explained by each variable and the distance of the variable from the next cluster, or a combination of the two measures. In addition, business considerations should also be used in selecting variables from this exercise, so that the final variables chosen are consistent with business reality.

PROC VARCLUS is better than using simple correlation figures, as it considers collinearity as well as correlation, and is therefore a better approach to choosing variables for scorecard development. This is consistent with the overall objective, which is the development of a scorecard, not just a correlation exercise.

Multicollinearity (MC), is not a significant concern when developing models for predictive purposes with large datasets. The effects of MC in reducing the statistical power of a model can be overcome by using a large enough sample such that the separate effects of each input can still be reliably estimated. In this case, the parameters estimates obtained through Ordinary Least Squares (OLS) regression will be reliable.[2]

Identifying correlation can be performed before or after initial characteristic analysis, but before the regression step. Both the correlation and grouping steps provide valuable information on the data at hand, and are more than just statistical exercises. While reducing the number of characteristics to be grouped (by checking for correlation first) is a time saver, one is also deprived of an opportunity to look at the nature of the relationship between many characteristics and performance. Therefore, the best approach is likely a combination of eliminating some characteristics and choosing more than one characteristic from each correlated "cluster" based on business and operational intuition. This serves to balance the need for efficiency with the opportunity to gain insights into the data.

Initial Characteristic Analysis

Initial characteristic analysis involves two main tasks. The first step is to assess the strength of each characteristic individually as a predictor of performance. This is also known as univariate screening, and is done to screen out weak or illogical characteristics.

The strongest characteristics are then grouped. This applies to

attributes in both continuous and discrete characteristics, and is done for an obvious reason. The grouping is done because it is required to produce the scorecard format shown in Exhibit 1.1.

Scorecards can also be, and are, produced using continuous (ungrouped) characteristics. However, grouping them offers some advantages:

- It offers an easier way to deal with outliers with interval variables, and rare classes.
- Grouping makes it easy to understand relationships, and therefore gain far more knowledge of the portfolio. A chart displaying the relationship between attributes of a characteristic and performance is a much more powerful tool than a simple variable strength statistic. It allows users to explain the nature of this relationship, in addition to the strength of the relationship.
- Nonlinear dependencies can be modeled with linear models.
- It allows unprecedented control over the development process—by shaping the groups, one shapes the final composition of the scorecard.
- The process of grouping characteristics allows the user to develop insights into the behavior of risk predictors and increases knowledge of the portfolio, which can help in developing better strategies for portfolio management.

Once the strongest characteristics are grouped and ranked, variable selection is done. At the end of initial characteristic analysis, the Scorecard Developer will have a set of strong, grouped characteristics, preferably representing independent information types, for use in the regression step.

The strength of a characteristic is gauged using four main criteria:

- Predictive power of each attribute. The weight of evidence (WOE) measure is used for this purpose.
- The range and trend of weight of evidence across grouped attributes within a characteristic.

- Predictive power of the characteristic. The Information Value (IV) measure is used for this.
- Operational and business considerations (e.g., using some logic in grouping postal codes, or grouping debt service ratio to coincide with corporate policy limits).

Some analysts run other variable selection algorithms (e.g., those that rank predictive power using Chi Square or R-Square) prior to grouping characteristics. This gives them an indication of characteristic strength using independent means, and also alerts them in cases where the Information Value figure is high/low compared to other measures.

The initial characteristic analysis process can be interactive, and involvement from business users and operations staff should be encouraged. In particular, they may provide further insights into any unexpected or illogical behavior patterns and enhance the grouping of all variables.

The first step in performing this analysis is to perform initial grouping of the variables, and rank order them by IV or some other strength measure. This can be done using a number of binning techniques. In SAS Credit Scoring, the Interactive Grouping Node can be used for this.

If using other applications, a good way to start is to bin nominal variables into 50 or so equal groups, and to calculate the WOE and IV for the grouped attributes and characteristics. One can then use any spreadsheet software to fine-tune the groupings for the stronger characteristics based on principles to be outlined in the next section. Similarly for categorical characteristics, the WOE for each unique attribute and the IV of each characteristic can be calculated. One can then spend time fine-tuning the grouping for those characteristics that surpass a minimum acceptable strength. Decision trees are also often used for grouping variables. Most users, however, use them to generate initial ideas, and then use alternate software applications to interactively fine-tune the groupings.

Statistical Measures

Exhibit 6.2 shows a typical chart used in the analysis of grouped characteristics. The example shows the characteristic "age" after it has been

EXHIBIT 6.2 ANALYSIS OF GROUPED VARIABLES

Age	Count	Tot Distr	Goods	Distr Good	Bads	Distr Bad	Bad Rate	WOE
Missing	1,000	2.50%	860	2.38%	140	3.65%	14.00%	−42.719
18–22	4,000	10.00%	3,040	8.41%	960	25.00%	24.00%	−108.980
23–26	6,000	15.00%	4,920	13.61%	1,080	28.13%	18.00%	−72.613
27–29	9,000	22.50%	8,100	22.40%	900	23.44%	10.00%	−4.526
30–35	10,000	25.00%	9,500	26.27%	500	13.02%	5.00%	70.196
35–44	7,000	17.50%	6,800	18.81%	200	5.21%	2.86%	128.388
44+	3,000	7.50%	2,940	8.13%	60	1.56%	2.00%	164.934
Total	**40,000**	**100%**	**36,160**	**100%**	**3,840**	**100%**	**9.60%**	

Information Value = 0.668

grouped. In the exhibit, "Tot Distr," "Distr Good," and "Distr Bad" refer to the column-wise percentage distribution of the total, good, and bad cases, respectively. For example, 17.5% of all cases, 18.81% of goods, and 5.21% of bads fall in the age group 35–44.

A few things to note in Exhibit 6.2:

- "Missing" is grouped separately. The weight of this group implies that most of the missing data comes from an age group between 23 and 29.
- A general "minimum 5% in each bucket" rule has been applied to enable meaningful analysis.
- There are no groups with 0 counts for good or bad.
- The bad rate and WOE are sufficiently different from one group to the next (i.e., the grouping has been done in a way to maximize differentiation between goods and bads). This is one of the objectives of this exercise—to identify and separate attributes that differentiate well. While the absolute value of the WOE is important, the difference between the WOE of groups is key to establishing differentiation. The larger the difference between subsequent groups, the higher the predictive ability of this characteristic.
- The WOE for nonmissing values also follows a logical distribution, going from negative to positive without any reversals.

The WOE, as mentioned previously, measures the strength of each attribute, or grouped attributes, in separating good and bad accounts. It is a measure of the difference between the proportion of goods and bads in each attribute (i.e., the odds of a person with that attribute being good or bad). The WOE is based on the log of odds calculation:

$$\text{(Distr Good / Distr Bad)}$$

which measures odds of being good (e.g., for the 23–26 attribute above, this would be 13.61/28.13 = 0.48). A person aged 23–26 has 0.48:1 odds of being good.

A more user-friendly way to calculate WOE, and one that is used in Exhibit 6.2, is:

$$[\, ln\left(\frac{\text{Distr Good}}{\text{Distr Bad}}\right)] \times 100.$$

For example, the WOE of attribute 23–26 is:

$$ln\left(\frac{0.1361}{0.2813}\right) \times 100 = -72.613.$$

Multiplication by 100 is done to make the numbers easier to work with. Negative numbers imply that the particular attribute is isolating a higher proportion of bads than goods.

Information Value, or total strength of the characteristic, comes from information theory,[3] and is measured using the formula:

$$\sum_{i=1}^{n} (\text{Distr Good}_i - \text{Distr Bad}_i) * ln\left(\frac{\text{Distr Good}_i}{\text{Distr Bad}_i}\right)$$

Note that "Distr Good" and "Distr Bad" are used in this formula in decimal format, for example, 0.136 and 0.28.

Based on this methodology, one rule of thumb regarding IV is:

- Less than 0.02: unpredictive
- 0.02 to 0.1: weak
- 0.1 to 0.3: medium
- 0.3 +: strong

Characteristics with IV greater than 0.5 should be checked for over-predicting—they can either be kept out of the modeling process, or used in a controlled manner, such as will be described later in the "Preliminary Scorecard" section.

IV is a widely used measure in the industry, and different practitioners have different rules of thumb regarding what constitutes weak or strong characteristics.

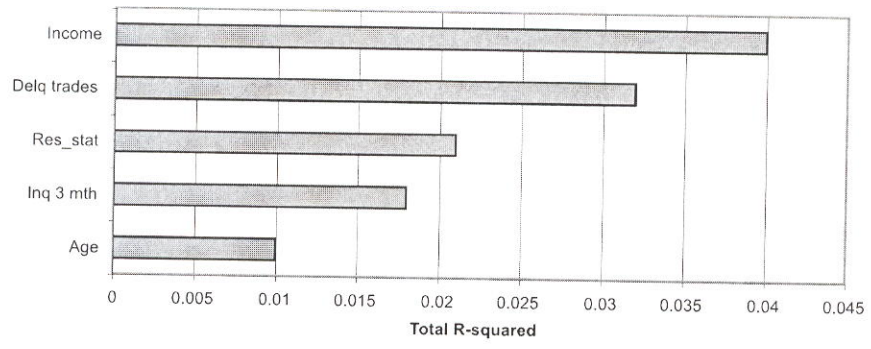
Where the scorecard is being developed using nongrouped characteristics, statistics to evaluate predictive strength include R-square and Chi-square. Both these methods use goodness-of-fit criteria to evaluate characteristics. The R-squared technique uses a stepwise selection method that rejects characteristics that do not meet incremental R-square increase cutoffs. A typical cutoff for stepwise R-squared is 0.005. Chi-square operates in a similar fashion, with a minimum typical cutoff value of 0.5. The cutoffs can be increased if too many characteristics are retained in the model. As with the technique using grouped variables, the objective here is to select characteristics for regression (or another modeling step).

Again, it is important to note that univariate screening, whether using grouping or not, does not account for partial associations and interactions among the input characteristics. Partial association occurs when the effect of one characteristic changes in the presence of another. Multivariate methods that consider joint subsets may be preferable in this case. In any case, the purpose of doing the exercise is the same—choosing a set of strong variables for input into regression (or another technique, as appropriate).

Some modeling software offers options to group characteristics for the R-square and Chi-square methods, and to test interactions for categorical inputs. Examples of two-way interactions that can be tested are income*residential status, age*income, and so forth. This methodology goes beyond individual characteristic analysis and can produce more powerful results by considering interactions between characteristics. Interaction terms are also a way of dealing with segmentation.

A typical output from an R-square analysis is shown in Exhibit 6.3, where the incremental increase in R-square value is shown as characteristics are added to the model starting with age and ending with income.

EXHIBIT 6.3 MODEL CHARACTERISTICS



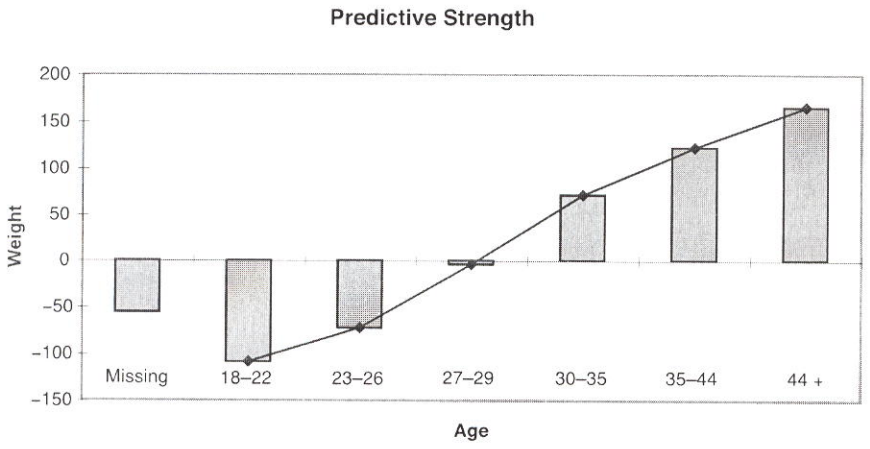
Logical Trend

The statistical strength, measured in terms of WOE and IV, is, however, not the only factor in choosing a characteristic for further analysis, or designating it as a strong predictor. In grouped scorecards, the attribute strengths must also be in a logical order, and make operational sense. For example, the distribution of attribute weight for age, from Exhibit 6.2, is plotted in Exhibit 6.4.

As can be clearly seen, apart from "missing," the other groupings in this characteristic have a linear relationship with WOE; that is, they denote a linear and logical relationship between the attributes in age and proportion of bads. This confirms business experience both in the credit and insurance sectors that younger people tend to be, in general, of a higher risk than the older population. Establishing such logical (not necessarily linear) relationships through grouping is the purpose of the initial characteristic analysis exercise. The process of arriving at a logical trend is one of trial and error, in which one balances the creation of logical trends while maintaining a sufficient IV value.

Experimenting with different groupings mostly eliminates reversals (where the trend reverses itself) and other illogical relationships. General trends can be seen by looking at the relationship between WOE and raw (ungrouped) attributes—grouping merely smoothes out

EXHIBIT 6.4 LOGICAL WOE TREND FOR AGE



the curve. In some cases, however, reversals may be reflecting actual behavior or data, and masking them can reduce the overall strength of the characteristic. These should be investigated first, to see if there is a valid business explanation for such behavior. In general, grouping serves to reduce "overfitting," whereby quirks in the data are modeled rather than the overall trend in predictiveness. Where valid nonlinear relationships occur, they should be used if an explanation using experience or industry trends can be made. Again, what needs to be confirmed is that an overall trend or profile is being modeled, and not data quirks. Business experience is the best test for this. For example, in North America, "revolving open burden" (utilization on revolving trades) has a banana-shaped curve with respect to WOE. Very low utilization accounts are higher risk, then the risk decreases up to a point, and finally risk starts increasing as utilization increases. Other valid relationships may be "U" shaped, and these should be kept as that, as long as the relationship can be explained.

Nominal variables are grouped to put attributes with similar WOE together, and, as with continuous variables, to maximize the difference from one group to the next.

Clearly, this process can be abused when it is done by someone who is not familiar with the business, which again underscores the need for it to be a collaborative process with other project team members.

Exhibit 6.5 illustrates an example of an illogical trend. In this particular dataset, this characteristic is weak and shows no logical relationship between age and good/bad performance.

Exhibit 6.6 shows two WOE relationships, both of which are logical. However, the steeper line (square markers) represents a stronger predictive relationship between age and performance. This will be reflected in its IV number.

Initial characteristic analysis involves creating business logical relationships through grouping of attributes that exceed minimum IV criteria. The alternate, purely statistical approach involves establishing relationships that only maximize IV or other measures, whether grouped or not. The business-based approach is better for several reasons, including:

- Logical relationships ensure that the final weightings after regression make sense. This also ensures that when attributes are allocated points to generate a scorecard, these points are logical

EXHIBIT 6.5 ILLOGICAL WOE TREND FOR AGE

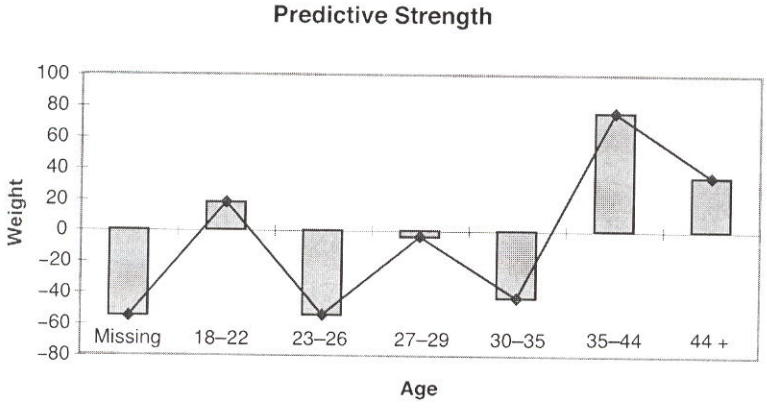
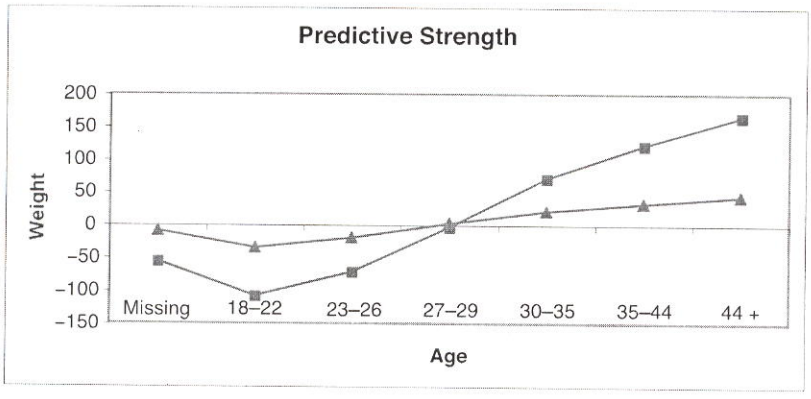


EXHIBIT 6.6 LOGICAL TREND AND STRENGTH



(e.g., an older person gets higher points than a younger person always).

- Logical relationships ensure buy-in from internal end users and operations departments. When the scorecard confirms general experience, it provides a higher level of confidence in automated decision making.
- Logical relationships confirm business experience, thus going one step further than a purely statistical evaluation. This allows the usage of business experience to enhance predictive modeling, and makes it relevant to business usage.
- Most important, generalizing relationships by grouping them in a logical fashion reduces overfitting. You are no longer modeling every quirk in the data by assigning an unlimited number of weights to ungrouped attributes. You are now risk ranking and modeling trends, so that the scorecard can now be applied to an incoming population with some elasticity (able to withstand some changes in the population), and that will remain stable for a longer period of time. A legitimate concern here would be that of over-generalization, whereby the model will seem to work even when the changes in the population dictate otherwise. The solution to

this issue is to build a widely based risk profile, and not a scorecard with a limited number of characteristics. The long-term application differentiates credit risk scorecard development from marketing models, which are often built for specific campaigns and then discarded. Therefore, one cannot afford to model quirks.

Business/Operational Considerations

Statistical considerations and business logic have been discussed as measures used to group attributes. The third consideration is business or operational relevance.

For nonnumerical discrete—that is, nominal—values, such as postal codes or lifestyle code, the groupings are normally done based on similar weights to produce a logical trend (i.e., attributes with similar weights are grouped together). Groupings should also be investigated based on provincial, regional, urban/rural, and other operational considerations such as corporate business regions. For example, if you are building a scorecard to predict default for mortgages, grouping of postal codes should be done by similar real estate markets. It may be that the risk associated with borrowers is dependent on the real estate market, which tends to differ in large urban areas and rural areas. For example, in the United States, it may not make sense to group by New York or California as state or region, as the housing market is not uniform. Grouping New York City, Los Angeles, and San Francisco together, and rural areas in both states together, makes far more sense.

In some cases it also makes sense to have breaks concurrent with policy rules. For example, if a company policy requires loans with debt service ratios greater than 42% to be referred, then that debt service ratio should be grouped with a break at 42%. The benefits of grouping in such a way is that the distortion caused by the policy rule on the scorecard is minimized, since those affected by the policy rule are now isolated somewhat. Such groupings can also test conventional wisdom and previous policies—for example, to see if the 42% rule makes sense at that point, or if it would be better situated at a higher debt service ratio to maximize risk discrimination.