Autorização automática de débito em contas com saldo insuficiente

Artigo

A Tripartite Scorecard for the Pay/No pay Decision-Making in the Retail Banking Industry

de

Maria Rocha Sousa e Joaquim Pinto da Costa

Analisado por Hélia Monteiro da Costa

Seminário de Modelação 2011/2012

18 de janeiro de 2012

Agenda

Sumário

Recolha de dados

Minimização das perdas

Modelo binário

Modelo ternário

Discussão

Bibliografia

Sumário

Enquadramento:

Os bancos suportam a decisão de crédito em modelos que preveem o incumprimento a 6 ou mais meses.

O débito ou não de uma transação bancária numa conta com saldo insuficiente deveria ser decidido com base na probabilidade de o cliente incumprir a 30 dias.

Objetivo:

Criar modelo para decidir se é efetuado ou não o débito de uma transação numa conta bancária com saldo insuficiente, minimizando as perdas.

Processo:

1. Modelos binários: usar vários modelos para distribuir os clientes por duas classes, MAU e BOM, mediante a probabilidade de incumprirem.
2. Modelos ternários: usar vários modelos para distribuir os clientes por três classes, MAU, BOM e para REVISÃO (decisão manual).

Resultado:

Obtida uma automatização de 87%, que compara favoravelmente com 79% do processo anterior.

Agenda

Sumário

Recolha de dados

Minimização das perdas

Modelo binário

Modelo ternário

Discussão

Bibliografia

Recolha de dados

Janela de observação:

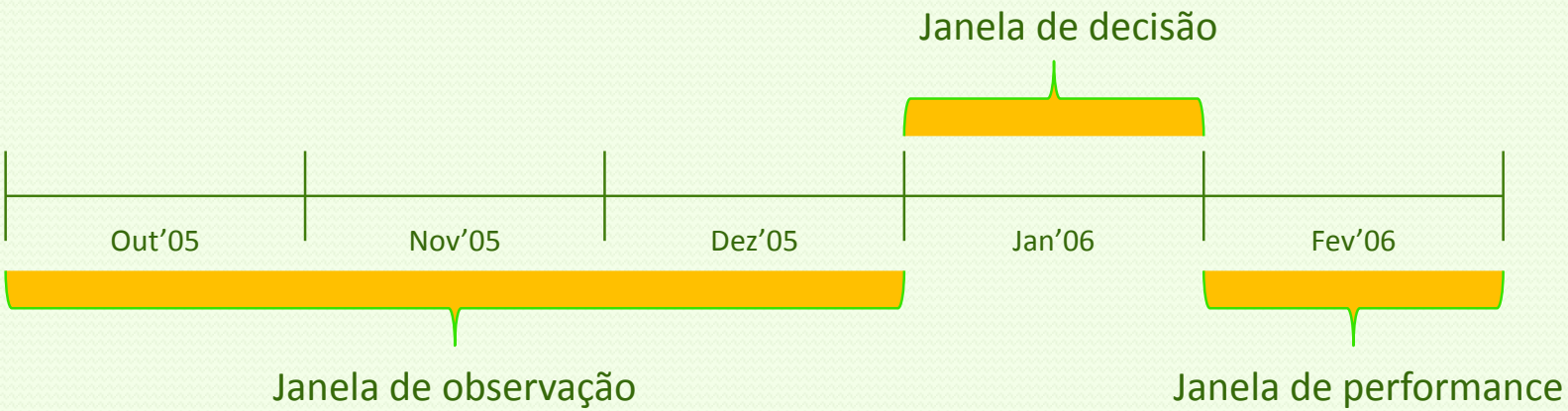
Período associado à informação histórica que permitirá caracterizar o cliente.

Janela de decisão:

Período associado à decisão de débito ou não de uma transação na conta com saldo insuficiente.

Janela de performance:

Período durante o qual será avaliada a capacidade de o cliente regularizar a conta a descoberto em 30 dias.

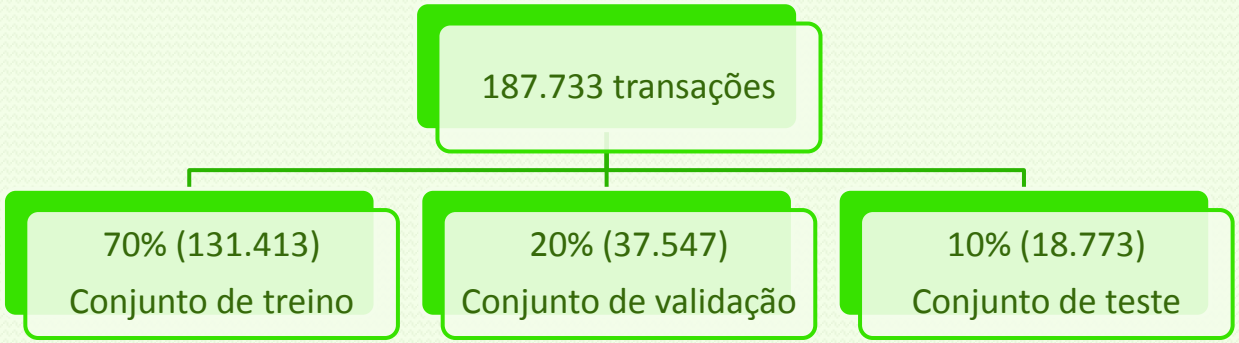


Recolha de dados

Atendendo às janelas definidas e usando o software *SAS Enterprise Miner* identificaram-se:

- 187.733 transacções cujo lançamento nas contas provocaria um descoberto e que requerem decisão de aprovação ou não de débito em conta.
- 47 características/atributos dos clientes e contas associadas às transacções acima referidas.
- 82% das transacções eram de clientes que conseguiam regularizar as contas em menos de 30 dias.

Para a criação e validação de modelos foi efetuada a seguinte distribuição por conjuntos:



Agenda

Sumário

Recolha de dados

Minimização das perdas

Modelo binário

Modelo ternário

Discussão

Bibliografia

Minimização das perdas

Ao prever a classe do cliente (BOM ou MAU), podemos definir a matriz genérica de perdas, L:

		PREVISÃO	
		MAU	BOM
REAL	MAU	l_1	l_2
	BOM	l_3	l_4

Para **determinar os parâmetros da matriz de perdas L**, tem-se em conta que:

- A **aprovação** do débito implica cobrança de **comissões pelo serviço** e **geração/ aumento de descobertos**.
- Os **cheques recusados** implicam também a cobrança de uma **comissão** e o **montante médio** destas transações e comissões é **superior** ao das restantes.
- **Não existem perdas** quando a **classificação é correta**.
- Classificar como BOM quando é MAU, gera um descoberto na conta que não será regularizado nos 30 dias seguintes e classificar como MAU quando é BOM, conduz à perda de comissões/juros do serviço.

Minimização das perdas

Definimos então:

M = montante (comissões mais juros) a pagar pela aprovação de qualquer transação;

Pc = proporção de cheques;

Po = 1- P_c = proporção de outras transacções (não cheques);

Lc = descoberto provocado pela aprovação de cheques;

Lo = descoberto provocado pela aprovação de outras transacções;

F+ = comissão adicional por cada aprovação de cheque;

F- = comissão adicional por cada devolução de cheque.

Usando os parâmetros anteriormente definidos obtemos a matriz:

		PREVISÃO	
		MAU	BOM
REAL	MAU	0	$P_c L_c + (1 - P_c)L_o$
	BOM	$P_c(F_+ - F_-) + M$	0

Minimização das perdas

Nestas condições, **o mais importante não é minimizar o número de erros, mas sim minimizar as perdas/prejuízos**.

Usando uma amostra com decisões históricas e as respectivas comissões e montantes, chegou-se à matriz:

		PREVISÃO	
		MAU	BOM
REAL	MAU	0	49
	BOM	1	0

A matriz de perdas obtida reflete os critérios de negócio atualmente em vigor. No entanto, deve ser sempre levada em conta a **variabilidade** destes critérios na construção do modelo de decisão.

Agenda

Sumário

Recolha de dados

Minimização das perdas

Modelo binário

Modelo ternário

Discussão

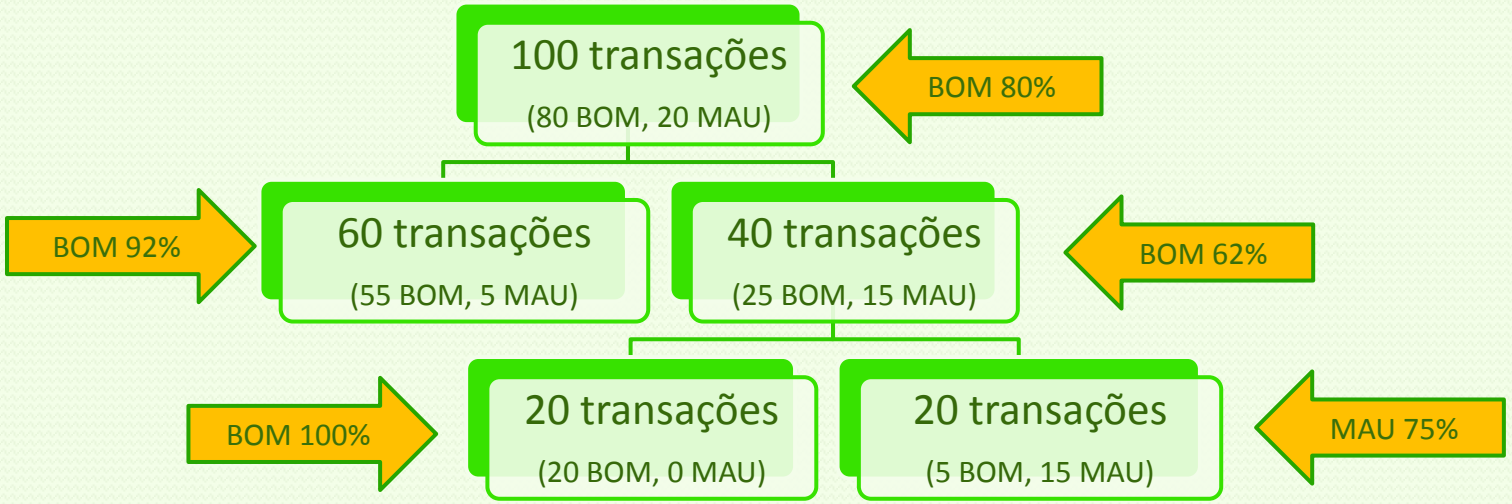
Bibliografia

Modelo binário

Pretende-se que o **resultado do modelo** contenha:

- a divisão da população em **duas classes** (BOM, MAU);
- **a probabilidade à posteriori em cada classe**.

Na prática, pretende-se dividir a população em grupos e, em função da probabilidade associada à classe dominante em cada grupo, definir como classificar os elementos desse grupo, minimizando os custos/perdas associadas a essa previsão.



Modelo binário

Serão abordadas duas estratégias:

A

Os dois tipos de **erros** **têm o mesmo peso** na decisão e incorporar os conceitos do negócio numa decisão baseada na probabilidade à posteriori de cada classe.

- Vantagem: fácil adaptação do modelo a variações dos conceitos do negócio;
- Desvantagem: se os conceitos do negócio conduzirem a tipos de erros muito assimétricos poderá diminuir a performance do modelo inicial.

		PREVISÃO	
		MAU	BOM
REAL	MAU	0	1
	BOM	1	0

B

Considerar **diferentes pesos para os diferentes erros**, incorporando assim os conceitos do negócio no modelo em si.

- Vantagem: modelo mais adaptado aos conceitos do negócio;
- Desvantagem: variações nos conceito de negócio implicam revisão do modelo.

		PREVISÃO	
		MAU	BOM
REAL	MAU	0	49
	BOM	1	0

Modelo binário

Dado que serão usadas **duas estratégias** e que irão ser usados **três tipos de modelos**: regressão logística, árvores de classificação e redes neuronais, será necessário analisar o poder discriminativo de cada modelo e comparar resultados. Para tal define-se a **matriz de confusão**, **C**, (ou tabela de contingência) como sendo:

		PREVISÃO		
		MAU	BOM	
REAL	MAU	P(RM,PM)	P(RM,PB)	P(RM)
	BOM	P(RB,PM)	P(RB,PB)	P(RB)
		P(PM)	P(PB)	1

Onde :

P(RM,PM) - probabilidade de o modelo prever MAU quando o cliente é MAU na realidade;

P(RM,PB) - probabilidade de o modelo prever BOM quando o cliente é MAU na realidade;

P(RB,PM) - probabilidade de o modelo prever MAU quando o cliente é BOM na realidade;

P(RB,PB) - probabilidade de o modelo prever BOM quando o cliente é BOM na realidade.

Modelo binário

Usando as matrizes L e C definidas anteriormente:

		PREVISÃO	
		MAU	BOM
REAL	MAU	l_1	l_2
	BOM	l_3	l_4

		PREVISÃO	
		MAU	BOM
REAL	MAU	P(RM,PM)	P(RM,PB)
	BOM	P(RB,PM)	P(RB,PB)

Podemos constatar que o valor esperado para as perdas resulta da soma dos pesos associados a cada tipo de perda (l_i) multiplicados pela probabilidade de ocorrerem, ou seja:

$$E[L \cdot C] = l_1 \cdot P(RM,PM) + l_2 \cdot P(RM,PB) + l_3 \cdot P(RB,PM) + l_4 \cdot P(RB,PB)$$

$$= P(RM)\left((l_1 - l_2) \cdot \underbrace{\frac{P(RM,PM)}{P(RM)}}_{\text{Sensibilidade}} + l_2 \right) + P(RB) \cdot (l_3 - l_4)\left(\underbrace{1 - \frac{P(RB,PB)}{P(RB)}}_{\text{1- Especificidade}} \right) + P(RB) \cdot l_4$$

Modelo binário

Então E[LC] pode ser escrita como uma função linear da forma:

$$E[LC] = a * (1 - \text{Especificidade}) + b * (\text{sensibilidade}) + c$$

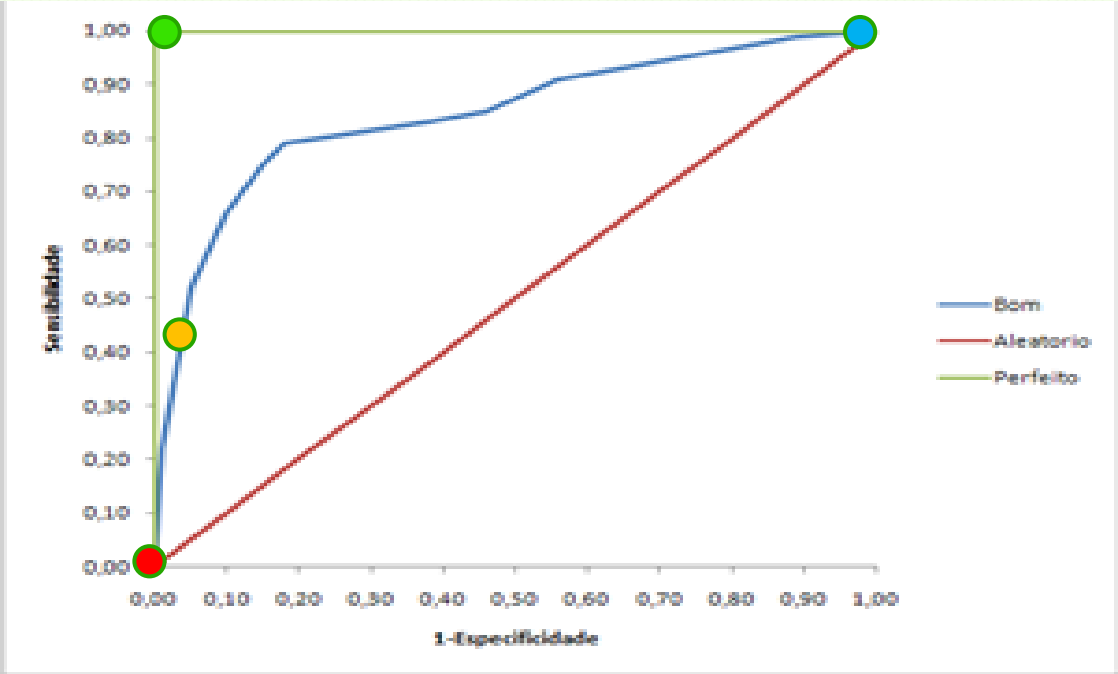
onde a, b e c são parâmetros dependentes das entradas da matriz L e da proporção de clientes da classe MAU na amostra.

A **performance** de um modelo pode então ser **analisada num espaço a duas dimensões**.

Para efetuar esta análise usaremos **curvas ROC** (Receiver Operating Characteristic).

Modelo binário

Abaixo é apresentado um exemplo de curvas ROC para diferentes classificadores.



- Ponto ótimo
- P = 90%
- P = 100%
- P = 0%

Cada ponto da curva representa uma probabilidade a partir da qual classificamos o cliente como MAU. Pontos no canto superior direito correspondem a probabilidades reduzidas e no canto inferior esquerdo correspondem a probabilidades elevadas.

Modelo binário

Os resultados obtidos foram os seguintes:

A Pesos iguais

Model	Loss	Specificity	Sensitivity	Error rate
Logistic Regression	0.094	98.4%	55.2%	9.4%
Decision Tree	0.083	97.8%	64.5%	8.3%
Neural Network	0.089	98.1%	59.6%	8.9%
Naive 1	0.820	0.0%	100.0%	82.0%
Naive 2	0.180	100.0%	0.0%	18.0%

B Pesos diferenciados, atendendo aos conceitos do negócio.

Model	Loss	Specificity	Sensitivity	Error rate
Logistic Regression	0.724	27.2%	98.6%	59.9%
Decision Tree	0.697	28.4%	98.8%	58.9%
Neural Network	0.697	34.3%	98.2%	54.1%
Naïve 1	0.820	0.0%	100.0%	82.0%
Naïve 2	8.820	100.0%	0.0%	18.0%

Agenda

Modelo ternário

Os **modelos binários não descriminam de forma satisfatória** os clientes. Para melhorar os resultados, permitir-se-á uma **terceira classe**, a **REVISÃO**. Esta classe implica que as transações necessitarão de **análise manual**, com eventual recolha de dados adicionais.

O número de clientes nesta classe deve ser reduzido, atendendo aos custos e tempo que tal processo exige.

Em analogia ao que foi efetuado para o modelo binário, pode-se então definir a matriz de confusão:

		PREVISÃO		
		MAU	REVISAO	BOM
REAL	MAU	P_1	P_2	P_3
	BOM	P_4	P_5	P_6

Objetivos:

- Minimizar P_3 e P_4 -> Diminuir os erros e as perdas;
- Minimizar P_2 e P_5 -> Aumentar a decisão automática.

Modelo ternário

Os resultados obtidos foram os seguintes:

Model	Cutoff Low	Cutoff High	Specificity	Sensitivity	Approved defaulters	Error rate	Automation
Logistic Regression	10.5%	30.5%	92.6%	84.8%	3.5%	7.2%	82.1%
Decision Tree	7.0%	23.7%	93.2%	80.2%	4.9%	8.0%	86.6%
Neural Network	11.0%	27.3%	92.8%	84.2%	3.7%	7.4%	85.0%

- **Maior equilíbrio entre Sensibilidade e Especificidade;**
- **Previsões mais assertivas;**
- **15% das transações são decididas manualmente.**

P_3/P_6

P_3+P_4

$P_1+P_3+P_4+P_6$

		PREVISÃO		
		MAU	REVISAO	BOM
REAL	MAU	P_1	P_2	P_3
	BOM	P_4	P_5	P_6

Agenda

Sumário

Recolha de dados

Maximização do lucro

Modelo binário

Modelo ternário

Discussão

Bibliografia

Discussão

Com o **objetivo** de criar um modelo para decidir autorizar ou não o débito de transações em contas com saldo insuficiente foram desenvolvidos dois tipos de modelos, binários e ternários.

Os resultados apresentados com **modelos de classificação binários** revelam que, apesar do estudo intensivo efetuado, nenhum dos modelos discriminou de forma satisfatória:

- Com pesos bastante assimétricos para os dois tipos de erro, a decisão foca-se na mitigação de erro associado às perdas mais elevadas.
- Com pesos iguais, a decisão foca-se apenas na previsão da classe dominante.

Para melhorar estes resultados, construíram-se **modelos de classificação ternários**, que possibilitavam a existência da classe REVISÃO, entre as classes BOM e MAU.

O modelo final obtido permitia uma **automatização de 87%**, com uma taxa de **erro de 8%**.

A incorporação futura dos critérios usados na revisão manual, poderá melhorar os resultados.

Agenda

Sumário

Recolha de dados

Maximização do lucro

Modelo binário

Modelo ternário

Discussão

Bibliografia

Bibliografia

[1] Maria R. SOUSA & Joaquim P. Costa. A Tripartite Scorecard for the Pay/No pay Decision-Making in the Retail Banking Industry. Applications of Data Mining in E-Business and Finance C. Soares et al. (Eds.) IOS Press, 2008, doi:10.3233/978-1-58603-890-8-45.

[2] Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook. Credit Scoring and its applications. SIAM, 2002.

[3] Roger M. Stein. The relationship between default prediction and lending profits: Integrating roc analysis and loan pricing. Journal of Banking & Finance, 29:1213–1236, 2005.

[4] J. R. Quinlan. Induction of decision trees. Machine Learning, 1(1):81–106, 1986.

[5] B.D. Ripley. Pattern Recognition and Neural Networks. Cambridge University Press, 1996.

[6] C. M. Bishop. Neural Networks for Pattern Recognition. Oxford University Press, 1995.

[7] K. B. Schebesch and R. Stecking. Support vector machines for classifying and describing credit applicants: detecting typical and critical regions. Journal of the Operational Research Society, 56:1082–1088, 2005.

[8] David J. Hand, So Young Sohn, and Yoonseong Kim. Optimal bipartite scorecards. Expert Systems with Applications, 29:684–690, 2005.

Questões