Paper 6

An Interactive GUI Front-End for a Credit Scoring Modeling System

Jeffrey Morrison, Futian Shi, and Timothy Lee
Knowledge Sciences & Analytics,
Equifax Credit Information Services, Inc.

Abstract
The need for statistical modeling has been on the rise in recent years. Banks, retailers, and even telecommunication companies are looking for better ways to reduce risks from new and existing customers. Software technology like SAS AF/FRAMES has stepped up to help meet this demand through GUI front end development, reducing programming errors and learning curves for new employees.

Consulting companies providing modeling services will be most successful if they can efficiently integrate these technologies with their model building processes. The following paper discusses how the development of a AF/Frames application using SCL, ACCESS, Share, and Graphics enabled ***Knowledge Sciences & Analytics*** *to merge these technologies into a client server platform aiding the following management initiatives…*

- *Reducing the Impact of Employee Turnover*
- *Ramping Up Training of New Employees*
- *Increasing Efficiency*
- *Integrating New Ideas Quickly*
- *Developing a Modeling Database*
- *Standardizing Modeling Methodologies*

The modeling system, through a collection of clickable buttons, list boxes, and other objects dynamically generates SAS code, enabling the user to interactively make modeling decisions based upon statistical output. The foundation of the system is heavily built around DATA LIST STRUCTURES which allow for more complex and efficient programming designs needed for the model building process.

Background and Motivation:
Knowledge Sciences & Analytics is the Consulting and Analytic branch of Equifax, Inc., one of the leading repositories of credit information in the U.S.. The group's model development process involves frequent interaction between its technical consultants, project managers, statistical model builders and the customer's project management team. Many of these models use credit scoring (classification) techniques which are specially designed for models using a zero / one dependent variable. For example, the three most common models tend to be…

- Generic and custom risk assessment models for account acquisition and account management

- Bankruptcy predictors
- Response to pre-approved credit offerings

Model development within the organization has historically been done by about 10-20 statistical analysts using work stations tied to a SUN platform through a UNIX operating system. The primary engine for statistical analysis and model building is base SAS. For the statistician, there are a number of SAS "standard programs" available for use. These programs reflect code written by a variety of individuals over the years for model development. However, these programs are essentially open code and require special editing and text substitutions at different locations before they can be used on new projects. Output from one program is often needed as input from others. Hundreds of predictor variables from the credit file are available for use in the modeling process requiring extensive storage capacity on the SUN system.

Although the majority of the programs reside in a common directory, many statisticians have made their own "versions" of these programs because (a) they couldn't get the original programs to work the way they wanted (b) didn't understand or need various portions of the code (c) varying levels of SAS programming background (d) they wanted the programs to do a few things a little different. Because of a lack of a true standard platform, modeling methodologies had the tendency to diverge and become inconsistent, depending on the direction of various managers. In additions, statisticians had a difficult time getting new modeling ideas integrated into the process because no true standardized platform existed. Often, analysts "reinvented the wheel", operating independently from one another. Combined with the employee turnover associated with the industry, management became concerned that the organization could be losing intellectual knowledge and efficiency surrounding some of the standard programs and what essential changes were necessary to make them work properly.

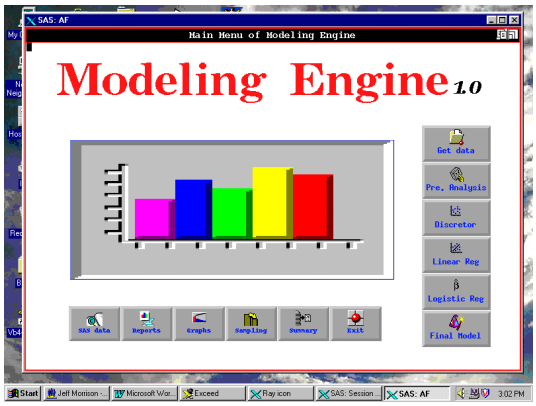
Features & Benefits:

As a result, a process improvement plan was proposed to create a GUI interface that, among other things, should perform text substitutions (passing parameters) to and from the existing standard programs and perform specialized interactive routines that assist the user in making modeling decisions. The system was designed to…

- More fully standardize programs and methodologies
- Create a platform where new programs and ideas can be universally deployed throughout the department.
- Increase process efficiency and reduce ad hoc programming.
- Minimize the impact of turnover by retaining intellectual knowledge.
- Allow more time for analysis by minimizing errors of code editing.
- Construct a database to retain key modeling statistics, dataset names, variables used, system generated SAS code, etc..
- Reduce learning curves for new statisticians who may be trained in other software packages.

The Application:

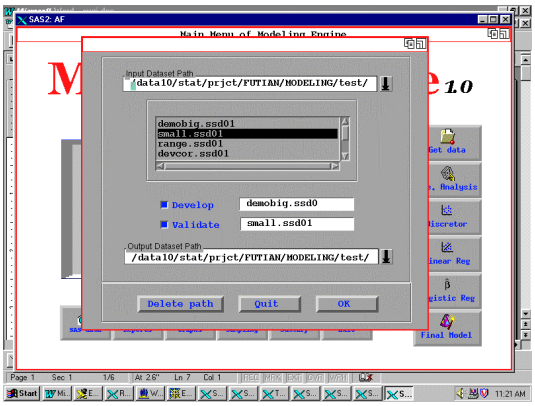
I). Main Menu



The main menu consists of two sets of toolbars which help direct the model building process. The vertical toolbar contains 6 buttons which should be followed in sequence to complete a model for a dichotomous dependent variable. The horizontal toolbar contains buttons representing special utilities used at various times in the model building process. For example, these built-in utilities enable the user to sample down or to navigate SAS data set before starting a modeling process, to view or print any external flat files (such as *.sas, *.log, and *.lst) instantly through the catalog output object.

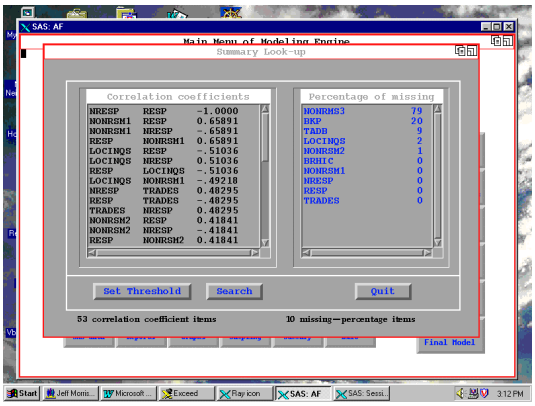
II) Input File Name(s)

The starting point for model development begins with the first button of vertical toolbar on the Main Menu. It is based on the assumption that the data is already in a SAS ready format (*.ssd01). After the user specifies the path name of the UNIX directory using an *extended input box*, a list of files with a SAS dataset extension is displayed.



The analyst simply chooses which dataset is desired for model development and which is to be used for validation. Once these choices are made, the other buttons on the MAIN MENU become active and available for use.

III). Preliminary Analysis

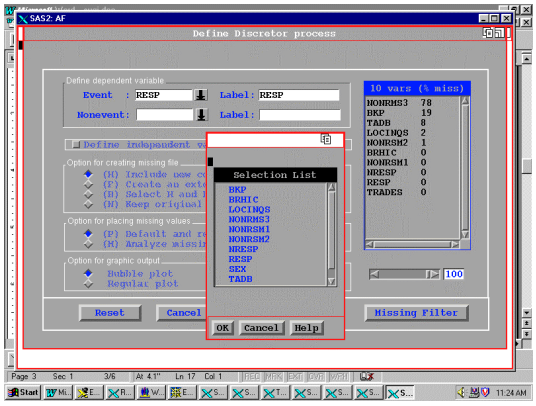


The next step would be to perform a preliminary analysis on the data - i.e. evaluate missing values and create pairwise and bivariate correlations. The second button on the vertical toolbar performs this function requiring no interaction from the user. This is an example of the system's ability to launch a stand alone program in a batch environment. Once the program is finished running, a message is sent to the user via UNIX mail indicating

completion. Output may be viewed through the Summary button in the utility toolbar through the use of 2 *list boxes*. Clickable options allow the user to display portions of the information based on % missing or correlation thresholds. String search routines are also available to help the user locate specific information quickly and efficiently.

IV). Outlier and Data Smoothing

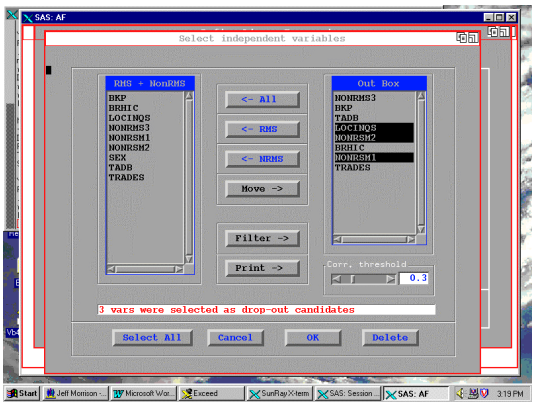
The next step is to launch a data smoothing algorithm which minimizes the impact of outliers, resulting in a more accurate model. The input required from the user is captured through a set of check boxes, radio boxes, list boxes, and other objects which interactively scan the data file for variable names and information from the preliminary analysis.



For example, the user can select which variables to apply the smoothing algorithm to based on criteria such as percent missing. If the variable has a large percentage of the information missing, then efficiency gains could be realized by eliminating it from the data smoothing and model building process. The process is made user friendly through a screening filter (horizontal scroll bar) which accesses missing value information and eliminates all variables over a certain missing threshold.

V). Regression Based Modeling

Model development consists of interactively selecting a set of predictor variables from a pool of about 500 through a mix of various stepwise regression procedures. These procedures are made available to the user through buttons on the vertical toolbar object. Again, through a combination of list boxes and check boxes, the type of procedure is selected along with p-values and type of regression. Furthermore, an interactive session may be selected which makes recommendations as to which predictor variables to exclude from the regression because of collinearity issues.



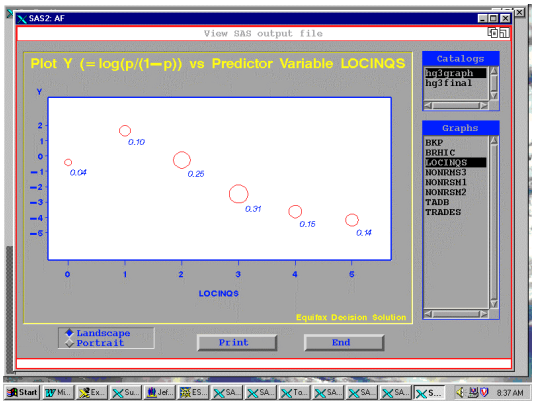
This session launches a background program which takes the user defined subset of variables and determines which variables are correlated above a certain threshold. These correlated variables are then compared one by one to the dependent variable. The variable that is least correlated to the dependent variable is then highlighted in the list box as a possible drop candidate.

As with the preliminary analysis, the completion of the regression run is signaled by a mail message being sent to the user through UNIX. Using the utility buttons on the main menu, the analyst can view the SAS code that was dynamically generated and examine the .log and .lst files produced in the batch environment. Upon examination of the .lst file, the program will do a string search for the word “ERROR” in the .log file indicating the successful completion of the. If an error was found, the program would open a frame window and pull in a copy of the log file, automatically bringing the user to the location of the first occurrence of the error. Otherwise, a “nobug” icon appears on the screen, indicating the successful completion of the program.

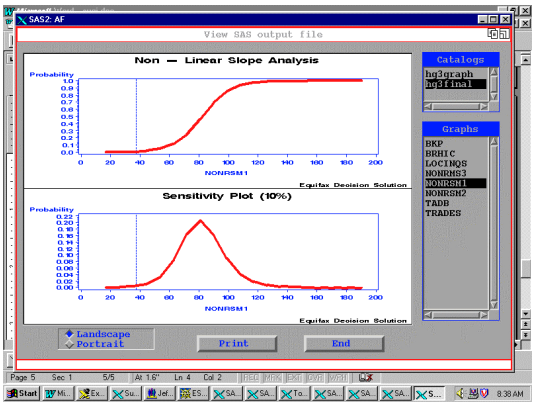
Since modeling is not an exact science, the analyst would have to estimate a number of models and compare results to decide on the optimal specification. The GUI system is designed to keep track of multiple runs for the same procedure. Every time a certain function is submitted, the output file names would be tracked by attaching a counter variable to the file name. The user could then compare the current model with historical results.

VI). Graphics

During the modeling process, a graphics utility becomes available, grouped in a graphical catalog object. If the first catalog is chosen, then a list box appears with all variables to which the data smoothing algorithm was applied. As each variable is selected from the list box, a bubble graph is displayed (using SAS GRAPHICS) showing the relationship between the logit of the dependent variable and discretized values of each predictor variable.



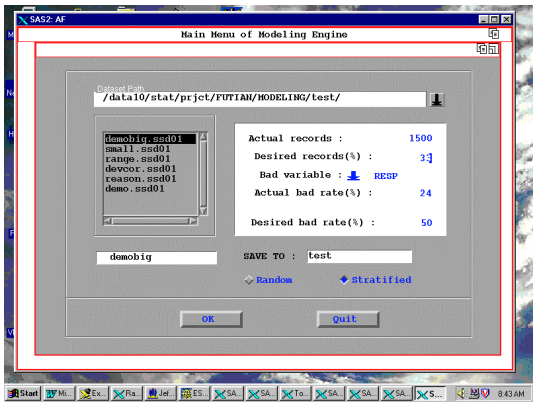
The size of the bubbles indicate the percentage of observations that fall into the discretized interval. If the second catalog is selected, a list box appears containing two graphs.



The top graph shows the relationship between the model ‘s predicted probabilities across the range of the variable selected. The bottom graph shows the areas along the variable’s range where the model is most sensitive to slight changes in the predictor variable.

VII). Sampling Utilities

In the world of modeling, sample size and design is extremely important. The modeling platform provides the user with an interactive method for creating additional samples from the original SAS dataset - i.e. smaller subsets of the original data.



The user can select from two sampling schemes - random or stratified. In response modeling, for example, the number of observations for the "event" variable is usually scarce. Therefore, the analyst may wish to create a stratified sample where there are more "event" values than in the real population. Through list boxes and other objects, the system prompts the user to pick the stratification variable and automatically calculates the maximum possible event percentages.

VIII). Database Information

Finally, after the final model is selected, the user clicks the last button on the vertical toolbar. This procedure writes out specific information needed in the auditing process and populates an Oracle database (SAS ACCESS) with a variety of information. This could range from the type of model estimated, the specific variables used, the analyst's i.d., statistical measures of precision, file names, .sas batch files, etc. As a result, management has the ability to collect historical information about the modeling effort over time and determine realistic performance benchmarks across a variety of modeling applications.

IX. Concluding Remarks

As in any process improvement plan, especially those involving computer technology, full management support and involvement is essential. Management needs to clearly state the goal and importance of the process improvements at all levels and ensure that users become involved in system design and debugging. Guidelines and incentives also need to be implemented to encourage proper use of the system and a continual feedback process for improvement. Therefore, the system described in this paper provides only a first generation modeling platform. Improvements, suggestions, and ways to enhance the system and make it more flexible are already underway.

Authors' Addresses

Knowledge Sciences & Analytics
Equifax Credit Information Services, Inc.
1525 Windward Concourse
Mail Drop 42-S
Alpharetta, Georgia 30005
Email: jeff.morrison@equifax.com

Trademarks