IV Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Os segredos da *Lifetime Probability of Default* através do modelo de Cox

Raphael Chaves[1], Ricardo Ehlers[2]
ICMC-USP

1 Introdução

A probabilidade de um devedor inadimplir ao longo da vida de uma determinada operação de crédito, ou *Lifetime Probability of Default* (LTPD), é essencial para uma gestão eficaz e prospectiva do risco de crédito. Tradicionalmente, a LTPD tem sido estimada por meio de modelos estatísticos de modelagem para o risco de crédito baseados na Regressão Logística, que consideram apenas um instante fixo no tempo e assumem uma relação estática entre variáveis explicativas e a probabilidade de inadimplência. No entanto, a abordagem de análise de sobrevivência oferece uma perspectiva mais flexível e dinâmica, ao considerar não apenas a ocorrência do evento de inadimplência, mas também o momento em que ele ocorre. Dentro desse contexto, o modelo de *Cox Proportional Hazard Regression* (CPH), especialmente quando estimado sob uma formulação bayesiana, destaca-se como uma alternativa poderosa e moderna à Regressão Logística tradicional.

Com base no acima exposto, o presente trabalho propõe o desenvolvimento e a aplicação de uma abordagem prática para estimar a LTPD utilizando técnicas de análise de sobrevivência, com ênfase na formulação bayesiana do modelo CPH. Espera-se, assim, obter estimativas mais estáveis e interpretáveis do risco de inadimplência ao longo da vida da operação de crédito, superiores à Regressão Logística.

2 Materais e métodos

2.1 Base de dados

Para o desenvolvimento deste trabalho, foi utilizado um conjunto de dados em formato painel, contemplando 23 variáveis que abrangem informações relativas à originação e ao desempenho de 50.000 contratos de empréstimos residenciais dos Estados Unidos da América (EUA), acompanhados ao longo de 60 meses. Os dados foram selecionados de maneira aleatória a partir de portfolios

[1] raphaelchaves@usp.br
[2] ehlers@icmc.usp.br

vinculados a títulos lastreados em hipotecas residenciais nos EUA *(Residential Mortgage-Backed Securities – RMBS)*, disponibilizados pela *International Financial Research*. Destaca-se que este conjunto de dados integra o acervo utilizado no livro: *Credit Risk Analytics: Measurement Techniques, Applications and Examples in SAS*, como em [2].

2.2 Divisão dos dados

O conjunto de dados mencionado na seção 2.1 foi dividido em duas amostras distintas: 80% dos dados foram destinados ao treinamento dos modelos, enquanto os 20% restantes foram reservados para validação e teste final. Essa divisão tem como objetivo garantir uma avaliação imparcial e robusta do desempenho preditivo dos modelos, simulando sua aplicação em dados inéditos e não utilizados durante o processo de estimação.

2.3 Transformação de variáveis

Adotou-se uma técnica de agrupamento para as quatro variáveis preditivas consideradas nos modelos, combinada com a transformação conhecida como Peso da Evidência, ou *Weight of Evidence* (WOE), como em [1]. Tal abordagem consiste em segmentar os valores contínuos dessas variáveis em intervalos discretos, ou *binnings*, e calcular, para cada *bin*, o logaritmo da razão entre as proporções de bons e maus pagadores, conforme ilustrado nas figuras 1, 2, 3 e 4.

LTV_time

Bin	# Goods	# Bads	%Dist. Total	WOE	IV	Bad Rate
01 (-Inf,43.2995)	26.414	177	5,3%	1,3163	0,0524	0,7%
02 [43.2995,54.5122)	32.122	292	6,5%	1,0113	0,0425	0,9%
03 [54.5122,69.6708)	84.556	871	17,1%	0,8863	0,0906	1,0%
04 [69.6708,73.42)	28.042	339	5,7%	0,7262	0,0216	1,2%
05 [73.42,78.1867)	37.565	536	7,6%	0,5605	0,0186	1,4%
06 [78.1867,81.0774)	27.233	436	5,6%	0,4453	0,0090	1,6%
07 [81.0774,88.2272)	54.473	1.300	11,2%	0,0461	0,0002	2,3%
08 [88.2272,93.5798)	33.403	947	6,9%	-0,1261	0,0012	2,8%
09 [93.5798,108.2867)	81.725	3.058	17,0%	-0,4036	0,0337	3,6%
10 [108.2867,113.8083)	30.582	1.413	6,4%	-0,6145	0,0328	4,4%
11 [113.8083,Inf)	49.648	2.772	10,5%	-0,8038	0,1013	5,3%
Missing Values	246	5	0,1%	0,2067	0,0000	2,0%
	486.009	**12.146**	**100,0%**		**0,4040**	

Figura 1: *Loan-to-Value* no período de observação.

gdp_time

Bin	# Goods	# Bads	%Dist. Total	WOE	IV	Bad Rate
01 (-Inf,-3.3395)	26.983	1.401	5,7%	-0,7312	0,0437	4,9%
02 [-3.3395,-0.2411)	42.638	1.895	8,9%	-0,5757	0,0393	4,3%
03 [-0.2411,0.893)	29.876	1.180	6,2%	-0,4577	0,0163	3,8%
04 [0.893,1.2292)	43.141	1.268	8,9%	-0,1622	0,0025	2,9%
05 [1.2292,2.1514)	118.826	2.879	24,4%	0,0310	0,0002	2,4%
06 [2.1514,2.8364)	117.077	2.374	24,0%	0,2090	0,0095	2,0%
07 [2.8364,3.0695)	50.526	574	10,3%	0,7884	0,0447	1,1%
08 [3.0695,Inf)	56.942	575	11,5%	0,9062	0,0633	1,0%
	486.009	**12.146**	**100,0%**		**0,2196**	

Figura 2: Crescimento do PIB no período de observação.

FICO_orig_time

Bin	# Goods	# Bads	%Dist. Total	WOE	IV	Bad Rate
01 (-Inf,613)	95.730	3.527	19,9%	-0,3881	0,0363	3,6%
02 [613,655)	88.806	2.854	18,4%	-0,2515	0,0131	3,1%
03 [655,687)	81.330	2.227	16,8%	-0,0914	0,0015	2,7%
04 [687,705)	44.723	1.006	9,2%	0,1053	0,0010	2,2%
05 [705,731)	56.808	1.123	11,6%	0,2344	0,0057	1,9%
06 [731,747)	31.410	510	6,4%	0,4312	0,0098	1,6%
07 [747,765)	33.230	432	6,8%	0,6536	0,0214	1,3%
08 [765,781)	26.265	250	5,3%	0,9653	0,0323	0,9%
09 [781,Inf)	27.707	217	5,6%	1,1603	0,0454	0,8%
	486.009	**12.146**	**100,0%**		**0,1665**	

Figura 3: Pontuação de crédito FICO na originação.

Interest_Rate_orig_time

Bin	# Goods	# Bads	%Dist. Total	WOE	IV	Bad Rate
01 (-Inf,6.4)	258.786	4.633	52,9%	0,3336	0,0504	1,8%
02 [6.4,6.8)	47.349	983	9,7%	0,1855	0,0031	2,0%
03 [6.8,7.13)	38.592	1.127	8,0%	-0,1557	0,0021	2,8%
04 [7.13,7.69)	39.726	1.390	8,3%	-0,3365	0,0110	3,4%
05 [7.69,Inf)	101.556	4.013	21,2%	-0,4582	0,0556	3,8%
	486.009	**12.146**	**100,0%**		**0,1222**	

Figura 4: Taxa de juros vigente no período de observação.

Observa-se que essa abordagem contribui para a formação de agrupamentos que apresentam uma relação monotônica entre o índice WOE e a taxa de maus pagadores, ou *Bad Rate*, agregando clareza interpretativa ao processo de modelagem. Tal característica é particularmente vantajosa para variáveis como a pontuação de crédito FICO (veja Figura 3), na qual se verifica que, à medida que a pontuação de crédito aumenta, há uma redução progressiva da *Bad Rate*, em consonância com a expectativa teórica de risco. Adicionalmente, para cada variável agrupada, foi calculada a estatística conhecida como Valor da Informação, ou *Information Value* (IV), a qual mensura o poder discriminatório individual das variáveis na identificação dos maus pagadores. Via de regra, valores elevados de IV indicam maior capacidade explicativa da variável na distinção entre bons e maus pagadores (veja Tabela 1), como em [3]:

Tabela 1: Interpretação do poder preditivo utilizando IV.

IV	Poder Preditivo
< 0.02	Não preditivo
$>= 0.02$ e < 0.10	Fraco para predição
$>= 0.10$ e < 0.30	Médio para predição
$>= 0.30$	Forte para predição

3 Desempenho dos modelos

A avaliação do desempenho dos modelos na amostra de teste evidencia diferenças relevantes entre as abordagens. Conforme ilustrado a seguir, o modelo CPH sob uma formulação bayesiana apresentou aderência superior à curva observada de *Bad Rate*, em especial evidenciado pela curva ROC, onde o modelo CPH alcançou uma área sob a curva (AUC) de 0,7289, superior à Regressão

Logística com AUC de 0,7148. Para quantificar a aderência, calculou-se a Raiz do Erro Quadrático Médio (RMSE), onde o modelo CPH obteve 0,3754%, inferior ao valor de 0,7623% da Regressão Logística.

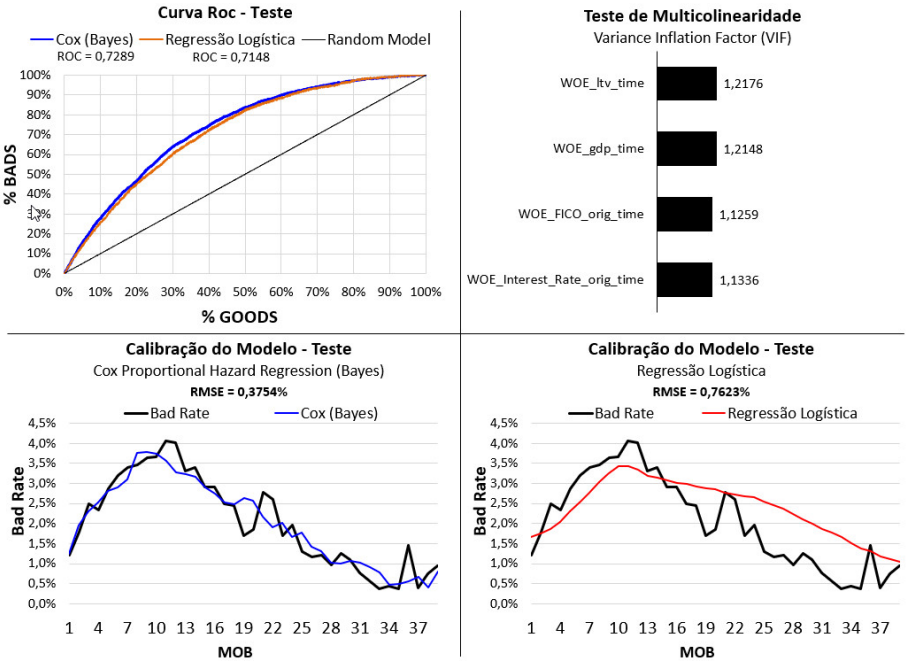

Figura 5: Desempenho dos Modelos.

Adicionalmente, é ilustrado o teste de multicolinearidade com base no Fator de Inflação da Variância, ou *Variance Inflation Factor* (VIF), o qual apresentou valores próximos a 1, indicando baixa correlação entre as variáveis explicativas e reforçando a robustez do modelo. A análise visual dos gráficos ilustrados acima reforça que o modelo CPH acompanha de modo mais preciso a evolução temporal da taxa de inadimplência ao longo dos meses de operação (MOB), apresentando menor distância para os valores reais, inclusive nos movimentos cíclicos e nos pontos de inflexão da série. Em contrapartida, o modelo de Regressão Logística tende a suavizar demais as variações, perdendo acurácia e capacidade de capturar oscilações relevantes do risco de crédito.

Referências

[1] A. Raymond. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. 1a. edição*. Oxford University Press, USA, 2007.

[2] B. Bart, R. Daniel, and S. Harald. *Credit risk analytics: Measurement techniques, applications, and examples in SAS. 1a. edição*. John Wiley & Sons, 2016.

[3] R. Mamdouh. *Credit risk scorecard: development and implementation using SAS. 1a. edição*. Lulu.com, 2011.