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