XXI Congresso de Iniciação Científica da Unicamp

Um estudo comparativo sobre a utilização de modelos LOGIT e PROBIT na previsão de insolvência

Maria Clara Girotto de Faria; Ivette Luna
Palavras-chave: Regressão Logística – Modelo Probabilístico – Insolvência
CNPQ/PIBIC

Introdução

O crédito constitui um elemento indispensável para o desenvolvimento econômico. Associado à concessão de recursos tem-se o risco de crédito que está presente em praticamente todas as operações bancárias, e está relacionado a possíveis perdas quando um dos contratantes não honra com seus compromissos.

A fim de minimizar essas perdas e fornecer maior suporte à tomada de decisão, percebe-se a necessidade de se utilizar a econometria e modelos preditivos associados a uma prévia análise do perfil do tomador de crédito.

Modelos utilizados

O modelo LOGIT pode ser descrito segundo a equação:

$$\mathrm{Pi} = \mathrm{E}(\mathrm{Y} = 1|\mathrm{Xi}) = \frac{1}{1+\mathrm{e}^{(\beta 1+\beta 2\mathrm{Xi})}}$$

Onde:

- ✓ Pi indica a probabilidade de possuir algo;
- ✓ β1 corresponde ao intercepto, ou seja, uma constante e;
- ✓ β2 indica o valor dos coeficientes relacionados às variáveis relevantes na tomada de decisão.

De forma a comparar o desempenho obtido na previsão de solvência dos consumidores analisados, foi utilizado também o modelo PROBIT:

$$\mathrm{Pi} = \Phi\left(\mathrm{Xi}\frac{\beta}{\sigma}\right)$$

Onde:

- ✓ Φ é a função distribuição acumulada da distribuição normal;
- ✓ β é o vetor de coeficientes estimados da função e;
- ✓ Xi é a matriz (n,k), em que n representa o número de observações e k os atributos característicos das observações.

Caracterização das Variáveis

A base de dados utilizada nesse projeto foi retirada do repositório *UCI Machine Learning*, sendo denominada *German Credit Data Set* [1]. Consiste em uma base de dados multivariados, subdivididos em dados numéricos (total de 7) e categóricos (total de 13), totalizando 20 atributos, representando um total de 1000 perfis de consumidores.

Vale citar que cada padrão está associado a um perfil de consumidor, estando este perfil relacionado a uma classificação (adimplente ou inadimplente), dependendo dos atributos que compõem tal padrão.

Das 20 variáveis iniciais foram consideradas 11 na estimação – duração da conta (meses), montante, porcentagem do salário utilizado para quitar dívidas, status e sexo, histórico de pagamentos, idade, propósito na tomada de crédito, outros_finan (outros financiamentos já realizados), outras_div (se existem outras dívidas), telefone e estrangeiro (se é ou não).

Metodologia

Como explicado durante a caracterização das variáveis, das 20 que inicialmente compunham a base de dados, apenas 11 foram escolhidas para o modelo final, em razão da significância que as mesmas apresentaram.

A fim de que o modelo não se mostrasse tendencioso, 70% dos dados foram utilizados para a estimação dos parâmetros do modelo e o restante, 300 padrões, para a validação.

O ajuste dos parâmetros dos modelos foi realizado via método dos Mínimos Quadrados Ordinários (MQO), através do software Eviews.

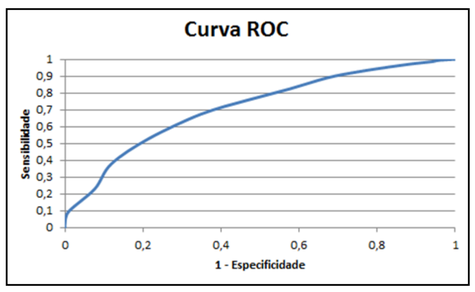
Para fins de verificação construiu-se uma matriz de confusão com um *cut-off* de 0,5, visto a necessidade em determinar a sensibilidade e especificidade do modelo. Além disso, para validação escolheu-se a curva ROC, que é uma medida da capacidade de discriminação do modelo. Vale citar que o cálculo da área sob a curva foi feito de maneira aproximada, através da Lei dos Trapézios.

Resultados

Matriz de confusão para o Logit

	Estimated Equation		
	Dep=0	Dep=1	Total
P(Dep=1)<=C	52	36	88
P(Dep=1)>C	158	454	612
Total	210	490	700
Correct	52	454	506
% Correct	24.76	92.65	72.29
% Incorrect	75.24	7.35	27.71
Total Gain*	24.76	-7.35	2.29
Percent Gain**	24.76	NA	7.62

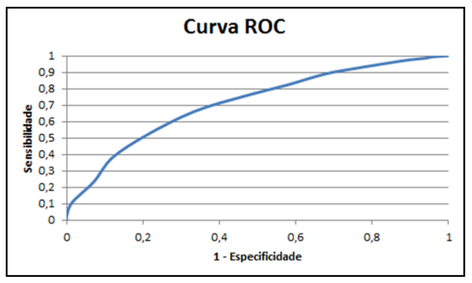
Curva ROC para o Logit



Matriz de confusão para o Probit

	Estimated Equation		
	Dep=0	Dep=1	Total
P(Dep=1)<=C	51	37	88
P(Dep=1)>C	159	453	612
Total	210	490	700
Correct	51	453	504
% Correct	24.29	92.45	72.00
% Incorrect	75.71	7.55	28.00
Total Gain*	24.29	-7.55	2.00
Percent Gain**	24.29	NA	6.67

Curva ROC para o Probit



Para o modelo LOGIT, o valor encontrado para a área sob a curva ROC foi de 0,7144 e no PROBIT 0,7135, o que significa que o modelo possui uma boa capacidade preditiva.

Conclusões

Através dos resultados obtidos foi possível perceber que a capacidade preditiva dos modelos analisados é bem similar, inclusive por serem, no que tange à teoria, bem próximos. Embora apresentem resultados satisfatórios, foi possível observar que o desempenho fica aquém do esperado, quando comparado com metodologia similar. Possível causas levantadas remetem ao tratamento empregado aos dados e às simplificações impostas durante os cálculos.

Bibliografia

[1] HOFMANN, H. Statlog (German Credit Data Set). Universit”at Hamburg. 17 Nov. 1994 [acesso em Outubro de 2012].