation of the predicitive capacity of a classification model. *Rev. Bras. Biom.,* São Paulo, v.26, n.2, p.83-91, 2008.

- *ABSTRACT: In this article we present some of the usual measures considered in the evaluation of the predicitve capacity of a classification model, among the which, the sensibility, the specificity, the negative and positive predictive values and to acuracy. Also we describe the ROC curve, that consists of a graphic representation of the performance of a model, according to the assembly of its possible answers. The procedures are illustrated in an real data set considering a logistics regression model.*

- *KEYWORDS: Classification model; predictive capacity measure; sensibility; specificity; negative and positive predictive values; ROC curve.*

Referências

ABREU, H. J. *Aplicação de análise de sobrevivência em um problema de credit scoring e comparação com a regressão logística.*, 2004. Dissertação (Mestrado em Estatística), Universidade Federal de São Carlos, São Carlos, 2004.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning*: data mining, inference and prediction. Spring. 2001.

HOSMER, W. D.; LEMESHOW, S. *Applied logistic regression.* Wiley. 1989.

MARTINEZ, E. Z.; LOUZADA-NETO, F. Metodologia estatística para testes diagnósticos e laboratoriais com respostas dicotomizadas. *Rev. Mat. Estat.*, São Paulo, v.18 , p.83-101, 2000.

MEDRONHO, R. A. *Epidemiologia.* Atheneu. 2002.

THOMAS, L. C.; EDELMAN D. B.; CROOK J. N. *Credit scoring and its applications.* SIAM. 2002.

ZWEIG, M. H.; CAMPBELL, G. Receiver-operating characteristic (ROC) plots. *Clin. Chem.*, Baltimore, v.29, p.561-577, 1993.

WEBB, A. *Statistical pattern pecognition*, 2a. ed., Wiley. 2002.

Recebido em 02.05.2007.

Aprovado após revisão em 20.06.2008.