

## Introduction

*Data Mining Using SAS Enterprise Miner* introduces the readers to data mining using SAS Enterprise Miner v4. This book will reveal the power and ease of use of the powerful new module in SAS that will introduce the readers to the various configuration settings and subsequent results that are generated from the various nodes in Enterprise Miner that are designed to perform data mining analysis. This book consists of step-by-step instructions along with an assortment of illustrations for the reader to get acquainted with the various nodes and the corresponding working environment in SAS Enterprise Miner. The book provides an in-depth guide in the field of data mining that is clear enough for novice statisticians or the die-hard expert.

The process of extracting information from large data sets is known as data mining. The objective in data mining is making discoveries from the data. That is, discovering unknown patterns and relationships by summarizing or compressing the data in a concise and efficient way that is both understandable and useful to the subsequent analysis. Extracting information from the original data that will result in an accurate representation of the population of interest, summarizing the data in order to make statistical inferences or statements about the population from which the data was drawn and observing patterns that seem most interesting, which might lead you to discover abnormal departures from the general distribution or trend in the data; for example, discovering patterns between two separate variables with an usually strong linear relationship, a combination of variables that have an extremely high correlation on a certain variable, or grouping the data to identify certain characteristics in the variables between each group, and so on. In predictive modeling, it is important to identify variables to determine certain distributional relationships in the data set in order to generate future observations or even discover unusual patterns and identifying unusual observations in the data that are well beyond the general trend of the rest of the other data points.

The basic difference between data mining and the more traditional statistical applications is the difference in size of the data set. In traditional statistical designs, a hundred observations might constitute an extremely large data set. Conversely, the size of the data mining data set in the analysis might consist of several million or even billions of records. The basic strategy that is usually applied in reducing the size of the file in data mining is sampling the data set into a smaller, more manageable subset that is an accurate representation of the population of interest. The other strategy is summarizing the variables in the data by their corresponding mean, median or sum-of-squares. Also, reducing the number of variables in the data set is extremely important in the various modeling designs. This is especially true in nonlinear modeling where an iterative grid search procedure must be performed in finding the smallest error from the multidimensional error surface.

The potential of data mining used in statistical analysis is unlimited. However, it would be a mistake to depend on data mining in providing you with the final solution to all the questions that need to be answered. In addition, the accuracy of the final results in data mining analysis depends on the accuracy of the data. The most common term that is often used is called “garbage in – garbage out”. This is particularly true in data mining analysis where the chance of this phenomenon of happening will occur more often than not due to the enormous amount of data at your disposal. In some instances, discovering patterns or relationships in the data might be due to the measurement inaccuracies, distorted samples, or some unsuspected difference between the erroneous data set and the actual representation of the population of interest.

Often, the choice of the statistical method to apply depends on the main objective of the analysis. For example, in marketing research with the disposal of customer and transactions data, it is usually important to interpret the buying behaviors of the various customers in order to focus our attention on the more effective promotions that will result in an increase of sales that can be accomplished by performing association analysis. It might be important in identifying and profiling the various buying habits of these same customers who might be placed into separate homogenous groups based on the total sales of the various items purchased, which can be performed by applying cluster analysis. Maybe the goal to the analysis might be predicting or forecasting future sales, which can be performed by applying regression modeling. Nonlinear modeling such as neural network modeling might be considered, which does not require a functional form between the predictor or input variables and the response, outcome or target variable to the model, with the added flexibility of handling an enormous number of input variables in the model. Two-stage modeling might be performed by classifying customers in purchasing certain items, then predicting these same customers based on their corresponding estimated probabilities, which are then applied in the subsequent modeling fit to predict the amount of total sales of these same items. Maybe it might be important to the business in developing a credit scoring system in

order to establish whether or not to extend credit to customers based on certain numerical scores that are above or below some threshold value, which can be performed by applying interactive grouping analysis.

SAS summarizes data mining with the acronym SEMMA, which stands for sampling, exploring, modifying, modeling, and assessing data as follows:

**Sample** the data from the input data set that is so large that a small proportion of the data can be analyzed at any one time, yet large enough that it contains a significant amount of information to perform the analysis.

**Explore** the data to statistically, and visually discover expected relationships and unexpected trends while at the same time discovering abnormalities in the data to gain a better understanding of the data.

**Modify** the data in creating, selecting, and transforming the variables or even the entire incoming data set to perform various modeling selections or certain statistical modeling techniques in preparation for the subsequent data mining analysis.

**Model** the data by applying various modeling techniques in seeking a certain combination of variables that reliability predicts the outcome response.

**Assess** the data by evaluating the usefulness and reliability of the results from the data mining process.

Data mining is an iterative process. That is, there is a sequential order to the listed categories in the SEMMA data mining analysis. Typically, you would first **sample** the data to target our population or determine the population of interest to the analysis. Second, you might **explore** the data to visualize the distribution of each variable in the analysis to validate the statistical assumptions such as normality in the distribution of the selected variable or determine patterns and relationships between the variables. Third, you might **modify** the variables to prepare the data for analysis by transforming the variables to satisfy the various assumptions that are critical to many of the statistical methods so that these same variables may be used in the analysis. The next step is that you might **model** the data by fitting a statistical model to generate predictions and forecasts. And, finally, you might **assess** the accuracy, interpreting the results and comparing the predictability of the various modeling designs in selecting the best predictive or classification model. Once the best model is selected, then you might want to generate prediction or forecasting estimates through the use of a scoring function that can be applied to a new set of values that might not necessarily consist of the target variable in the data. It is important to realize that many of the previously mentioned data mining steps might not be applied at all or some of these steps might be applied any number of times before the goal of the data mining analysis is finally achieved.

The SEMMA design and data mining analysis is constructed within the process flow diagram. Enterprise Miner is designed so that very little SAS programming experience is needed in constructing a well-built SAS reporting system. The reason is that the process of constructing the process flow diagram within Enterprise Miner is performed by simply dragging icons on to a GUI interface desktop window, then connecting these same icons to one another, all within the same graphical diagram workspace. And yet a data mining expert can specify various option settings in the design in order to fine-tune the configuration settings and the corresponding listings and results. SAS Enterprise Miner is a very easy to learn and very easy to use. You do not even need to know SAS programming and can have very little statistical expertise in designing an Enterprise Miner project in order to develop a completely comprehensive statistical analysis reporting system, whereas an expert statistician can make adjustments to the default settings and run the Enterprise Miner process flow diagram to their own personal specifications. Enterprise Miner takes advantage of the intuitive point-and-click programming within a convenient graphic user interface. The diagram workspace or the process flow diagram has the look and appearance much like the desktop environment in Microsoft Windows. Enterprise Miner is built around various icons or nodes at your disposal that will perform a wide variety of statistical analysis.

Each chapter of the book is organized by the SEMMA acronym based on the various nodes that are a part of the data mining acronym. Each section to the book is arranged by the way in which the node appears within the hierarchical listing that is displayed in the **Project Navigator** within the Enterprise Miner window. Each section of the book will begin by providing the reader with an introduction to the statistics and some of the basic concepts in data mining analysis with regard to the corresponding node. The book will then explain the various tabs along with the associated option settings that are available within each tab of the corresponding node, followed by an explanation of the results that are generated from each one of the nodes. In the first chapter, the book will begin by explaining the purpose of the various sampling nodes that are a part of the Sample section in the SEMMA design. The first section will introduce the readers to the way in which to

read the data in order to create the analysis data set to be passed on to the subsequent nodes in the process flow diagram. The following section will allow the readers to learn how to both randomly sample and partition the analysis data set within the process flow. This is, sampling the analysis data set into a smaller, more manageable sample size that is an accurate representation of the data that was randomly selected, or even split the analysis data set into separate files that can be used in reducing the bias in the estimates and making an honest assessment in the accuracy of the subsequent predictive model.

The second chapter will focus on the various nodes that are designed to discover various relationships or patterns in the data that are a part of the Explore section in the SEMMA design. The first couple nodes that are presented are designed to generate various frequency bar charts or line graphs in order to visually observe the univariate or multivariate distribution of the variables in the data. In addition, the readers will be introduced to the **Insight** node, which can perform a wide range of statistical analysis through the use of the synchronized windows. The following section will explain the purpose of association analysis that is widely used in market basket analysis in discovering relationships between the different combinations of items that are purchased. The next section will introduce the readers to the variable selection procedure that can be applied within the process flow, which is extremely critical in both predictive or classification modeling designs that are designed to select the best set of input variables among a pool of all possible input variables. The chapter will conclude with the readers getting familiar with link analysis which automatically generates various link graphs in order to view various links or associations between the various class levels that are created from the analysis data set.

The third chapter will introduce readers to the various nodes that are designed to modify the analysis data sets that are a part of the Modify section of the SEMMA design. The chapter will allow the readers to realize how easy it is to modify the various variable roles or level of measurements that are automatically assigned to the analysis data set once the data set is read into the process flow. In addition, the subsequent section will allow readers to transform the variables in the data set in order to meet the various statistical assumptions that must be satisfied in data mining analysis. The following section will allow readers to understand both the importance and the process of filtering, excluding, and removing certain problematic observations from the data set that might otherwise influence the final statistical results. The next section will explain the procedure of imputing missing values and replacing undesirable values in the data set and the subsequent analysis. The following two sections will explain both clustering and SOM/Kohonen analysis, which are designed to group the data into homogenous groups in order to profile or characterize the various groups that are created. The next section will explain the way in which to transform the data set in preparation to repeated measures or time series analysis. And, finally, the chapter will conclude by introducing readers to interactive grouping analysis that is designed to automatically create separate groups from the input variables in the data set based on the class levels of the binary-valued target variable. These input variables may then be used as input variables in subsequent classification modeling designs such as fitting scorecard models.

The fourth chapter will present the readers to the various modeling nodes in Enterprise Miner that are a part of Model section in the SEMMA design. The chapter will begin with traditional least-squares modeling and logistic regression modeling designs that are based on either predicting an interval-valued or a categorically-valued target variable of the model. The following section will introduce the readers to the **Model Manager** which is available in any one of the modeling nodes. The purpose of the **Model Manager** is to store and list the various models that are created within each modeling node. Each model that is displayed is based on the different settings that were previously specified each time you saved the corresponding changes. The **Model Manager** will allow you to select the number of observations from each partitioned data set to be passed along for interactive assessment in evaluating the accuracy of the modeling fit from the **Assessment** node. In addition, the **Model Manager** will allow you to specify if the various diagnostic plots or performance charts will be available for viewing within the **Assessment** node for each partitioned data set. The next section will introduce the readers to decision tree modeling, which is based on a recursive splitting process in which binary splits are automatically performed, and where the average value of the interval-valued target variable is commonly used as the standard cutoff point or the separate class levels of the categorically-valued target variable that are repeatedly divided through the decision tree based on the corresponding range of values or the separate class levels of the input variables in the model. Neural network modeling will then be introduced. This is essentially nonlinear modeling of the process flow that has the flexibility of interpolating many different functional forms with extreme accuracy or approximating many different classification boundaries with great precision. Neural network modeling is built around a multilayered design in which the linear combination of input variables and weight estimates are transformed through the layers where the weight

estimates must be solved by some type of iterative grid search or line search routine. The next section will explain principal components analysis to readers; this is designed to reduce the number of variables in the data based on the linear combination of input variables and components in the model, where the components are selected to explain the largest proportion of the variability in the data. User-defined modeling will be introduced that will allow you to incorporate a wide variety of modeling techniques within the process flow that are unavailable in Enterprise Miner and the various modeling nodes. Furthermore, the node will allow you to create your own scoring code that will generate the standard modeling assessment statistics that can be passed along the process flow in order to compare the accuracy between the other modeling nodes. The book will provide readers with a powerful modeling technique called ensemble modeling that either averages the prediction estimates from various models or averages the prediction estimates based on successive fits from the same predictive model, where the analysis data set is randomly sampled any number of times. The following section will introduce the readers to memory-based reasoning or nearest neighbors modeling that is essentially nonparametric modeling in which there are no distributional assumptions that are assumed in the model in which the fitted values are calculated by averaging the target values or determined by the largest estimated probability based on the most often occurring target category within a predetermined region. The chapter will conclude with two-stage modeling that fits a categorically-valued target variable and an interval-valued target variable in succession, where the categorically-valued target variable that is predicted in the first-stage model will hopefully explain a large majority of the variability in the interval-valued target variable to the second-stage model.

The fifth chapter will explain the various nodes in Enterprise Miner that are a part of Assess section to the SEMMA design. These nodes will allow you to evaluate and assess the results that are generated from the various nodes in the process flow. The chapter will begin by introducing the readers to the **Assessment** node that will allow you to evaluate the accuracy of the prediction estimates from the various modeling nodes based on the listed assessment statistics that are automatically generated and the numerous diagnostic charts and performance plots that are created for each model. The node will allow you to evaluate the accuracy of the modeling estimates by selecting each model separately or any number of models simultaneously. The **Reporter** node will be introduced, which is designed to efficiently organize the various option settings and corresponding results that are generated from the various nodes within the process flow into a HTML file that can be viewed by your favorite Web browser.

The sixth chapter will introduce the readers to the **Score** node that manages, exports, and executes the SAS scoring code in order to generate prediction estimates from previously trained models. In addition, the node will allow you to write your own custom-designed scoring code or scoring formulas that can be applied to an entirely different sample drawn in order to generate your own prediction or classification estimates.

The final chapter of the book will conclude with the remaining nodes that are available in Enterprise Miner. These nodes are listed in the Utility section within the **Project Navigator**. The chapter will begin by explaining the purpose of the **Group Processing** node that will allow you to transform variables by splitting these same variables into separate groups. In addition, the node is used with the **Ensemble** node that determines the way in which the various prediction estimates are formed or combined within the process flow. The subsequent section will explain the purpose of the data mining data set that is designed to accelerate processing time in many of the nodes within the process flow. This is because the data mining data set contains important metadata information to the variables in the analysis, such as the variable roles, level of measurement, formats, labels, range of values, and the target profile information to name a few. The next section will explain the importance of the **SAS Code** node. The node is one of the most powerful nodes in Enterprise Miner. This is because the node will allow you to incorporate SAS programming within the process flow, thereby enabling you to write data step programming within the process flow in order to manipulate the various data sets, write your own scoring code, or access the wide variety of procedures that are available in SAS. The final two sections will explain the purposes of the **Control point** node and the **Subdiagram** node that are designed to manage and maintain the process flow more efficiently.

For anyone looking for a new edge in the field of statistics, *Data Mining Using SAS Enterprise Miner*, offers the inside track to new knowledge in the growing world of technology. I hope that reading the book will help statisticians, researchers, and analysts learn about the new tools that are available in their arsenal making statisticians and programmers aware of this awesome new product that reveals the strength and easy use of Enterprise Miner for creating a process flow diagram in preparation for data mining analysis.