RESUMO

O trabalho reporta-se a uma tese de doutoramento realizada no ISCTE sobre o tema Credit Scoring.

O objetivo desta investigação foi determinar as principais características dos utilizadores de cartões de crédito que pudessem antecipar a classificação de bom ou mau pagador.

Foi selecionado um modelo estatístico (Regressão Logística Binária) para testar uma base de dados de 4 000 utilizadores de cartões de crédito, e verificar a robustez preditiva do modelo.

Após a construção do modelo e a sua aplicação à base de dados verificou-se que o modelo conseguia acertar em mais de 90% dos casos.

Autor: António Manuel Sarmento Batista

Cédula Profissional: 3346

1 - INTRODUÇÃO

O trabalho que vou apresentar baseia-se num caso prático, que utiliza o modelo de Regressão Logística Binária, que é um algoritmo clássico de Machine Learning.

Machine Learning é um subconjunto de ***Inteligência Artificial*** (IA) inserindo-se o presente trabalho numa das 3 temáticas propostas.

O tema é um breve resumo da minha tese de doutoramento realizada em 2009 no ISCTE – INSTITUTO UNIVERSITÁRIO DE LISBOA com o título «CREDIT SCORING – Uma metodologia de gestão para a prevenção e redução do crédito malparado».

A motivação que inspirou este trabalho foi a constatação do aumento de crédito malparado em Portugal, que as instituições financeiras registavam ao longo de diversos anos. Este problema já há muito que tinha sido reconhecido e consagrado pelos Acordos de Basileia que entre várias recomendações sugeriu aos Bancos formas mais rigorosas de controlar o risco de crédito.

Para mitigar este problema foram propostas diversas práticas, entre elas a quantificação probabilística do incumprimento traduzida por uma pontuação de risco, cuja identificação na gíria do discurso financeiro se designa por *Scoring* ou *Credit Scoring*.

Neste contexto, o objetivo deste estudo foi identificar fatores explicativos capazes de prever a probabilidade de um devedor ser no futuro um Bom ou Mau pagador e avaliar a robustez preditiva do modelo utilizado para este efeito.

O projeto de investigação incidiu sobre o crédito ao consumo tendo a identificação daqueles fatores explicativos sido feita através da utilização de uma base de dados de 4 000 utilizadores de cartões de crédito, cujos hábitos de pagamento se conhecem *a priori*.

A metodologia de investigação empírica consistiu na aplicação do modelo de Regressão Logística Binária aos dados em análise, por ser especialmente adequado ao estudo em causa e devido à sua simplicidade.

A identificação dos fatores explicativos (atributos) mais relevantes foi realizada através do método iterativo *forward stepwise* (Likelihood Ratio) e que consiste em selecionar entre as variáveis independentes aquelas cuja capacidade preditiva do comportamento de Bom ou Mau pagador são estatisticamente significativas.

EMPRÉSTIMOS AO COSUMO EM PORTUGAL			
	Empréstimo	**Cob.Duvidosa**	**%**
dez/04	9 059 000 000	454 000 000	5,01%
dez/05	9 406 000 000	292 000 000	3,10%
dez/06	11 379 000 000	369 000 000	3,24%
dez/07	13 790 000 000	505 000 000	3,66%
dez/08	15 452 000 000	759 000 000	4,91%
dez/09	15 731 000 000	1 032 000 000	6,56%
dez/10	15 484 000 000	1 237 000 000	7,99%
dez/11	14 987 000 000	1 477 000 000	9,86%
dez/12	13 371 000 000	1 580 000 000	11,82%
dez/13	12 075 000 000	1 407 000 000	11,65%
dez/14	12 099 000 000	1 297 000 000	10,72%
dez/15	12 183 000 000	1 123 000 000	9,22%
dez/16	12 375 000 000	790 000 000	6,38%
dez/17	13 857 000 000	659 000 000	4,76%
dez/18	15 311 000 000	563 000 000	3,68%
dez/19	19 226 000 000	811 000 000	4,22%
dez/20	19 166 000 000	?	#VALOR!
dez/21	19 210 000 000	?	#VALOR!
dez/22	20 666 000 000	?	#VALOR!

Fonte: Boletim Estatístico do BdP (B.4.1.4)

2 - METODOLOGIA

A metodologia seguida é segmentada em 6 momentos:

O primeiro momento caracterizou-se pela obtenção de uma amostra que contivesse o mesmo número de bons e maus pagadores. Assim, dos 4 000 utilizadores de cartões de crédito, 2 000 tinham um perfil de bom pagador e os outros 2 000 um perfil de mau pagador.

O segundo momento baseou-se em dividir a amostra dos 4 000 utilizadores de cartões de crédito em duas subamostras: uma com 3 200 titulares (1 600 bons e 1 600 maus) e outra amostra com 800 (400 bons e 400 maus). A primeira amostra denominou-se por *in-sample* e que contem 80% dos utentes, e foi utilizada para estimar os parâmetros do modelo logístico. A segunda amostra, designada por *out-of-sample* ou *holdout sample*, contendo 20% dos utilizadores, serviu para validar fora da amostra os resultados obtidos pela primeira.

O terceiro momento consistiu em estabelecer critérios e métodos que guiassem toda a investigação empírica, Assim, criou-se a variável dependente dicotómica que se designou por COD que assume o valor 1 quando o modelo prevê uma conta boa e o valor 0 quando prevê uma conta má.

No quarto momento traduziu-se pela caracterização da amostra *in-sample*, recorrendo-se às medidas de estatística descritiva para obter mais conhecimento sobre os utilizadores que fizeram parte da amostra, nomeadamente as características pessoais (idade, género, estado civil, ocupação, etc.) e outras de índole comportamental, fruto da experiência que a instituição de crédito tinha já construído sobre o indivíduo (pontualidade nos pagamentos, incidentes, saldo médio, entre muitas outras).

No quinto momento estimou-se o modelo de regressão logística recorrendo aos programas estatísticos E-Views, SAS e SPSS, obtendo-se a informação sobre a qualidade de previsão do modelo.

No sexto momento efetuou-se a validação estatística do modelo através da *holdout sample* que confirmou a robustez preditiva do modelo previamente estimado.

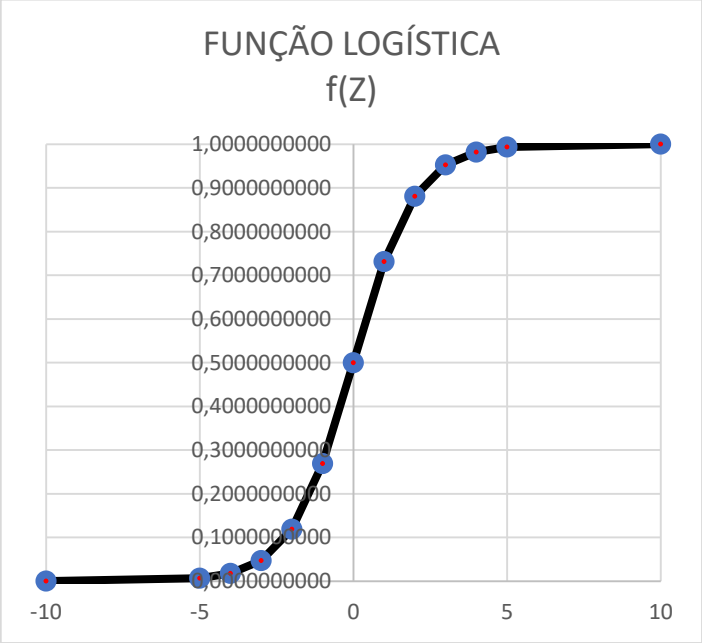
3- REGRESSÃO LOGÍSTICA

A regressão logística é uma técnica estatística que produz a partir de um conjunto de informações, um modelo que permite a predição de valores tomados por uma variável categórica binária, a partir de um conjunto de variáveis independentes (ou explicativas).

A popularidade da regressão logística fundamenta-se na função logística que descreve a forma matemática, na qual o modelo logístico se baseia. Esta função traduz-se pela equação:

$$f(z) = \frac{1}{1 + e^{-z}}$$

Para se obter o traçado da função logística foram dados valores a z para se obter $f(z)$



O modelo logístico nunca poderá obter uma probabilidade de risco superior a 100% ou inferior a 0%. Assim, a probabilidade de incumprimento situa-se entre 0 e 1. A forma de um alongado S descreve o efeito combinado de vários fatores de risco de incumprimento. Os programas de regressão logística preveem o grupo dicotómico da variável dependente, baseado num valor típico de corte de 0,5 ou superior são classificados por 1 e aqueles casos com probabilidades abaixo de 0,5 são classificados com 0.

O modelo logístico foi obtido a partir da função logística, traduzindo z como uma soma linear de z:

$$Z_i = \alpha_0 + \beta_1 X_{1_i} + \cdots + \beta_n X_{n_i}$$

Em que:

α_0 – É o parâmetro do modelo designado por CONSTANTE (porque não depende de X);

β_n – É o parâmetro do modelo designado por COEFICIENTE DA VARIÁVEL X;

X_{n_i} - É a VARIÁVEL INDEPENDENTE.

Substituindo Z na função logística, obtém-se:

$$f(z) = \frac{1}{1 + e^{-(\alpha_0 + \beta_1 X_{1_i} + \cdots + \beta_n X_{n_i})}}$$

4 – ANÁLISE DE DADOS

De entre as características referentes aos 4 000 utentes de cartões de crédito, identificaram-se 21 que devem constituir (total ou parcialmente) as variáveis independentes ou explicativas do modelo de regressão logística.

	Variável Independente
1	Scoring (Pontuação)
2	Limite de Crédito
3	Saldo Atual
4	Código do Estado da Conta
5	Código da Classificação da Conta
6	Revolving
7	Rendibilidade Mensal da Conta
8	Ano e Mês (a que respeitam os dados)
9	Descrição do Código de Classificação
10	Descrição do Código do Estado da Conta

	Variável Independente
11	Delinquente (> 60 dias)
12	Delinquente (> 90 dias)
13	Género (M/F)
14	Código Postal
15	Data de Nascimento
16	Estado Civil
17	Habilitações Literárias
18	Região Geográfica
19	Profissão
20	Idade
21	Rendimento

Algumas destas variáveis não necessitam de quaisquer explicações prévias (idade, género, rendimento, etc.) embora outras, pelo seu conteúdo mais específico, requeiram um esclarecimento preliminar, nomeadamente:

- Scoring: Corresponde a uma determinada pontuação que permite posicionar o utilizador num referencial cartesiano, em que nas abcissas é mostrada a sua pontuação (score) e nas ordenadas o número de utilizadores que constituem a carteira;

- Limite de crédito: É o montante máximo que o utente pode utilizar com o seu cartão de crédito. Esta variável é calculada com base em determinados critérios dos quais se destacam o rendimento, o número de anos no atual emprego, estado civil, responsabilidades em outras instituições de crédito, etc.;

- Saldo atual: É o valor da dívida que o utente tem no momento final do ciclo de faturação;

- Estado da conta: É identificado por um código numérico;

- Código da Classificação da conta: Esta identificação é uma combinação com outras classificações;

- Revolving: É o valor que ficou por pagar, no extrato do mês anterior.

- Rendibilidade: Mede o desempenho da conta;

- Delinquência: Delito no atraso de pagamento

As restantes variáveis identificam datas, códigos de classificação, códigos postais.

- AGRUPAMENTO DAS VARIÁVEIS SEGUNDO A SUA NATUREZA

Variáveis Qualitativas	**Variáveis Quantitativas**	**Variáveis "Data"**
Código do Estado da Conta (4)	Scoring (1)	Data Nascimento (15)
Descrição do Código do Estado da Conta (10)	Limite de Crédito (2)	Ano e Mês (a que respeitam os dados) (8)
Descrição do Código de Classificação (9)	Saldo Atual (3)	
Código da Classificação da Conta (5)	Revolving (6)	
Delinquente (>60 dias) (11)	Rendibilidade Mensal da Conta (7)	
Delinquente (>90 dias) (12)	Idade (20)	
Género (M/F) (13)	Rendimento (21)	
Código Postal (14)		
Estado Civil (16)		
Habilitações Literárias (17)		
Região Geográfica (18)		
Profissão (19)		

5 – APLICAÇÃO DO MODELO AOS DADOS *IN-SAMPLE*

Como ponto de partida foram incluídas no modelo todas as 21 variáveis explicativas inicialmente consideradas. Em etapas seguintes foram excluídas 11, cujos coeficientes não se revelaram estatisticamente significativos para um nível de significância de 10%. Apresentam-se a seguir os resultados da estimação após 23 iterações, indicando aquelas que permaneceram no modelo.

Dependent Variable: COD
Method: ML-Binary Logit (Quadratic hill climbing)
Sample: 1 3200
Included observations: 3200
Convergence achieved after 23 iterations
Covariance matrix computed using second derivatives

Variable	Coeficient	Std Error	z-Statistic	Pro
Scoring	0.066500	0.003795	17.52427	0.0000
Limite Crédito	0.000261	2.05E-05	12.74539	0.0000
Saldo Atual	0.000499	0.000110	4.526730	0.0000
Rendibilidade	-0.008396	0.002934	-2.861424	0.0042
Idade	0.018678	0.007091	2.634029	0.0084
Habil. (Dummy)	-0.602927	0.132102	-4.564096	0.0000
Class. (Dunmmy)	-1.434070	0.349123	-4.107634	0.0000
Revolving	-0.000384	0.000120	-3.201036	0.0014
Est.Civ. (Dummy)	-0.578892	0.134513	-4.303619	0.0000
Género (Dummy)	0.808648	0.144411	5.599634	0.0000
Constante	-49.89290	2.752544	-18.12610	0.0000

McFadden R-squared	0.653876	Mean dependente var	0.500000
S.D. dependent.var	0.500078	S.E. regression	0.260019
Alcaike info criterion	0.486704	Sum squared resid	215.6077
Schwarz criterion	0.507573	Log likelihood	-767.7269
Hannan-Quinn crit.	0,494186	Restr.log likelihood	-2218.071
LR statistic	2900.688	Avg.log likelihood	-0.239915
Prob(LR statistic)	0.000000		

Obs with Dep=0	1600	Total obs	3200
Obs with Dep=1	1600		

A tabela anterior dá-nos também informação sobre a contribuição ou importância de cada uma das variáveis explicativas. As estimativas para os coeficientes β (2ª coluna da tabela) são os valores que foram utilizados na função de distribuição logística para estimar a probabilidade de um cliente ser bom pagador.

Ilustra-se a seguir o *output* do E-Views referentes aos dados da amostra *in-sample* mostrando uma taxa de acerto geral de 90,84%.

Expectation-Prediction Evaluation for Binary Specification
Success cutoff C=0,5

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	1433	126	1559	1600	1600	3200
P(Dep=1)>C	167	1474	1641	0	0	0
Total	1600	1600	3200	1600	1600	3200
Correct	1433	1474	2907	1600	0	1600
%Correct	89.56	92.13	90,84	100.00	0.00	50.00
%Incorrect	10.44	7.87	9,16	0.00	100.00	50.00
Total Gain	-10.44	92.13	81.69			

A informação dada na tabela anterior pode ser resumida na tabela seguinte:

Classification Table

			Predicted		
			COD		
			0	1	% Correct
Observed	COD	0	1433	126	89,56
		1	167	1474	92,13
		Overall Per.			90,84
The cut value is 0,5					

A validação do modelo foi efetuada através de uma amostra (*out-of-sample*) contendo 800 utentes (400 bons e 400 maus) e aplicada ao modelo de regressão logística de acordo com a seguinte equação:

$$p(Y = 1) = \frac{1}{1 + e^{-Z_i}}$$

Em que: $Z_i = \alpha_0 + \beta_1 X_{1_i} + \cdots + \beta_n X_{n_i}$

Os valores obtidos estão resumidos na tabela seguinte, o que comprova a robustez preditiva do modelo:

In-sample				
		Predicted		
Observed		COD		
		0	1	% Corect
COD	0	1433	126	**89,56**
	1	167	1474	**92,13**
Overall %				**90,84**

Out-of-sample				
		Predicted		
Observed		COD		
		0	1	% Correct
COD	0	356	43	**89,00**
	1	27	374	**93,50**
Overall %				**91,25**

6- CONCLUSÕES

Da similitude dos resultados obtidos nas duas amostras (*in sample* e *out-of-sample*) demonstrou-se que o modelo oferece uma considerável robustez preditiva, permitindo antecipar o resultado Bom ou Mau e, desta forma, prever e reduzir os créditos de cobrança duvidosa.