Preparativos para o Novo Acordo da Basiléia Parte 2: Estratégias de Modelagem

Jeffrey S. Morrison

Preparing for Basel II Part 2: Modeling Strategies

Este artigo dá continuidade à série que trata dos requisitos do Novo Acordo da Basiléia e da construção de modelos PI e PCI, oferecendo algumas dicas sobre estratégia de construção de modelos para que o analista possa extrair o maior número de dados disponíveis. O texto concentra-se na escolha das preditivas, variáveis *dummy*, transformações e interações corretas, na avaliação de *outliers* influentes e no desenvolvimento de esquemas de segmentação.

Continuing the series dealing with requirements for Basel II and building PD and LGD models, this article offers some tips in model-building strategy so the analyst can get most out of the data available. The focus is on selecting the right predictors, dummy variables, transformations, and interactions, evaluating influential outliers, and developing segmentation schemes.

Foram escritos inúmeros livros sobre a probabilidade de inadimplência (PI), a perda em caso de inadimplência (PCI) e as técnicas estatísticas usadas para a sua estimativa. Há, contudo, escassez de literatura sobre conselhos práticos adequadas para que o analista possa produzir, agilmente, melhores modelos. Há diversidade de abordagens para cada tópico — muitas podendo funcionar igualmente bem. Independentemente de estar o analista construindo um modelo de PI ou de PCI, algumas estratégias práticas podem dar um bom impulso inicial no alinhamento de seu processo de modelagem ao

Volumes of literature have been written on probability of default, loss given default, and the statistical techniques used to estimate them. There's a deficit of literature, however, on practical advice so the analyst can produce better models more quickly. For each topic there are a variety of approaches — many of which can work equally well. Regardless of whether the analyst is building a PD or LGD model, some practical strategies can provide a good jump-star in aligning your modeling process to the New Basel Capital Accord. In

this article, we will refer to a hypothetical dataset to model the probability of default using such predictive attributes as LTV, payment history, bureau scores, and income. And although these techniques can be implemented in just about any statistical language, examples will be given using the SAS software language as well as S-PLUS.

Selecting the Right Predictors

Althogh good modeling data may be initially hard to find, once your bank has put together the informational infrastructure for the New Capital Accord, you might have more data than you know what to do with. In building a PD model using a regression technique, typically only 10 or 15 predictor variables are finally chosen. However, you may have 200 or more potentially predictive attributes coming from your loan accounting systems, not to mention outside vendors such as economic service providers. So how do you narrow down all that information? Sound judgment, combined with knowledge of simple correlations in your data, is a good starting point. Luckily, there are some additional tools to make the model-building process less stressful.

The first is the automatic variable selection routines found in most statistical pa-

Variable clustering identifies similar groups.

Novo Acordo de Capital da Basiléia. Neste artigo, nos referimos a um conjunto de dados hipotético para modelar a probabilidade de inadimplência, usando variáveis preditivas como a razão entre empréstimo e valor (LTV), histórico de pagamentos, *scores* de *credit bureau* e renda. Embora essas técnicas possam ser implementadas em praticamente qualquer linguagem estatística, os exemplos serão dados na linguagem do *software* SAS e em S-PLUS.

O agrupamento de variáveis identifica grupos semelhantes.

Escolhendo as Preditivas Corretas

Apesar da dificuldade em encontrar bons dados para a modelagem, depois de um banco ter estabelecido a estrutura informacional necessária para o Novo Acordo de Capital, haverá sempre dados supérfluos. Quando se constrói um modelo de PI, por meio de uma técnica de regressão, normalmente são escolhidas apenas 10 ou 15 variáveis preditivas. Mas poderá haver mais de 200 atributos potencialmente preditivos, vindos dos sistemas de contabilidade de empréstimos, além de fornecedores externos, como os prestadores de serviços de economia. Como proceder para separar todas essas informações? Um bom ponto de partida é um julgamento sólido, aliado ao conhecimento de correlações simples entre seus dados. Felizmente, há mais algumas ferra-

mentas que tornam o processo de construção de modelos menos estressante.

Em primeiro lugar aparecem as rotinas de seleção automática de variáveis encontradas na maioria dos pacotes estatísticos. Normalmente, são conhecidas como procedimentos *stepwise* adiante, *stepwise* à ré e melhor encaixe. A seleção, a seguir, constrói o modelo de baixo para cima, adicionando uma nova variável eliminando quaisquer variáveis anteriores que não demonstrem ser estatisticamente significativas de acordo com um critério prestabelecido. O procedimento à ré funciona da maneira oposta. Parte de todas as preditivas disponíveis e as elimina uma a uma com base em seu nível de significância. Os procedimentos de melhor encaixe podem consumir mais tempo de computação porque experimentam muitas variáveis diferentes para maximizar alguma medida de encaixe, em vez do nível de significância de cada uma. O que fazer? É necessário experimentar todos os procedimentos, mas não será surpresa se o procedimento *stepwise* à ré funcionar um pouco melhor no longo prazo. Porém é preciso atenção: só devem ser incluídas no modelo variáveis que estejam de acordo com o ponto de vista do negócio, independentemente do resultado de qualquer procedimento automático.

Binning helps increase a model's predictive power.

Binning ajuda a aumentar o poder preditivo do modelo.

ckages. Generally, they are called forward stepwise, backward stepwise, and best-fit procedures. Forward selection builds the model from the ground up, entering a new variable and eliminating any previous variables not found statistically significant according to a specified criteria rule. The backward stepwise procedure works just the opposite. It starts with all available predictors and eliminates them one by one based upon their level of significance. Best-fit procedures may take much more computer time because they may try a variety of different variables to maximize some measure of fit rather than the significance level of each variable. What's the advice? Experiment with them all, but don't be surprised if the backward stepwise procedure works a little better in the long run. However, a word of caution–only include variables in your model that make business sense, regardless of what comes out of any automatic procedure.

One major drawback of these procedures is that they can lead to variables in your final model that are too correlated with one another. For example, of local area employment and income are highly correlated with one another, the stepwise procedures may not be able to correctly iso-

late their individual influence. As a result, they could both end up in your model. This is a condition called multicollinearity and may cause problems in the model-building process. One way of avoiding this problem is to perform a correlation analysis before you run any regression procedure. The advice–only include predictor variables in your regression that have a correlation with each other less than .75 in absolute values. If you do run into variables that are too highly correlated with one another, choose the one that is the most highly correlated with your dependent variable and drop the other.

A second tool is called variable clustering, a set of procedures that can help the analyst jump some of the hurdles associated with using an automatic selection routine. This procedure seeks to find groups or clusters of variables that look alike. The optimal predictive attributes would be those that are highly correlated with variables within the cluster, but correlated very little with variables in other clusters. SAS, for example, has an easy-to-use procedure called PROC VARCLUS that produces a table such as the one shown in Figure 1.

Clustering routines do not offer any insight into the relationship of the individual

CART is an exploratory analysis tool.

O CART é uma ferramenta de análise exploratória.

O ponto fraco desses procedimentos é que eles podem conduzir a variáveis excessivamente correlacionadas entre si em seu modelo final. Por exemplo, se o nível de emprego e o de renda da área estiverem fortemente correlacionados, os procedimentos *stepwise* podem não ser capazes de isolar corretamente a influência individual de cada um. Com isso, os dois poderiam acabar sendo incluídos no modelo. Essa é uma situação chamada multicolinearidade, capaz de causar problemas no processo de construção de modelos. Um meio de evitar esse problema é fazer uma análise de correlação, antes de rodar qualquer rotina de regressão. Assim, só se deve incluir na regressão variáveis preditivas que tenham entre si correlação inferior a 0,75, em termos absolutos. E, se forem encontradas variáveis com correlação excessiva entre si, a escolha deve recair naquela mais fortemente correlacionada com sua variável dependente, abandonando-se a outra.

Outra ferramenta é o agrupamento de variáveis, um conjunto de procedimentos que pode ajudar a superar alguns dos obstáculos associados ao uso de uma rotina de seleção automática. Esse procedimento tem por objetivo identificar grupos de variáveis semelhantes entre si. Os atributos preditivos ideais são aqueles forte-

mente correlacionados com variáveis de seu próprio agrupamento, mas pouco correlacionados com as variáveis de outros agrupamentos. O SAS, por exemplo, tem um procedimento de fácil uso chamado PROC VARCLUS que produz uma tabela como a da Figura 1.

As rotinas de agrupamento não oferecem qualquer *insight* sobre a relação entre as variáveis individuais e a variável dependente (PI ou PCI). Mas oferecem uma maneira muito mais fácil de estudar os atributos preditivos. Pode ser que o Agrupamento 1 represente dados sobre delinqüência em hábitos de pagamento. O Agrupamento 2 poderia refletir dados geográficos ou econômicos. Assim, num procedimento de regressão, deve-se testar as variáveis de cada agrupamento. Como decidir qual deve ser a escolhida? Uma maneira é escolher as variáveis com os menores valores-índice constantes da Coluna D — uma medida facilmente calculada pela maior parte dos *softwares* estatísticos populares, como o SAS. Em nosso exemplo, poderíamos escolher a VAR7 do Agrupamento 1 e a VAR5 do Agrupamento 2 para uso numa regressão stepwise.

Variáveis Dummy

Grande parte das informações à disposição sobre uma conta qualquer encontra-se sob a forma de variáveis categóricas, como, por exemplo, tipo de produto, código da garantia ou o estado em que se encontram o cliente ou a garantia real. Por exemplo, digamos que se tenha um código de tipo de produto armazenado como um número inteiro entre 1 e 4. Se a variável tiver sido colocada num modelo de regressão sem qualquer alteração, o significado de seu coefi-

variables and our dependent variable (PD or LGD). What they do offer, however, is a much easier way of looking at your predictive attributes. It may be that Cluster 1 represents delinquent payment behavior data. Cluster 2 could reflect geographic or economic data. Therefore, in a regression procedure you should be sure to test variables from each cluster. So how do you know which to pick? One way is to choose those variables with the lowest ratio values as indicated in Column D — a measure easily computed by most popular statistical software such as SAS. In our example, we might select VAR7 from Cluster 1 and VAR5 from Cluster 2 to try in a stepwise regression.

Dummy Variables

Much of the information you might have on an account is in the form of a categorical variable. These are variables such as product type, collateral code, or the state

Figure 1

Variable Clustering

Variable A	B	C	D
	R-Squared Own Cluster	R-Squared Next Cluster	1-R-Squared Ratio
Cluster 1			
VAR1	0.334	0.544	1.460
VAR2	0.544	0.622	1.206
VAR7 -->	**0.322**	**0.242**	**0.894**
Cluster 2			
VAR5 -->	**0.988**	**0.433**	**0.021**
VAR8	0.544	0.231	0.592
VAR3	0.764	0.434	0.416
VAR4	0.322	0.511	1.386

in which the customer or collateral resides. For example, let's say you have a product type code stored as an integer ranging from 1 to 4. If the variable was entered into a regression model without any changes, then its coefficient's meaning might not make much sense, let another be predictive. However, if we recoded it as follows, then the regression could pick up the differences each product category makes in the PD or LGD, all other things remaining equal:

IF PRODUCT_CODE=1 THEN DUMMY_1=1;
ELSE PRODUCT CODE=0;
IF PRODUCT_CODE=2 THEM DUMMY_2=1;
ELSE PRODUCT CODE=0;
IF PRODUCT_CODE=3 THEN DUMMY_3=1;
ELSE PRODUCT CODE=0;
OF PRODUCT_CODE=4 THEN DUMMY_4=1;
ELSE PRODUCT CODE=0;

What you now have are four product code dummy variables that will allow the model to estimate the default impact among different product types. By convention, one of the variables has to be left out in the

Figure 2

Step 1 - Discretizing your Data

Grouping Range	SAS Code
455-569	IF B_SCORE>455 AND B_SCORE<=569 THEN B_SCORE=569;
570-651	IF B_SCORE>569 AND B_SCORE<=651.5 THEN B_SCORE=651;
652-695	IF B_SCORE>651 AND B_SCORE<=695 THEN B_SCORE=695;
696-764	IF B_SCORE>695 AND B_SCORE<=764 THEN B_SCORE=764;
765-850	IF B_SCORE>764 AND B_SCORE<=850 THEN B_SCORE=849;

Figura 1

Agrupamento de Variáveis

Variável A	B	C	D
	R2 Próprio Agrupamento	R2 Agrupamento Seguinte	Razão 1-R2
Agrupamento 1			
VAR1	0,334	0,544	1,460
VAR2	0,544	0,622	1,206
VAR7 -->	**0,322**	**0,242**	**0,894**
Agrupamento 2			
VAR5 -->	**0,988**	**0,433**	**0,021**
VAR8	0,544	0,231	0,592
VAR3	0,764	0,434	0,416
VAR4	0,322	0,511	1,386

ciente pode ser absurdo, quanto mais ser preditivo. Contudo, se o registro se der da maneira abaixo, a regressão poderá captar a diferença que cada categoria de produto faz para a PI ou a PCI em igualdade das demais condições:

IF PRODUCT_CODE=1 THEN DUMMY_1=1;
ELSE PRODUCT CODE=0;
IF PRODUCT_CODE=2 THEM DUMMY_2=1;
ELSE PRODUCT CODE=0;
IF PRODUCT_CODE=3 THEN DUMMY_3=1;
ELSE PRODUCT CODE=0;
OF PRODUCT_CODE=4 THEN DUMMY_4=1;
ELSE PRODUCT CODE=0;

O que temos agora são quatro variáveis *dummy*, para o código do produto que permitirão ao modelo estimar o impacto de diferentes tipos de produto sobre a inadimplência. Por convenção, uma das variáveis precisa ser deixada fora do modelo de regressão, ou surgirá um erro. Assim, poder-se-ia incluir a *dummy*_1, a *dummy*_2, e a *dummy*_3 no modelo de regressão, em vez de uma só variável representando todos os valores do código do produto.

Binning de Variáveis

Partindo do conceito da variável *dummy*, às vezes um procedimento chamado *binning* pode ajudar o analista a aumentar ainda mais o poder preditivo do modelo. Embora haja diversas abordagens diferentes de *binning* disponíveis, todas partem do conceito de que segmentar uma variável em intervalos pode oferecer informações adicionais. Ao lidar com uma variável que tenha valores contínuos, como um *score* de credit bureau, pode ser usado um processo em duas etapas. Primeiro faz-se algo chamado de discretização, transformando o valor original em faixas para suavizar os dados — e criando um pequeno número de valores definidos. Um exemplo desse processo é apresentado na Figura 2, onde o score foi recodificado para que tivesse apenas cinco valores.

Figura 2

Passo 1 - Discretização dos Dados

Faixa de Agrupamento	Código SAS
455-569	IF B_SCORE>455 AND B_SCORE<=569 THEN B_SCORE=569;
570-651	IF B_SCORE>569 AND B_SCORE<=651,5 THEN B_SCORE=651;
652-695	IF B_SCORE>651 AND B_SCORE<=695 THEN B_SCORE=695;
696-764	IF B_SCORE>695 AND B_SCORE<=764 THEN B_SCORE=764;
765-850	IF B_SCORE>764 AND B_SCORE<=850 THEN B_SCORE=849;

Depois de reagrupar os dados em cinco faixas de valor, basta criar variáveis *dummy* a partir deles, como mostra a Figura 3.

Observe-se que a última variável de agrupamento não foi realizada porque ela seria automaticamente deixada fora do modelo, por convenção. A razão pela qual esse procedimento

regression model or an error will result. Therefore, you would include dummy_1, dummy_2, and dummy_3 in your regression model rather than a single variable presenting all values of the product code.

Variable Binning

Building on the dummy variable concept, a procedure called binning sometimes helps the analyst increase the model's predictive power even further. Although there are a variety of different binning approaches available, all leverage off the idea that breaking down a variable into intervals can provide additional information. When dealing with a variable that has continuous values, such as a bureau score, a two-step process can be implemented. First we do something called discretizing, where the original value is changed into ranges in order to smooth the data — creating a small number of discrete values. An example of this process is shown in Figure 2, where the bureau score was recoded into having only five values.

Now that you have regrouped the data

Figure 3

Step 2 – Dummy Variables / Binning

```
B_SCORE_1 =0;
IF B_SCORE=569 THEN B_SCORE_1 =1;
B_SCORE_2 =0;
IF B_SCORE =651 THEN B_SCORE_2 =1;
B_SCORE_3 =0;
IF B_SCORE= 695 THEN B_SCORE_3 =1;
B_SCORE_4 =0;
IF B_SCORE =764 THEN B_SCORE_4 =1;
```

into five range values, simply create dummy variables from them as shown in Figure 3.

Note that the last grouping variable was not done because we would automatically leave this out of the model by convention. The reason this procedure sometimes works to increase the accuracy of the model is because it allows the model to be more flexible in estimating the relationship between the bureau score, for example, and the default when the relationship may not be strictly linear. This is a great technique to try on other variables such as LTV, income, or months on books.

Transformations

Many times, simple transformations of the data end up making the model more predictive. If you are working with a PD model, one thing you could do as part of the preliminary analysis is to run a logistic regression using only a single predictor variable, calculating the log likelihood measure as shown in Figure 4. This measure is automatically produced by most statistical software packages.

Figure 4

Transformation Analysis–Logistic

Obs	_Status_	_Lnlike_	Name	T_Type	Recomm
1	0 Converged	-423.744	INCOME	XSQ	<-------*
2	0 Converged	-423.744	INCOME	QUAD	
3	0 Converged	-434.979	INCOME	NONE	
4	0 Converged	-443.669	INCOME	SQRT	
5	0 Converged	-454.364	INCOME	LOG	
6	0 Converged	-478.693	INCOME	INV	

às vezes aumenta a precisão do modelo é o fato de que ele permite que ele seja mais flexível ao estimar a relação entre o *score*, por exemplo, e a inadimplência, mesmo que a relação não seja estritamente linear. Essa é uma excelente técnica que pode ser experimentada com outras variáveis, como razão empréstimo/valor (LTV), renda ou meses de presença no balanço.

Figura 3

Passo 2 – Variáveis Dummy / Binning

```
B_SCORE_1 =0;
IF B_SCORE=569 THEN B_SCORE_1 =1;
B_SCORE_2 =0;
IF B_SCORE =651 THEN B_SCORE_2 =1;
B_SCORE_3 =0;
IF B_SCORE= 695 THEN B_SCORE_3 =1;
B_SCORE_4 =0;
IF B_SCORE =764 THEN B_SCORE_4 =1;
```

Transformações

Em muitos casos, simples transformações dos dados acabam aumentando o poder preditivo do modelo. Caso se trabalhe com um modelo de PI, o que pode ser feito como parte da análise preliminar é rodar uma regressão logística usando apenas uma variável preditiva, calculando a medida logarítmica probabilidade, como mostra a Figura 4. Essa medida é produzida automaticamente pela maioria dos pacotes de *software* estatístico.

Se for feito o mesmo com diferentes transformações, como um quadrado, a raiz quadrada e o logaritmo da variável, deve-se escolher a transformação em que a medida logarítmica da probabilidade esteja mais próxima a 0. É necessário incluir essa transformação "vencedora" no

Figura 4

Análise da Transformação – Logística

Obs	_Status_	_Lnlike_	Nome	T_Type	Recomm
1	0 Converged	-423,744	INCOME	XSQ	<-------*
2	0 Converged	-423,744	INCOME	QUAD	
3	0 Converged	-434,979	INCOME	NONE	
4	0 Converged	-443,669	INCOME	SQRT	
5	0 Converged	-454,364	INCOME	LOG	
6	0 Converged	-478,693	INCOME	INV	

modelo pleno de regressão, juntamente com as outras preditivas. Mas não há garantias de que tal transformação melhore o encaixe geral do modelo uma vez acrescentadas as outras variáveis preditivas. Deve-se ter sempre presente que a modelagem é uma arte, mas também é uma ciência!

Interações

Podem ser criadas variáveis adicionais que reflitam a interação entre duas outras. Essas novas variáveis podem, às vezes, acabar aumentando a precisão do modelo. Normalmente, a criação dessas variáveis decorre de um conhecimento prévio daquilo que deve influenciar as probabilidades de inadimplência ou a PCI. Para criar uma variável desse tipo, basta multiplicar aquelas que se deseja interagir e incluir a nova na regressão.

Outliers Influentes

As observações chamadas ponto de influência podem ter maior impacto sobre as estimativas de parâmetros do que outras. Durante o processo de desenvolvimento do modelo, é sempre bom produzir uma plotagem de influência para decidir que medidas corretivas são

If you do the same for various transformations such as the square, the square root, and the log of the variable, pick the transformation where the log likelihood measure is closest to 0. Include this "winning" transformation in your full regression model along with other predictors. However, there is no guarantee that such a transformation will improve the overall fit of the model once the other predictor variables have been added. Remember, modeling is as much of an art as it is a science!

Interactions

Additional variables can be created reflecting the interaction of two variables together that may sometimes end up increasing your model's accuracy. Usually the creation of these variables comes from prior knowledge as to what should influence default probabilities or LGD. To create such a variable, simply multiply together the variables you wish to interact and include the new variable in the regression.

Influential Outliers

Observations called influence points can impact the parameter estimates more than other. During the model development process, it is always good to produce an influence plot so you can decide what corrective action (if any) is necessary. Figure 5 is a bubble chart showing that observation #100 exerts significant influence on the size of your regression estimates. This type of analysis usually uses a shortcut appro-

ach to see how the predicted probabilities (and estimated parameters) would change if each observation were removed. The rule of thumb is to examine any observations that have a change in deviance greater then 4.0, as represented by the dotted horizontal line in Figure 5. Some software packages call these measures by different names, but most produce some option for looking at influential observations. In reality, there may always be some observations that show up on a graph of this type, but the worst ones should be reviewed and

Figure 5

Influence Plot

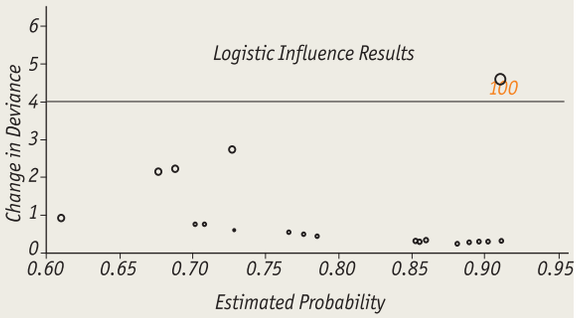


Figure 6

Typical CART. Segmentation Analysis in S-PLUS Software

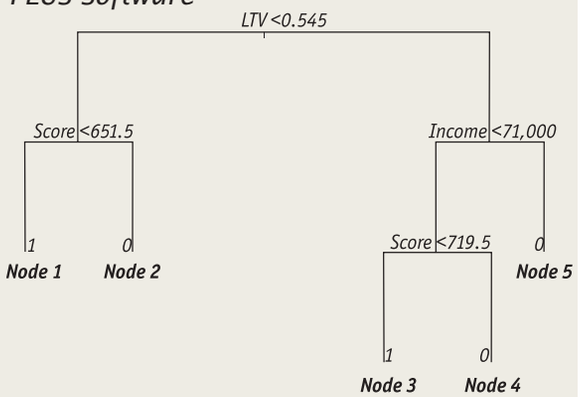
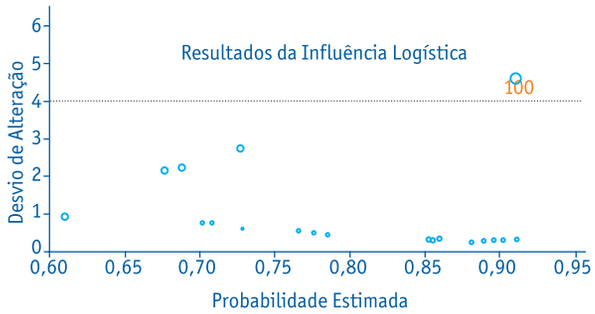


Figura 5

Plotagem de Influência



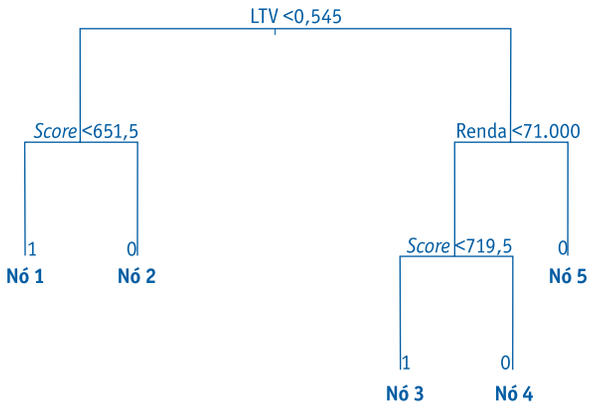
necessárias, eventualmente. A Figura 5 é um gráfico-bolha que mostra que a observação nº 100 exerce influência significativa sobre o porte das estimativas da regressão. Esse tipo de análise costuma usar um atalho para mostrar como as probabilidades previstas (e os parâmetros estimados) mudariam com a remoção de cada observação. A regra de bolso é examinar quaisquer observações que apresentem alteração do desvio superior a 4,0 conforme indica a linha horizontal pontilhada da Figura 5. Alguns pacotes de *software* denominam essas medidas diferentemente, mas a maioria produz alguma opção de análise de observações influentes. Na verdade, sempre pode haver alguma observação que surja num gráfico desse tipo, mas as piores devem ser analisadas e é preciso chegar a uma decisão sobre a eliminação de cada uma delas. No que se refere ao código de implementação de um modelo, é sempre bom limitar os valores extremos das preditivas, estabelecendo determinados tetos ou pisos. Por exemplo, pode ser interessante restringir o valor do LTV a algo entre 0 e 100 no algoritmo de codificação, antes de usar coeficientes ou ponderações para prever a probabilidade de inadimplência.

Segmentação

Às vezes é possível construir um melhor modelo de PI ou PCI se for feita uma análise de segmentação. Nesse contexto, a segmentação se refere ao desenvolvimento de mais de um modelo de PI ou PCI para melhorar a precisão geral. A eficácia de dedicar o tempo necessário para a construção de mais modelos, depende das informações preditivas disponíveis e de serem ou não as populações segmentadas realmente diferentes, umas das outras. Às vezes os grupos de segmentação são óbvios, como, por exemplo, modelar separadamente os empréstimos à pessoa física daqueles à pequena empresa. Entretanto, às vezes a segmentação não é tão evidente assim. Por exemplo, poder-se construir um modelo de PI para empréstimos com saldos elevados, médios e baixos, ou construir modelos separados para contas com diferentes níveis de renda ou de LTV.

Figura 6

CART Típico. Análise de Segmentação no Software S-PLUS Software



Então, como definir os pontos de segmentação entre alto, médio e baixo? Uma abordagem comum é chamada CART, que significa Classifica-

Segmentation

Sometimes a better overall PD or LGD model can be built if a segmentation analysis is performed. Segmentation in this context refers to the development of more than one PD or LGD model for better overall accuracy. The effectiveness of devoting the time needed to build additional models depends on the predictive information available and whether or not segmented populations are really that different from one another. Sometimes the segmentation groups are obvious, such as modeling consumer loans apart from loans to small businesses. Sometimes, the segmentation is not so obvious. For example, you could build a PD model for loans with high, medium, or low balances, or separate models for accounts with different levels of income or LTV.

So how do you know how to define the high, medium, or low segmentation breaks? One common approach is something called CART. CART stands for Classification and Regression Trees — a method quite different from regression analysis.

a decision made as to whether they should be deleted. With respect to a model's implementation code, it is always a good idea to limit extreme predictor values to certain maximums or minimums. For example, you might want to restrict an LTV value to be between 0 and 100 in the coding algorithm before coefficients or weights are used to predict the probability of default.

It uses an approach referred to as recursive partitioning to determine breaks in your data. It does this by first finding the best binary split in the data, followed by the next best binary split, and so forth.

CART is an exploratory analysis tool that attempts to describe the structure of your data in a tree-like fashion. S-Plus, one of the more popular CART vendors, uses deviance to determine the tree structure. As in regression, CART relates a single dependent variable (in our case, default or LGD) to a set of predictors. An example of this tree-like structure is shown in Figure 6. CART first splits on accounts where LTV is greater than and less than 54.5%. Therefore, you could design a PD model for accounts with LTVs below this cutoff and one for accounts greater than this cutoff. If you had even more time on your hands, you might want to consider using score and income to further define your segmentation schemes.

Sometimes it is possible to use CART analysis directly in a logistic regression model. This is a much quicker process than building and implementing separate regression models for each segment. So how can we integrate the CART methodology into a single PD model? The answer is to construct a variable that reflects the tree structure for the branches in the tree. Figure 6 shows that the tree structure produces five nodes. A node represents the end of the splitting logic and the segmentation branch of the tree. Therefore, in order to

tion and Regression Trees ('Árvores de Classificação e Regressão') – um método bem diferente da análise de regressão, que usa uma abordagem conhecida como particionamento recursivo para determinar os pontos de interrupção dos dados. Isso é feito identificando a melhor divisão binária dos dados, seguida pela segunda melhor divisão binária e assim por diante.

O CART é uma ferramenta de análise exploratória que procura descrever a estrutura de seus dados como se fosse uma árvore. A S-Plus, uma das mais conhecidas fornecedoras do CART, usa o desvio para determinar a estrutura da árvore. Como se fosse uma regressão, o CART relaciona uma única variável dependente (a inadimplência ou a PCI, em nosso caso) a um conjunto de variáveis preditivas. Um exemplo dessa estrutura em forma de árvore consta da Figura 6. O CART primeiro divide as contas entre as que têm LTV superior e inferior a 54,5%. Assim, pode-se projetar um modelo de PI para as contas com LTV abaixo desse valor de corte e outro para as contas com LTV maior. Caso haja tempo, poder-se-ia considerar o uso do *score* e da renda para definir ainda melhor o esquema de segmentação.

Às vezes é possível usar a análise CART diretamente num modelo de regressão logística. Esse processo é muito mais rápido do que construir e implementar modelos de regressão separados para cada segmento. Como, então, integrar a metodologia CART a um só modelo de PI? A resposta é: construir uma variável que reflita a estrutura da árvore para cada um de seus ramos. A Figura 6 mostra que a estrutura da árvore produz cinco nós. Cada nó representa o fim da lógica de divisão e o ramo de segmentação da árvore.

Assim, para criar uma variável de segmentação CART para nosso modelo de regressão, podemos usar um código como o seguinte:

IF LTV<.545 AND SCORE<651.6 THEN NODE=1;

IF LTV<.545 AND SCORE>651.5 THEN NODE=2;

Etc.

O mesmo tipo de codificação seria usado para os nós de 3 a 5. Para introduzir esse conteúdo informacional na regressão logística, criaríamos variáveis *dummy* para cada nó e as usaríamos como variáveis preditivas no modelo de PI.

Sumário

Extrair o máximo dos dados disponíveis é um dos maiores desafios com que se depara o construtor de modelos. Qualquer um pode construir um modelo. O desafio está em construir o modelo mais preditivo possível. Este artigo ofereceu algumas dicas sobre como "construir uma boa armadilha" e como proceder de maneira baseada em abordagens estatísticas sólidas que devem satisfazer a maioria dos reguladores. Por outro lado, é bom lembrar que construir modelos de PI e PCI é mais uma arte do que uma ciência. Os modeladores de hoje vêm de uma ampla gama de disciplinas, têm diferentes históricos acadêmicos e experiências e estão acostumados ao uso de determinadas ferramentas estatísticas. Independentemente dessas diferenças, não há substituto para o bom conhecimento dos próprios dados. Entender como foram desenvolvidos, que valores são ou não razoáveis como a empresa funciona é essencial para o processo de desenvolvimento de modelos.

create a CART segmentation variable for our regression model, we would code something like this:

IF LTV<.545 AND SCORE<651.6 THEN NODE=1;

IF LTV<.545 AND SCORE>651.5 THEN NODE=2;

Etc.

The same type of coding would be done for nodes 3-5. To introduce this information content into logistic regression, we would create dummy variables for each node and use them as predictors in our PD model.

Summary

Getting the most out of the data is one of the biggest challenges facing the model builder. Anyone can build a model. However, the challenge is building the most predictive model possible. This article has offered some tips on "building a better mouse-trap" and how to go about it in a way that is based on sound statistical approaches that should satisfy most regulators. On the other hand, it must be remembered that building PD and LGD models is more an art than a science. Today's model builders come from a variety of disciplines, have different academic training and experience, and are accustomed to using certain statistical tools. Regardless of these differences, there is simply no substitute for knowing your data. Understanding how it was developed, what values are reasonable, and how the business works is essential in the model development process.

References

[1]*MORRISON Jeffrey S., "Preparing for Modeling Requirements in Basel II, Part 1: Model Development," The RMA Journal. May 2003.*

[2]*HOSMER David and LEMESHOW Stanley. Applied Logistic Regression, 1989 John Wiley & Sons, Inc.*

[3]*NELSON Bryan D., "Variable Reduction for Modeling Using PROC VARCLUS," Paper 261-26. SAS User-group publication.*

2004 RMA. Jeffrey S. Morrison was vice-presidente Credit Metrics-PRISM Team, at Suntrust Banks Inc., Atlanta, Georgia. Morrison is currently senior manager Modeling Services for Transunion LLP in the Atlanta Georgia office. Transunion builds modeling solutions for both credit risk and marketing applications in addition to their core credit bureau products.
Contact Morrison at m_jeffer@bellsouth.net
RMA - Risk Management Association is an international association of financial services professionals. For membership information, e-mail acauley@rmahq.org ; to subscribe to The RMA Journal, visit www.rmahq.org/Ed_Opps/pubs/journalad.htm

Referências

[1]MORRISON Jeffrey S., "Preparing for Modeling Requirements in Basel II, Part 1: Model Development," The RMA Journal. May 2003.

[2]HOSMER David e LEMESHOW Stanley. Applied Logistic Regression, 1989 John Wiley & Sons, Inc.

[3]NELSON Bryan D., "Variable Reduction for Modeling Using PROC VARCLUS," Paper 261-26. Publicação para grupos de usuários SAS.

2004 RMA. Jeffrey S. Morrison foi vice-presidente de Medidas de Crédito – Equipe PRISM do Suntrust Banks Inc., Atlanta, Georgia. Atualmente ele é gerente sênior de Serviços de Modelagem do Transunion LLP em Atlanta, na Georgia. A Transunion constrói soluções em modelagem tanto para risco de crédito como para aplicações em marketing em seu escritório central de produtos de crédito.
Os contatos com Jefferson Morrison podem ser feitos pelo E-mail m_jeffer@bellsouth.net
A RMA - Risk Management Association é uma associação internacional de serviços financeiros profissionais. Para informações, e-mail acauley@rmahq.org ; Para assinar The RMA Journal visite o site www.rmahq.org/Ed_Opps/pubs/journalad.htm