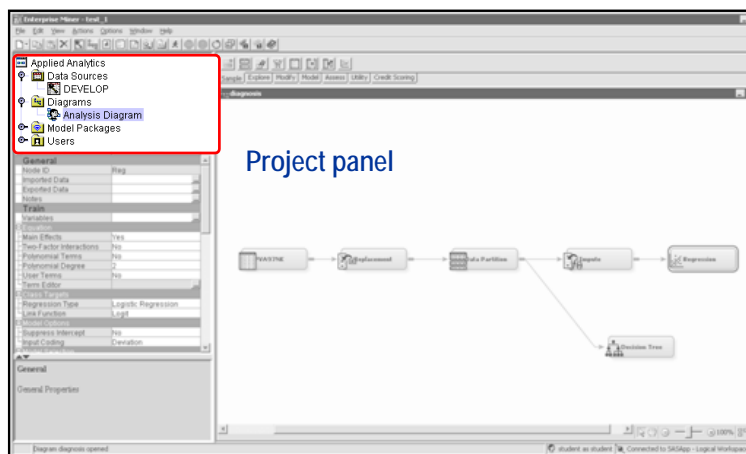


Chapter 1 Introduction

0.1	Introduction.....	Error! Bookmark not defined.
0.2	A Section Title.....	Error! Bookmark not defined.
	Demonstration: <Type title of demo here.>	Error! Bookmark not defined.
	Exercises	Error! Bookmark not defined.
0.3	Chapter Summary.....	Error! Bookmark not defined.
0.4	Solutions	Error! Bookmark not defined.
	Solutions to Exercises	Error! Bookmark not defined.
	Solutions to Student Activities (Polls/Quizzes)	Error! Bookmark not defined.

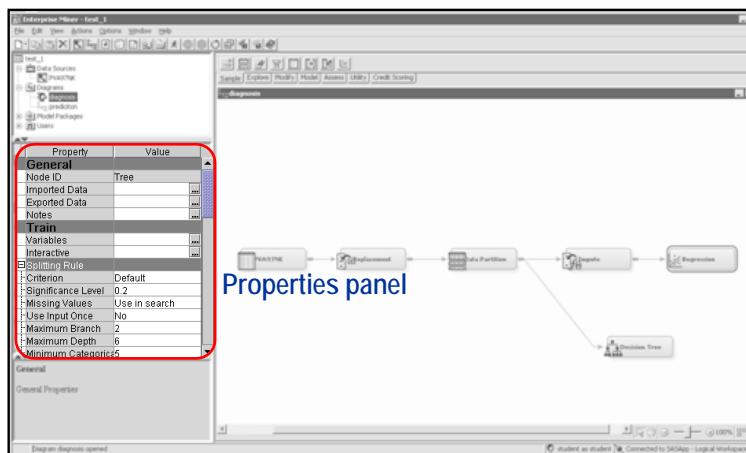
SAS Enterprise Miner – Interface Tour



4

The *Project panel* manages and views data sources, diagrams, results, and project users.

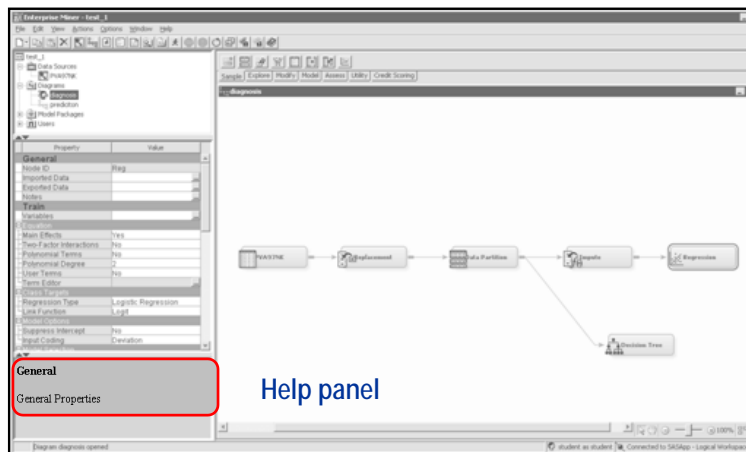
SAS Enterprise Miner – Interface Tour



5

The *Properties panel* enables you to view and edit the settings of data sources, diagrams, nodes, results, and users.

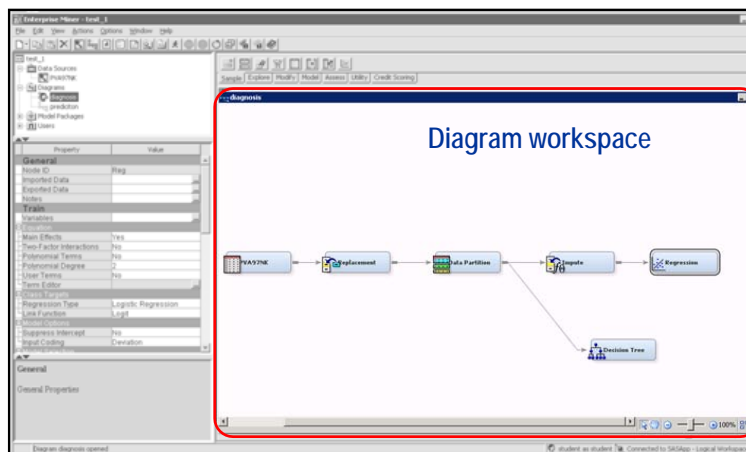
SAS Enterprise Miner – Interface Tour



6

The *Help panel* displays a short description of the property that you select in the Properties panel. Extended help can be found in the Help Topics selection from the Help main menu.

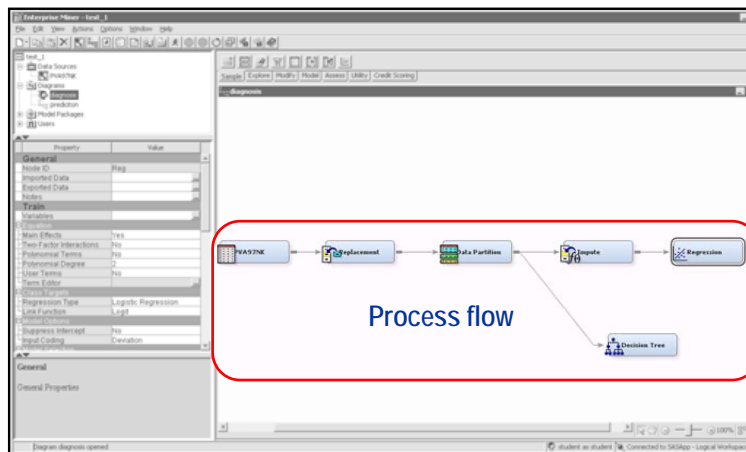
SAS Enterprise Miner – Interface Tour



7

In the *Diagram workspace*, process flow diagrams are built, edited, and run. The workspace is where you graphically sequence the tools that you use to analyze your data and generate reports.

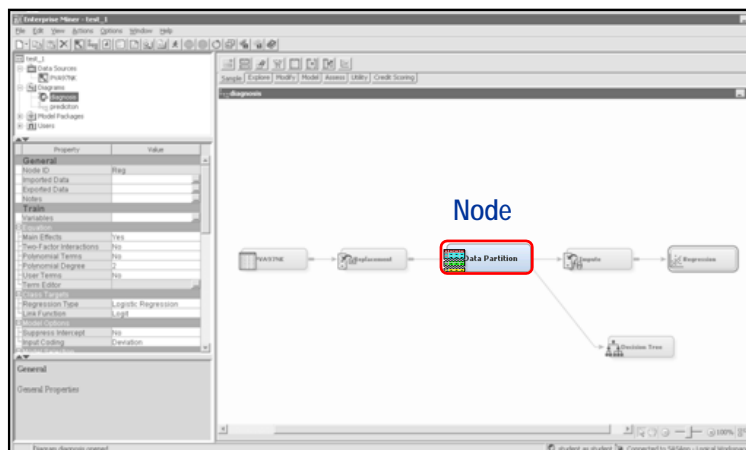
SAS Enterprise Miner – Interface Tour



8

The Diagram workspace contains one or more process flows. A *process flow* starts with a data source and sequentially applies SAS Enterprise Miner tools to complete your analytic objective.

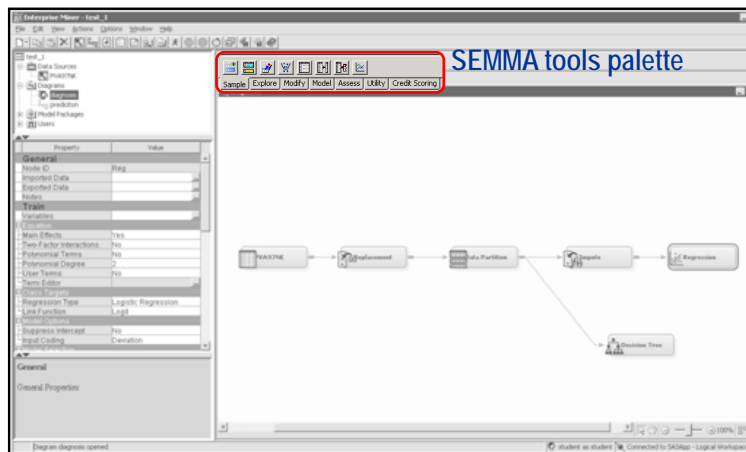
SAS Enterprise Miner – Interface Tour



9

A process flow contains several nodes. *Nodes* are SAS Enterprise Miner tools connected by arrows to show the direction of information flow in an analysis.

SAS Enterprise Miner – Interface Tour



10

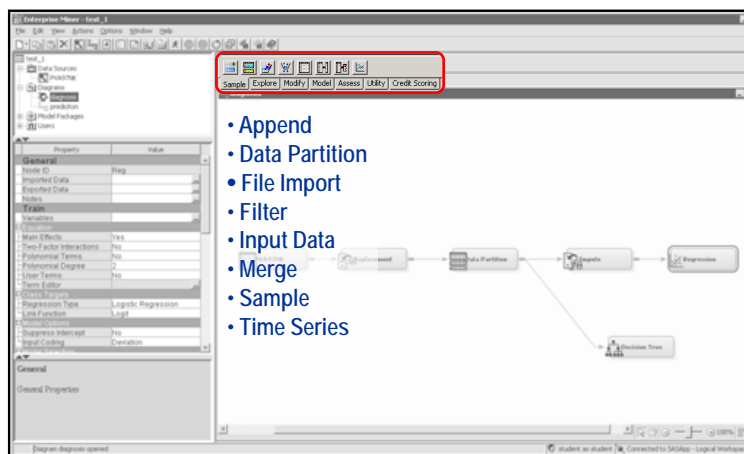
The SAS Enterprise Miner tools available to your analysis are contained in the *tools palette*. The tools palette is arranged according to a process for data mining, SEMMA.

SEMMA is an acronym for the following:

- Sample** You sample the data by creating one or more data tables. The samples should be large enough to contain the significant information, but small enough to process.
- Explore** You explore the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas.
- Modify** You modify the data by creating, selecting, and transforming the variables to focus the model selection process.
- Model** You model the data by using the analytical tools to search for a combination of the data that reliably predicts a desired outcome.
- Assess** You assess competing predictive models (build charts to evaluate the usefulness and reliability of the findings from the data mining process).

Additional tools are available under the Utility group and, if licensed, the Credit Scoring group.

SEMMA – Sample Tab



11

The tools in each Tool tab are arranged alphabetically.

The **Append** tool is used to append data sets that are exported by two different paths in a single process flow diagram. The Append node can also append train, validation, and test data sets into a new training data set.

The **Data Partition** tool enables you to partition data sets into training, test, and validation data sets. The *training data set* is used for preliminary model fitting. The *validation data set* is used to monitor and tune the model during estimation and is also used for model assessment. The *test data set* is an additional holdout data set that you can use for model assessment. This tool uses simple random sampling, stratified random sampling, or cluster sampling to create partitioned data sets.

The **File Import** tool enables you to convert selected external flat files, spreadsheets, and database tables into a format that SAS Enterprise Miner recognizes as a data source.

The **Filter** tool creates and applies filters to your training data set, and optionally, to the validation and test data sets. You can use filters to exclude certain observations, such as extreme outliers and errant data that you do not want to include in your mining analysis.

The **Input Data** tool represents the data source that you choose for your mining analysis and provides details (metadata) about the variables in the data source that you want to use.

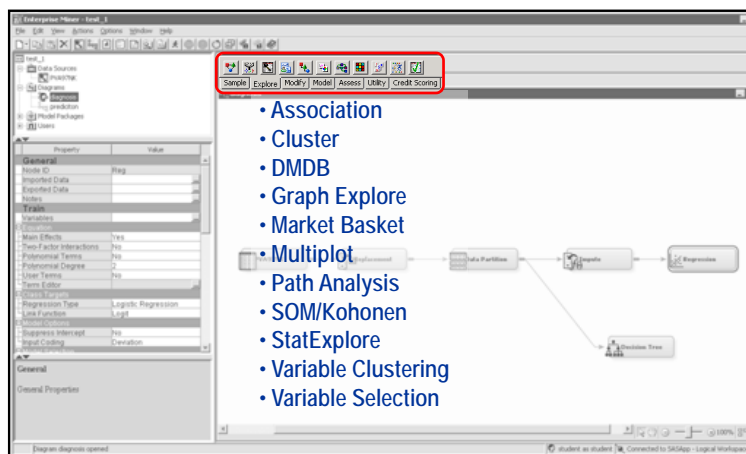
The **Merge** tool enables you to merge observations from two or more data sets into a single observation in a new data set. The Merge tool supports both one-to-one and match merging.

The **Sample** tool enables you to take simple random samples, n^{th} observation samples, stratified random samples, first- n samples, and cluster samples of data sets. For any type of sampling, you can specify either a number of observations or a percentage of the population to select for the sample. If you are working with rare events, the Sample tool can be configured for oversampling or stratified sampling.

Sampling is recommended for extremely large databases because it can significantly decrease model training time. If the sample is sufficiently representative, relationships found in the sample can be expected to generalize to the complete data set. The Sample tool writes the sampled observations to an output data set and saves the seed values that are used to generate the random numbers for the samples so that you can replicate the samples.

The **Time Series** tool converts transactional data to time series data. *Transactional data* is time-stamped data that is collected over time at no particular frequency. By contrast, *time series data* is time-stamped data that is summarized over time at a specific frequency. You might have many suppliers and many customers, as well as transaction data that is associated with both. The size of each set of transactions can be very large, which makes many traditional data mining tasks difficult. By condensing the information into a time series, you can discover trends and seasonal variations in customer and supplier habits that might not be visible in transactional data.

SEMMA – Explore Tab



12

The **Association** tool enables you to perform association discovery to identify items that tend to occur together within the data. For example, if a customer buys a loaf of bread, how likely is the customer to also buy a gallon of milk? This type of discovery is also known as *market basket analysis*. The tool also enables you to perform sequence discovery if a timestamp variable (a sequence variable) is present in the data set. This enables you to take into account the ordering of the relationships among items.

The **Cluster** tool enables you to segment your data; that is, it enables you to identify data observations that are similar in some way. Observations that are similar tend to be in the same cluster, and observations that are different tend to be in different clusters. The cluster identifier for each observation can be passed to subsequent tools in the diagram.

The **DMDB** tool creates a data mining database that provides summary statistics and factor-level information for class and interval variables in the imported data set.

The **Graph Explore** tool is an advanced visualization tool that enables you to explore large volumes of data graphically to uncover patterns and trends and to reveal extreme values in the database. The tool creates a run-time sample of the input data source. You use the Graph Explore node to interactively explore and analyze your data using graphs. Your exploratory graphs are persisted when the Graph Explore Results window is closed. When you reopen the Graph Explore Results window, the persisted graphs are re-created.

The experimental **Market Basket** tool performs association rule mining over transaction data in conjunction with item taxonomy. Transaction data contains sales transaction records with details about items bought by customers. Market basket analysis uses the information from the transaction data to give you insight about which products tend to be purchased together.

The **MultiPlot** tool is a visualization tool that enables you to explore large volumes of data graphically. The MultiPlot tool automatically creates bar charts and scatter plots for the input and target. The code created by this tool can be used to create graphs in a batch environment.

The **Path Analysis** tool enables you to analyze Web log data to determine the paths that visitors take as they navigate through a Web site. You can also use the tool to perform sequence analysis.

The **SOM/Kohonen** tool performs unsupervised learning by using Kohonen vector quantization (VQ), Kohonen self-organizing maps (SOMs), or batch SOMs with Nadaraya-Watson or local-linear smoothing. Kohonen VQ is a clustering method, whereas SOMs are primarily dimension-reduction methods. For cluster analysis, the Clustering tool is recommended instead of Kohonen VQ or SOMs.

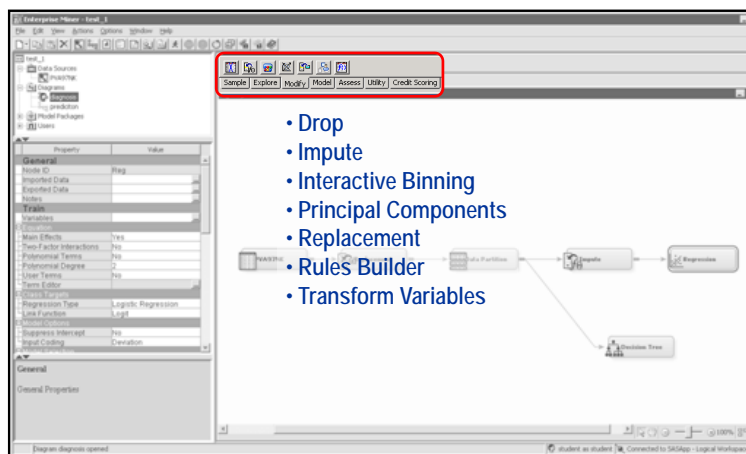
The **StatExplore** tool is a multipurpose tool used to examine variable distributions and statistics in your data sets. The tool generates summarization statistics. You can use the StatExplore tool to do the following:

- select variables for analysis, for profiling clusters, and for predictive models
- compute standard univariate distribution statistics
- compute standard bivariate statistics by class target and class segment
- compute correlation statistics for interval variables by interval input and target

The **Variable Clustering** tool is useful for data reduction, such as choosing the best variables or cluster components for analysis. Variable clustering removes collinearity, decreases variable redundancy, and helps to reveal the underlying structure of the input variables in a data set.

The **Variable Selection** tool enables you to evaluate the importance of input variables in predicting or classifying the target variable. To select the important inputs, the tool uses either an R-squared or a Chi-squared selection criterion. The R-squared criterion enables you to remove variables in hierarchies, remove variables that have large percentages of missing values, and remove class variables that are based on the number of unique values. The variables that are not related to the target are set to a status of *rejected*. Although rejected variables are passed to subsequent tools in the process flow diagram, these variables are not used as model inputs by more detailed modeling tools, such as the Neural Network and Decision Tree tools. You can reassign the input model status to rejected variables.

SEMMA – Modify Tab



13

The **Drop** tool is used to remove variables from scored data sets. You can remove all variables with the role type that you specify, or you can manually specify individual variables to drop. For example, you could remove all hidden, rejected, and residual variables from your exported data set, or you could remove only a few variables that you identify yourself.

The **Impute** tool enables you to replace values for observations that have missing values. You can replace missing values for interval variables with the mean, median, midrange, mid-minimum spacing, or distribution-based replacement, or you can use a replacement M-estimator such as Tukey's biweight, Huber's, or Andrew's Wave. You can also estimate the replacement values for each interval input by using a tree-based imputation method. Missing values for class variables can be replaced with the most frequently occurring value, distribution-based replacement, tree-based imputation, or a constant.

The **Interactive Binning** tool is an interactive grouping tool that you use to model nonlinear functions of multiple modes of continuous distributions. The Interactive tool computes initial bins by quantiles. Then you can interactively split and combine the initial bins. You use the Interactive Binning node to create bins or buckets or classes of all input variables, which include both class and interval input variables. You can create bins in order to reduce the number of unique levels as well as attempt to improve the predictive power of each input.

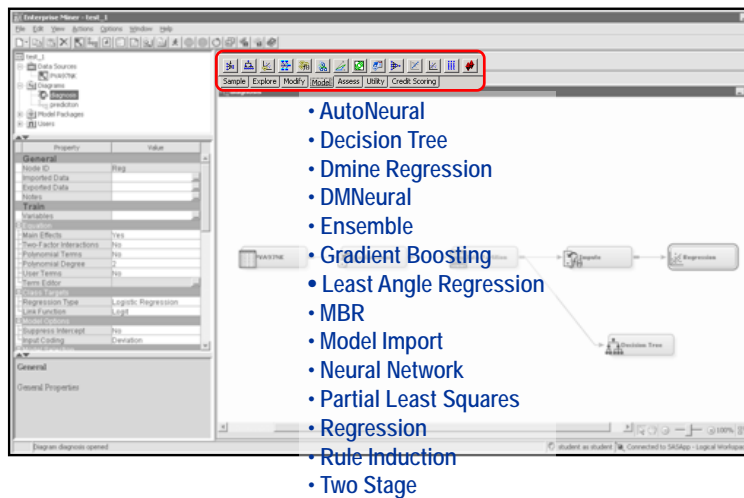
The **Principal Components** tool calculates eigenvalues and eigenvectors from the uncorrected covariance matrix, corrected covariance matrix, or the correlation matrix of input variables. Principal components are calculated from the eigenvectors and are usually treated as the new set of input variables for successor modeling tools. A principal components analysis is useful for data interpretation and data dimension reduction.

The **Replacement** tool enables you to reassign and consolidate levels of categorical inputs. This can improve the performance of predictive models.

The **Rules Builder** tool opens the Rules Builder window so that you can create ad hoc sets of rules with user-definable outcomes. You can interactively define the values of the outcome variable and the paths to the outcome. This is useful in ad hoc rule creation such as applying logic for posterior probabilities and scorecard values.

The **Transform Variables** tool enables you to create new variables that are transformations of existing variables in your data. Transformations are useful when you want to improve the fit of a model to the data. For example, transformations can be used to stabilize variances, remove nonlinearity, improve additivity, and correct nonnormality in variables. The Transform Variables tool supports various transformation methods. The available methods depend on the type and the role of a variable.

SEMMA – Model Tab



14

The **AutoNeural** tool can be used to automatically configure a neural network. It conducts limited searches for a better network configuration.

The **Decision Tree** tool enables you to perform multiway splitting of your database based on nominal, ordinal, and continuous variables. The tool supports both automatic and interactive training. When you run the Decision Tree tool in automatic mode, it automatically ranks the input variables based on the strength of their contributions to the tree. This ranking can be used to select variables for use in subsequent modeling. In addition, dummy variables can be generated for use in subsequent modeling. You can override any automatic step with the option to define a splitting rule and prune explicit tools or subtrees. Interactive training enables you to explore and evaluate a large set of trees as you develop them.

The **Dmine Regression** tool performs a regression analysis on data sets that have a binary or interval level target variable. The Dmine Regression tool computes a forward stepwise least squares regression. In each step, an independent variable is selected that contributes maximally to the model R-square value. The tool can compute all two-way interactions of classification variables, and it can also use AOV16 variables to identify nonlinear relationships between interval variables and the target variable. In addition, the tool can use group variables to reduce the number of levels of classification variables.



If you want to create a regression model on data that contains a nominal or ordinal target, then you would use the Regression tool.

The **DMNeural** tool is another modeling tool that you can use to fit a nonlinear model. The nonlinear model uses transformed principal components as inputs to predict a binary or an interval target variable.

The **Ensemble** tool creates new models by combining the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple predecessor models. The new model is then used to score new data. One common ensemble approach is to use multiple modeling methods, such as a neural network and a decision tree, to obtain separate models from the same training data set. The component models from the two complementary modeling methods are integrated by the Ensemble tool to form the final model solution. It is important to note that the ensemble model can only be more accurate than the individual models if the individual models disagree with one another. You should always compare the model performance of the ensemble model with the individual models. You can compare models using the Model Comparison tool.

The **Gradient Boosting** tool uses a partitioning algorithm described in "A Gradient Boosting Machine," and "Stochastic Gradient Boosting" by Jerome Friedman. A *partitioning algorithm* searches for an optimal partition of the data defined in terms of the values of a single variable. The optimality criterion depends on how another variable, the target, is distributed into the partition segments. When the target values are more similar within the segments, the worth of the partition is greater. Most partitioning algorithms further partition each segment in a process called *recursive partitioning*. The partitions are then combined to create a predictive model. The model is evaluated by goodness-of-fit statistics defined in terms of the target variable. These statistics are different than the measure of worth of an individual partition. A good model might result from many mediocre partitions.

The **Least Angle Regressions (LARS)** tool can be used for both input variable selection and model fitting. When used for variable selection, the LAR algorithm chooses input variables in a continuous fashion that is similar to Forward selection. The basis of variable selection is the magnitude of the candidate inputs' estimated coefficients as they grow from zero to the least squares' estimate. Either a LARS or LASSO algorithm can be used for model fitting. See Efron et al (2004) and Hastie, Tibshirani, and Friedman (2001) for further details.

The **Memory-Based Reasoning (MBR)** tool is a modeling tool that uses a k -nearest neighbor algorithm to categorize or predict observations. The k -nearest neighbor algorithm takes a data set and a probe, where each observation in the data set consists of a set of variables and the probe has one value for each variable. The distance between an observation and the probe is calculated. The k observations that have the smallest distances to the probe are the k -nearest neighbors to that probe. In SAS Enterprise Miner, the k -nearest neighbors are determined by the Euclidean distance between an observation and the probe. Based on the target values of the k -nearest neighbors, each of the k -nearest neighbors votes on the target value for a probe. The votes are the posterior probabilities for the class target variable.

The **Model Import** tool imports and assesses a model that was not created by one of the SAS Enterprise Miner modeling nodes. You can then use the Assessment node to compare the user-defined model(s) with a model(s) that you developed with a SAS Enterprise Miner modeling node. This process is called *integrated assessment*.

The **Neural Network** tool enables you to construct, train, and validate multilayer feed-forward neural networks. In general, each input is fully connected to the first hidden layer, each hidden layer is fully connected to the next hidden layer, and the last hidden layer is fully connected to the output. The Neural Network tool supports many variations of this general form.

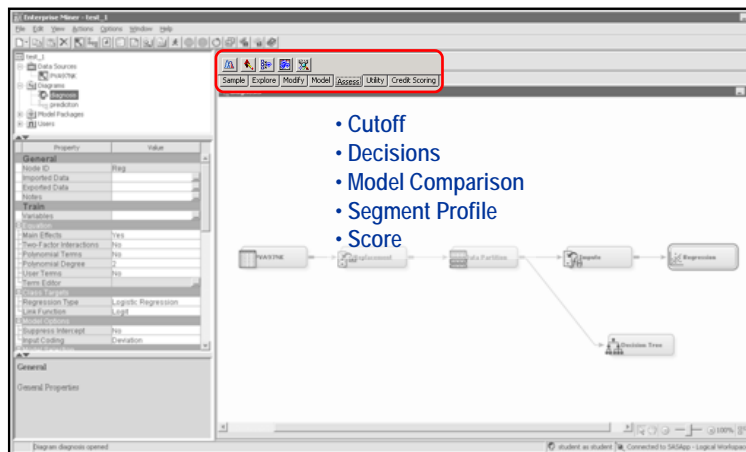
The **Partial Least Squares** tool models continuous and binary targets based on the SAS/STAT PLS procedure. The Partial Least Squares node produces DATA step score code and standard predictive model assessment results.

The **Regression** tool enables you to fit both linear and logistic regression models to your data. You can use continuous, ordinal, and binary target variables. You can use both continuous and discrete variables as inputs. The tool supports the stepwise, forward, and backward selection methods. The interface enables you to create higher-order modeling terms such as polynomial terms and interactions.

The **Rule Induction** tool enables you to improve the classification of rare events in your modeling data. The Rule Induction tool creates a Rule Induction model that uses split techniques to remove the largest pure split tool from the data. Rule induction also creates binary models for each level of a target variable and ranks the levels from the rarest event to the most common.

The **TwoStage** tool enables you to model a class target and an interval target. The interval target variable is usually the value that is associated with a level of the class target. For example, the binary variable **PURCHASE** is a class target that has two levels, Yes and No, and the interval variable **AMOUNT** can be the value target that represents the amount of money that a customer spends on the purchase.

SEMMA – Assess Tab



15

The **Cutoff** tool provides tabular and graphical information to assist users in determining appropriate probability cutoff point(s) for decision making with binary target models. The establishment of a cutoff decision point entails the risk of generating false positives and false negatives, but an appropriate use of the Cutoff node can help minimize those risks.

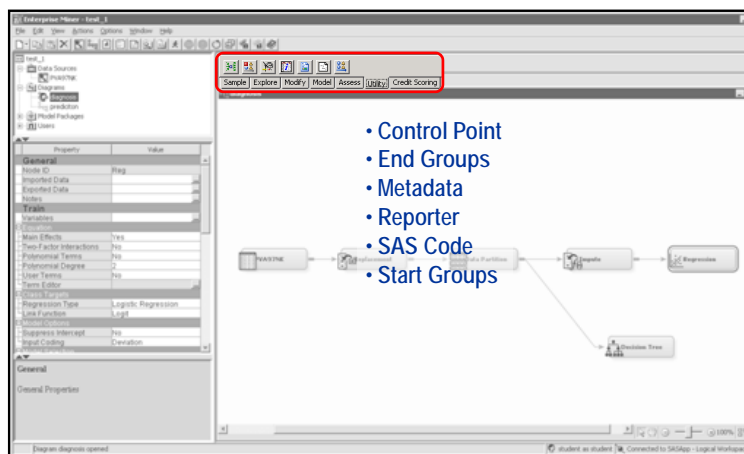
The **Decisions** tool enables you to define target profiles for a target that produces optimal decisions. The decisions are made using a user-specified decision matrix and output from a subsequent modeling procedure.

The **Model Comparison** tool provides a common framework for comparing models and predictions from any of the modeling tools. The comparison is based on the expected and actual profits or losses that would result from implementing the model. The tool produces several charts that help to describe the usefulness of the model, such as lift charts and profit/loss charts.

The **Segment Profile** tool enables you to examine segmented or clustered data and identify factors that differentiate data segments from the population. The tool generates various reports that aid in exploring and comparing the distribution of these factors within the segments and population.

The **Score** tool enables you to manage, edit, export, and execute scoring code that is generated from a trained model. Scoring is the generation of predicted values for a data set that might not contain a target variable. The Score tool generates and manages scoring formulas in the form of a single SAS DATA step, which can be used in most SAS environments even without the presence of SAS Enterprise Miner. The Score tool can also generate C score code and Java score code.

Beyond SEMMA – Utility Tab



16

The **Control Point** tool enables you to establish a control point to reduce the number of connections that are made in process flow diagrams. For example, suppose that three Input Data Source tools are to be connected to three modeling tools. If no Control Point tool is used, then nine connections are required to connect all of the Input Data Source tools to all of the modeling tools. However, if a Control Point tool is used, only six connections are required.

The **End Groups** tool is used only in conjunction with the Start Groups tool. The End Groups node acts as a boundary marker that defines the end-of-group processing operations in a process flow diagram. Group processing operations are performed on the portion of the process flow diagram that exists between the Start Groups node and the End Groups node.

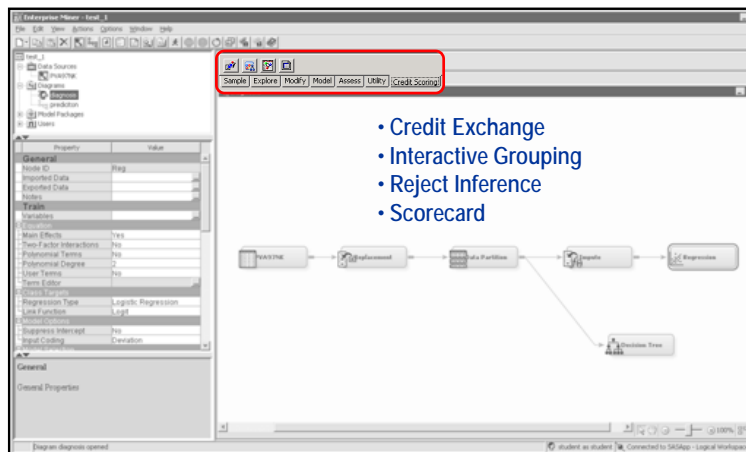
The **Metadata** tool enables you to modify the columns metadata information at some point in your process flow diagram. You can modify attributes such as roles, measurement levels, and order.

The **Reporter** tool uses SAS Output Delivery System (ODS) capability to create a single PDF or RTF file that contains information about the open process flow diagram. The PDF or RTF documents can be viewed and saved directly and are included in SAS Enterprise Miner report package files.

The **SAS Code** tool enables you to incorporate new or existing SAS code into process flow diagrams. The ability to write SAS code enables you to include additional SAS procedures into your data mining analysis. You can also use a SAS DATA step to create customized scoring code, to conditionally process data, and to concatenate or merge existing data sets. The tool provides a macro facility to dynamically reference data sets used for training, validation, testing, or scoring variables, such as input, target, and predict variables. After you run the SAS Code tool, the results and the data sets can then be exported for use by subsequent tools in the diagram.

The **Start Groups** tool is useful when your data can be segmented or grouped, and you want to process the grouped data in different ways. The Start Groups node uses BY-group processing as a method to process observations from one or more data sources that are grouped or ordered by values of one or more common variables. BY variables identify the variable or variables by which the data source is indexed, and BY statements process data and order output according to the BY group values.

Credit Scoring Tab (Optional)



17

The optional Credit Scoring tab provides functionality related to credit scoring.



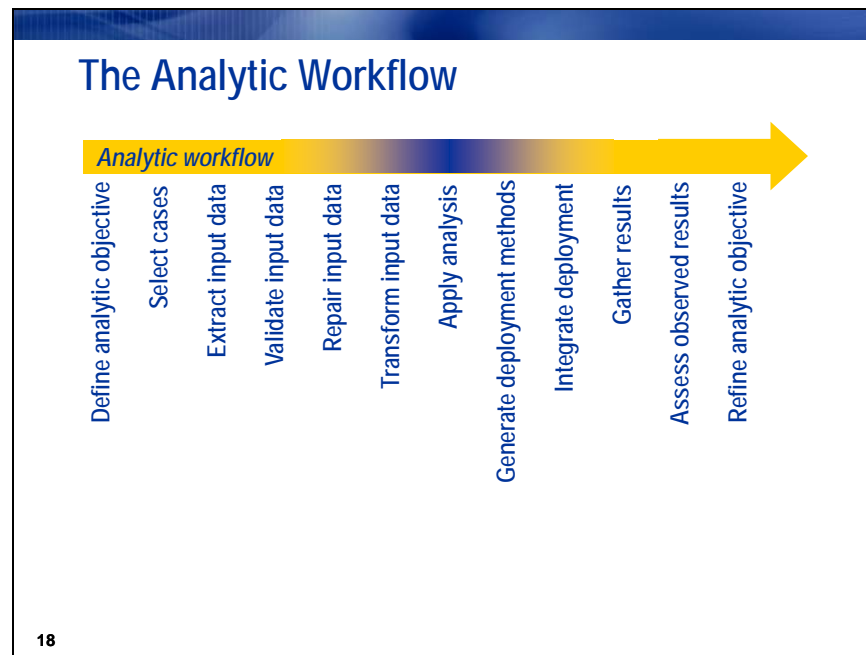
The Credit Scoring for the SAS Enterprise Miner solution is not included with the base version of SAS Enterprise Miner. If your site did not license Credit Scoring for SAS Enterprise Miner, the Credit Scoring tab and its associated tools do not appear in your SAS Enterprise Miner software.

The **Credit Exchange** tool enables you to exchange the data that is created in SAS Enterprise Miner with the SAS Credit Risk Management solution.

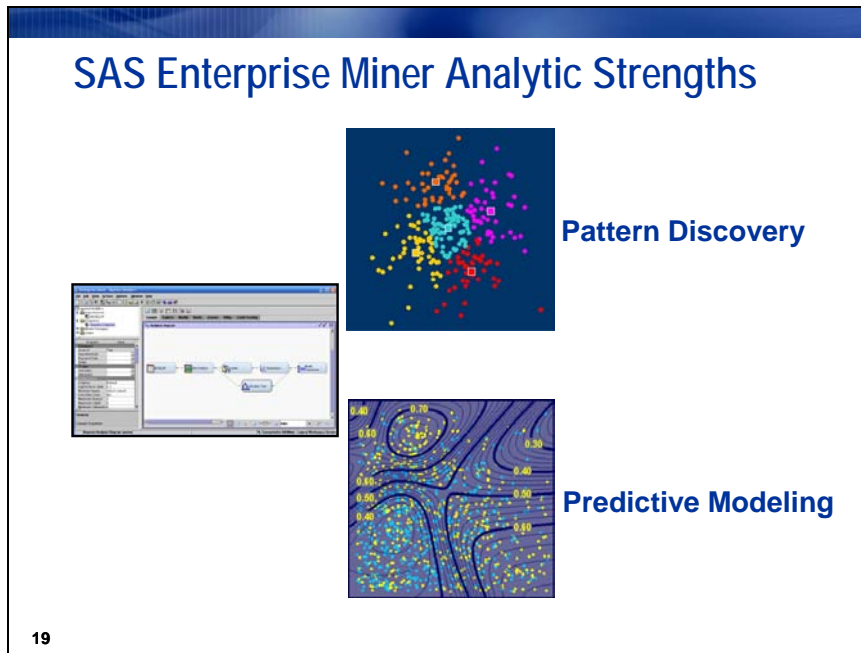
The **Interactive Grouping** tool creates groupings, or classes, of all input variables. (This includes both class and interval input variables.) You can create groupings in order to reduce the number of unique levels as well as attempt to improve the predictive power of each input. Along with creating group levels for each input, the Interactive Grouping tool creates Weight of Evidence (WOE) values.

The **Reject Inference** tool uses the model that was built using the accepted applications to score the rejected applications in the retained data. The observations in the rejected data set are classified as inferred “goods” and inferred “bads.” The inferred observations are added to the **Accepts** data set that contains the actual “good” and “bad” records, forming an augmented data set. This augmented data set then serves as the input data set of a second credit-scoring modeling run. During the second modeling run, attribute classification is readjusted and the regression coefficients are recalculated to compensate for the data set augmentation.

The **Scorecard** tool enables you to rescale the logit scores of binary prediction models to fall within a specified range.



The *analytic workflow* is the sequence of steps required to fulfill an applied analytic objective. The tools and capabilities of SAS Enterprise Miner occupy the central steps of this workflow. Before using SAS Enterprise Miner, you must carefully define your analytic objective, select analysis cases, and extract, validate, and possibly repair analysis data. SAS Enterprise Miner then enables you to further transform your data, apply the analysis of interest, and generate deployment methods. The analytic workflow then continues outside the competencies of SAS Enterprise Miner. Deployment methods must be integrated into production systems, and results from this integration must be captured, assessed, and used to refine the next iteration of analysis.

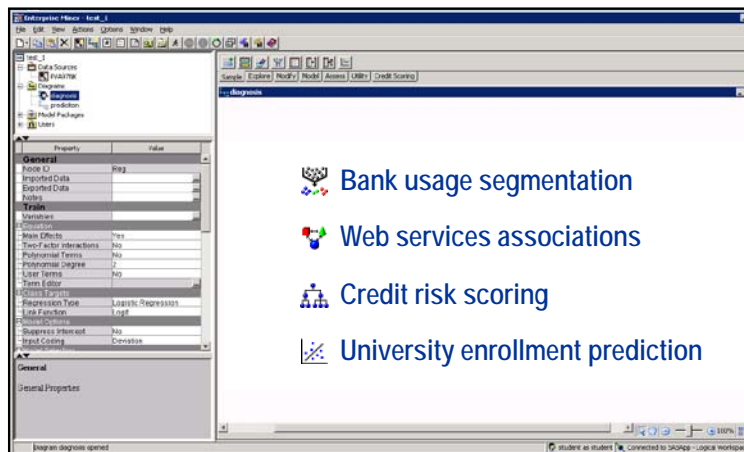


The analytic strengths of SAS Enterprise Miner lie in the realm traditionally known as *data mining*. There are a great number of analytic methods identified with data mining, but they usually fall into two broad categories: *pattern discovery* and *predictive modeling* (Hand 2005). SAS Enterprise Miner provides a variety of pattern discovery and predictive modeling tools.

Chapters 2 through 7 demonstrate SAS Enterprise Miner's predictive modeling capabilities. In those chapters, you use a rich collection of modeling tools to create, evaluate, and improve predictions from prepared data. Special topics in predictive modeling are featured in Chapter 9.

Chapter 8 introduces some of SAS Enterprise Miner's pattern discovery tools. In this chapter, you see how to use SAS Enterprise Miner to graphically evaluate and transform prepared data. You use SAS Enterprise Miner's tools to cluster and segment an analysis population. You can also choose to try SAS Enterprise Miner's market basket analysis and sequence analysis capabilities.

Applied Analytics Case Studies



20

Appendix A further illustrates SAS Enterprise Miner's analytic capabilities with case studies drawn from real-world business applications.

- A consumer bank sought to segment its customers based on historic usage patterns. Segmentation was to be used for improving contact strategies in the Marketing Department.
- A radio station developed a Web site to provide such services to its audience as podcasts, news streams, music streams, archives, and live Web music performances. The station tracked usage of these services by URL. Analysts at the station wanted to see whether any unusual patterns existed in the combinations of services selected.
- A bank sought to use performance on an in-house subprime credit product to create an updated risk model. The risk model was to be combined with other factors to make future credit decisions.
- The administration of a large private university requested that the Office of Enrollment Management and the Office of Institutional Research work together to identify prospective students who would most likely enroll as freshmen.