Predictive modeling with sas enterprise miner

More. Chapter 1: Research Strategy 1.1 Introduction 1.2 Entry Types 1.2.1 Variable Measurement Scales 1.2.2 Predictive Model With Text Data 1.3 Definition of Target 1.3.1 Forecasting Reaction to Direct Mail 1.3.2 Risk Forecasting in the Auto Insurance Industry 1.3.3.3 Forecasting The Sensitivity of Bank Deposit Products Rate 1.3.4 Customer Loss Forecasting 1.3.5 Forecasting Rated Categorical (Disordered Polyhotomy) Target 1.4 Data Modeling Sources 1.4.1 Comparison Between Sample and Target Universe 1.4.2 Surveillance Weight 1.5 Pre-Processing Data 1.5.1 Clearing Data Before Launch SAS Enterprise Miner 1.5.2 Clearing Data 1.5.2 Data After the launch of SAS Enterprise Miner 1.6 Alternative Simulation Strategies 1.6.1 Regression with a moderate number of input variables 1.6.2 Regression with a large number of input variables 1.7 Notes 1.1 Introduction This chapter discusses the planning and organization of the project predictive modeling. Planning includes tasks such as: - identifying and measuring the target variable in accordance with the business issue - data collection - comparing the distribution of key variables between the modeling dataset and the target population to verify that the sample adequately represents the target group of the population - determining the sample weights if necessary - performing the tasks of cleaning the data that needs to be performed before the launch of the SAS® Enterprise Miner™ Alternative Strategies for the Development of Predictive Models with the use of the projected. 1.2 Types of inputs in the predictive type can be used by different types of data. The general types of data are numerical and character. Measurement scales called SAS levels® Enterprise Miner™, predictive modeling set variables are defined in section 1.2.1. The use of text inputs in predictive modeling is discussed in section 1.2.2. 1.2.1 The I variable measurement scale will first determine the variable measurement scales used in this book. In general, I tried to follow the definitions given by Alan Agresti: The categorical variable is one for which the measurement scale consists of a set of categories. Categorical variables for which levels (categories) do not have a natural order are called nominal. The categorical variables that have the natural order of their levels are called serial. Interval variable is a difference that has numerical distances between any two levels of the scale.1 - Binary variable occupies only two values, such as 1 and 0, M and F, etc. According to the above definitions, income and AGE variables in tables 1.1 to 1.5 and BAL_AFTER in table 1.3 are variables of interval scale. Because the RESP variable in Table 1.1 is categorical and has only two levels, it is called a binary variable. The LOSSFR variable in Table 1.2 is orderly. (In SAS Enterprise you can change its measurement scale to an interval, but I left it as a serial.) Teh Teh PRIORPR and NEXTPR in Table 1.5 are nominal. Variable interval scales are sometimes referred to as continuous. Continuous variables are treated as interval variables. So I use the terms interval scale and continuous interchangeable. I also use terms, orderly polychotomous variables and serial variables interchangeable. Similarly, I use the terms disordered polychotomous variables and nominal variables interchangeable. 1.2.2 To develop a predictive model with text text data data can be used to develop predictive. To develop predictive models from text data, you first need to convert text data into numerical form. Text data is first arranged in a tablicular form, where each row of the table contains one complete document. Some examples of text data and how to convert text data into numerical form are discussed in Chapter 9. 1.3 Determining the target first step in any data collection project is to identify and measure the target variable that will be predicted by the model that flows from the data analysis. This section provides examples of this step that apply to five different business issues. 1.3.1 Direct Mail Forecasting In this example, a hypothetical auto insurance company wants to purchase customers by direct mail. The company wants to minimize the cost of mailing, focusing only on the most responsive customers. So the company decides to use a response model. The target variable for this model is the RESP, and it's binary, taking a value of 1 for the answer and 0 for the answer. Table 1.1 shows a simplified version of the dataset used to model the Binary Target Response (RESP). Table 1.1 In Table 1.1, AGE, INCOME, STATUS, PC, and NC variables are input variables (or explanatory variables). AGE and INCOME are numerical, and while they can theoretically be considered continuous, it is simply more practical to treat them as variables at intervals. The status variable is categorical and nominally scaled. Category S, if the client is not married and never married, MC if married with children, MNC, if married without children, W if widowed, and D if divorced. The variable PC is numerical and binary. It indicates whether customers own a personal computer, taking a value of 1 if they do and 0 if not. The NC variable represents the number of credit cards that customers have. You can decide whether this variable is orderly or interval-based. The target variable is RESP and takes a value of 1 if the customer responded, for example, to the campaign newsletter, and 0 otherwise. The binary goal can be both numerical and symbolic; I could have written down an answer like Y instead of 1, and a no-answer like N instead of 0, with little or no impact on the shape of the final equation. Please note that There are some extreme values. For example, the age of one customer is registered as 6 years. This is obviously a record error, and the age should be corrected to show the actual value if Revenue has missing values that are displayed as dots, while the status nominal variable has missing values that are represented by spaces. The SAS Enterprise Miner can be used to imperid such missing values. Read more in Chapters 2, 6 and 7. 1.3.2 Risk Forecasting in the auto insurance industry Auto insurance company wants to examine its customers' data and classify its customers into different risk groups. The goal is to bring premiums in line with their customers' risk rates. If high-risk customers are charged a low premium, the loss rates will be too high and the company will be driven out of the business. If low-risk customers are charged disproportionately, the company will lose customers from competitors. By accurately assessing the risk profiles of its customers, the company hopes to set customer premiums at the optimal level consistent with risk. The risk model is necessary to assign a risk assessment to each existing client. In the risk model, the frequency of losses can be used as a target variable. The frequency of losses is calculated as the number of losses as a result of an accident per car per year, where the year of the car equals the time from the moment due to the auto insurance policy expressed in years multiplied by the number of cars covered by the policy. The frequency of losses can be considered as a continuous (interval) variable or discrete (orderly) variable that categorizes each customer's losses into a limited number of cells. (See Chapters 5 and 7 for detailed information on bunkers.) For illustration purposes, I model the loss frequency as a continuous variable in Chapter 4 and as a discrete order variable in chapters 5 and 7. The frequency of losses is considered here the losses associated with the accident in which the customer was to blame, so it can also be called due to the frequency of accidents. I use loss frequency, frequency of claims and frequency of accidents interchangeable. Table 1.2 shows what a simulation dataset might look like to develop a loss-frequency model as an orderly target. Table 1.2 Target variable is LOSSFR, which is a one-year car-year accident incurred by a customer over a period of time. This variable is discussed in more detail in subsequent chapters of this book. At the moment, it is enough to note that this is a serial variable that assumes values 0, 1, 2 and 3. Input variables: AGE, INCOME and NPRVIO. The NPRVIO variable represents the number of previous violations that the customer had prior to the purchase of the insurance policy. 1.3.3 Forecasting the sensitivity of bank deposit rates To assess the sensitivity of customers to increase the interest rate on a savings account, the bank can conduct price tests. Suppose one of these tests includes a proposal high bet for a certain period of time, called the promotion window. To assess the customer's sensitivity to speed, you can attribute three types of models to the data generated by the experiment: the experiment: Response model for predicting the probability of the answer - a short-term demand model for predicting the expected change in deposits during the promotion period - a model of long-term demand to predict an increase in deposit levels after the promotion period Target variable for the binary response model: the answer or no answer. The target variable for the short-term demand model is to increase savings deposits during the promotion of net2 of any collateral downturns on other accounts. The target variable for the long-term demand model is the amount of increase left in customers' bank accounts after the promotion period. In the case of this model, the promotion window for analysis should be clearly defined, and only client transactions that occurred before the promotion window should be included as input in the modeling sample. Table 1.3 shows what a data set looks like to model a continuous target. Table 1.3 The data set shown in Table 1.3 is an attempt by a hypothetical bank to encourage its customers to increase their savings deposits by increasing the interest paid to them in advance by a certain number of basis points. This increased interest rate was offered (suppose) in May 2006. Customer deposits were then registered at the end of May 2006 and stored in a dataset shown in Table 1.3 under the variable name BAL_AFTER. The bank would like to know what type of customer is likely to increase its savings balances the most in response to a future incentive by the same amount. The target variable for this is a dollar change of balances from point to share period to point after the stock period. The target variable is continuous. Input or explanatory variables: AGE, INCOME, B_JAN, B_FEB, B_MAR, and B_APR. Variable B_JAN, B_FEB, B_MAR and B_APR relate to customer balances in all their accounts at the end of January, February, March and April 2006, respectively. 1.3.4 Predicting customer loss in the banking sector may mean that a customer closes a savings account, checking account or investment account. In the model for predicting depletion, the target variable can be both binary and continuous. For example, if a bank wants to identify customers who may terminate their accounts at any time for a predetermined period of time in the future, you can model exhaustion as a binary goal. However, if the bank is interested in predicting the specific time attrit the customer is likely to attrit, it is best to model depletion as a continuous goal - time attrit. In this example, depletion is modeled as a binary target. When you're depleting a binary goal, you need to determine the performance window during which you observe the occurrence or non-appearance of the event. If a customer is up during a performance window, the record shows 1 for the event and 0 otherwise. Any transactions (deposits, (deposits, and the transfer of funds) that are used as inputs to develop the model should occur in the period leading up to the performance window. The input window during which transactions are observed, the performance window during which the event occurs, and the operational lag, which is a delay in receiving input, are discussed in detail in Chapter 7, where the depletion model is being developed. Table 1.4 shows what a data set looks like to simulate customer depletion. Table 1.4 In the dataset shown in Table 1.4, the ATTR variable represents customer depletion observed in the performance window, consisting of the June, July, and August 2006 periods. The target variable takes the value of 1 if the client attrits during the performance window and 0 otherwise. Table 1.4 shows input variables for the model. This is AGE, INCOME, B_JAN, B_FEB, B_MAR and B_APR. Variable B_JAN, B_FEB, B_MAR and B_APR relate to customer balances in all their accounts at the end of January, February, March and April 2006, respectively. 1.3.5 Forecasting the nominal categorical (disordered polychotomy) target suggests that a hypothetical bank wants to predict based on the products the customer currently owns and other characteristics of which product the customer is most likely to buy next. For example, a client may currently have a savings account and checking account, and the bank would like to know if the client is more likely to open an investment account or IRA, or take out a mortgage. The target variable for this situation is nominal. Models with nominal goals are also used by market researchers who need to understand consumer preferences for different products or brands. Chapter 6 provides some examples of models with nominal goals. Table 1.5 shows what a data set might look like to model a nominal categorical goal. Table 1.5 Table 1.5 inputs include the PRIORPR variable, which indicates a product or product belonging to a hypothetical bank customer at the beginning of the performance window. The performance window, defined in the same way as in section 1.3.4, is the period of time a customer purchases are observed. Given that the customer owned certain products at the beginning of the performance window, we see the next product that the customer purchased during the performance window and specify it with the NEXTPR variable. For each customer, the PRIORPR variable indicates a product that belonged to the customer at the beginning of the performance window. Letter A can mean a savings account, B can stand for a deposit certificate, etc. Similarly, the value of the NEXTPR variable indicates the first product purchased by the customer during the performance window. For example, if a customer owned a B product at the beginning of the performance window and purchased X and q products in this order, Performance window time, then variable variable If a customer has purchased q and X, in this order, the NEXTPR variable accepts the value of q, and the PRIORPR variable accepts the B value in the customer's record. 1.4 There are two different scenarios for which data becomes available for modeling. For example, consider a marketing campaign. In the first scenario, the data is based on an experiment conducted through a marketing campaign for a well-designed sample of customers taken from the target population. In the second scenario, the data is a sample taken from the results of the previous marketing campaign, not from the target population. While the latter scenario is clearly less desirable, it is often necessary to deal with any available data. In such cases, some adjustments can be made using observational scales to compensate for the lack of ideal compatibility between model sampling and target population. In any case, for modeling purposes, we accustom the file with the results of the marketing campaign to the data on customer characteristics and customer operations. Although transaction data is not always available, it is generally key to predicting a depletion event. 1.4.1 Comparison between sample and Target Universe Before launching a simulation project, you need to make sure that the sample is a good representation of the target universe. This can be done by comparing the distribution of some key variables in the sample and the target universe. For example, if age and income are key characteristics, the age and income distribution between the sample and the target universe should be compared. 1.4.2 The weight of observations If the distribution of key characteristics in the sample and target population is different, sometimes the weight of observations is used to correct any biases. In order to detect the difference between the target population and the sample, you need to have some prior knowledge about the target population. Assuming that age and income are key characteristics, you can gain weights as follows: Divide income into, say, four groups and age into, say, three groups. Suppose the target universe has nij people in the ith age group and jth income group, and suppose that a sample of nij people in the same age group of income. In addition, suppose that the total number of people in the target population N, and the total number of people in the sample n. In this case, the corresponding weight of observation (Nij/N)/ (nij/n) for the person in the ith age group and jth income group in the sample. You have to build these observation scales and incorporate them for each entry into the simulation sample before launching THE SAS Enterprise Miner, effectively creating an additional variable in the dataset. In SAS Enterprise Miner, the frequency role is assigned to this variable so that modeling tools take these weights into account when evaluating models. This situation inevitably arises you don't have a scientific sample taken from the target population, which is very common. However, another source of bias is often deliberately introduced. This bias is due to an over-sampling of rare events. For example, when modeling responses, if the response level is very low, you need to include all available responders and only a random share of non-responders. Prejudice by making such over-sampling is corrected by adjusting the predicted probabilities with previous probabilities. These methods are discussed in section 4.8.2. 1.5 Pre-processing data pre-processing has several purposes: - eliminate clearly inappropriate data items, such as name, Social Security number, street address, etc., that clearly do not affect the target variable, convert the data into the appropriate scale of measurement, especially the conversion of categorical (nominal scaled) data into interval scale, when appropriate, eliminate variables with highly distorted distributions and eliminate variables with highly distorted distributions. disguised as input, and blame missing values, although you can perform many cleanup tasks in sas Enterprise Miner, 1.5.1 Clearing data before launching data providers SAS Enterprise Miner Data sometimes considers variables of interlude, such as date of birth or income, as variables. If a variable such as the date of birth is entered as a character variable, it is considered by SAS Enterprise Miner as a categorical variable with many categories. To avoid this situation, it's best to extract the numerical variable from the variable symbol and then discard the original variable symbol from the dataset. Similarly, income is sometimes presented as a variable symbol. Character A can cost $20K ($20,000), B for $30K, etc. Another situation that requires data purification that cannot be done in SAS Enterprise Miner occurs when a target variable masquerades as a input variable. For example, a financial institution wants to simulate the loss of clients in their brokerage accounts. The model should predict the probability of exhaustion during the three-month interval in the future. The agency decides to develop a model based on actual depletion within three months. The goal is to predict attritions based on the demographic and income profiles of clients, and to balance activity on their brokerage accounts before the window. The binary target variable takes the value of 1 if the client attrits and 0 otherwise. If the client's balance on his brokerage account is 0 for two consecutive months, then he is considered an attriter, and the target value is set 1. If the dataset includes as a target variable (attrition/no and balances during the performance window, the balances on the account can be inadvertently considered as input variables. To prevent this, inputs that are indeed targeted variables disguised as input variables must be removed before launching SAS Enterprise Miner. 1.5.2 Cleaning up data after the launch of SAS Enterprise Miner Display 1.1 shows an example of a variable that is highly distorted. The variable is MS, which indicates the marital status of the client. The RESP represents the customer's response to the mail. It takes a value of 1 if the customer responds, and 0 otherwise. In this hypothetical sample, there are only 100 clients with a marital status of M (married) and 2900 with S (one). None of the married clients is a defendant. An unusual situation like this can lead to a marital variable playing a much larger role in the predictive model than is really warranted, because the model usually concludes that all married clients were unresponsive because they were married. The real reason there were no defendants among them is simply that there were so few married clients in the sample. Display 1.1 These variables can give false results when used in a model. These variables can be identified using the StatExplore node, set their roles to deviate in the input node, and drop them off the table with Drop. The filter node can be used to eliminate sightings with extreme values, although I do not recommend eliminating observations. Fixing them or limiting them instead may be better to avoid introducing any biases into the model settings. The Impute site offers a variety of methods for compromising missing values. These nodes are discussed in the next chapter. The application of missing values is necessary when using regression or neural network nodes. 1.6 Alternative Modeling Strategies Choice Modeling Strategy depends on the modeling tool and the amount of input used for modeling. Here are examples of two possible strategies when using a regression node. 1.6.1 Regression with a moderate number of input variables Pre-process data: - Exclude clearly inappropriate variables. If necessary, convert nominal input with too many levels into numerical interval input. If necessary, create composite variables (such as the average savings account balance for six months prior to the promotional campaign) from the original variables. This can also be done with SAS Enterprise Miner using the SAS Code. Next, use SAS Enterprise Miner to do these tasks: Transform input variables. Divide the modeling dataset into train, check, and testing samples (when the available data is large enough). Separation can be done before the change and transformation, that SAS Enterprise Miner automatically applies them to all parts of the data. You run the regression node with Stepwise option. 1.6.2 Regression Regression A large number of input variables Pre-process data: - Eliminate clearly inappropriate variables. If necessary, convert nominal input with too many levels into numerical interval input. Combine variables if necessary. Next, use SAS Enterprise Miner to do these things: Make a preliminary selection of the variable. (Note: This step is not included in Section 1.6.1.) Group categorical variables (collapse levels). Converting interval inputs. Divide the data set into trains, test and test samples. You run the regression node with Stepwise option. The steps in sections 1.6.1 and 1.6.2 are just two of many possibilities. For example, you can use the Decision Tree node to make variable choices and create fictitious variables to then use regression in the node. 1.7 Notes 1. Alan Agresti, Categorical Data Analysis (New York, NY: John Wiley and sons, 1990), 2. 2. If the client increased the savings deposits by $100, but reduced the settlement deposits by $20, the net increase is $80. Here, net means exclusion. Chapter 2: Start with Predictive Modeling 2.1 Introduction 2.2 Opening SAS Enterprise Miner 14.1 2.3 Creating a New Project in SAS Enterprise Miner 14.1 2.4 Sas Enterprise Miner Window 2.5 Creating SAS Source Source 2.6 Creating Process Flow Chart 2.7 Node Node Input Data 2.7 2.7.2 Data Section Node 2.7.3 Filter Node 2.7.4 File Import Node 2.7.5 Time Series Nodes 2.7.6 Merger Node 2.7.7 Application Node 2.8 Tools for Initial Data Study 2.8.1 Stat Explore Knot 2.8.2 Multi-Density Knot 2.8.3 Chart Explore Node 2.8.4 Variable Cluster Node 2.8.5 Cluster Node 2.8.6 Variable Node Of Choice 2.9 Data Modification Tools 2.9.1 Drop Node 2.9.2 Replacement Node 2.9.2 3 Sane Node 2.9.4 Interactive Node binning 2.9.5 Key Components Node 2.9.6 Conversion Of Node Variables 2.10 Utility Knots 2.10.1 SAS Code Knot 2.11 Appendix to Chapter 2 2.11.1 Type, Measurement scale, and number of variable levels 2.11.2 Eigenvalues , Eigenvectors, and the main components of 2.11.3 Cramer's V 2.11.4 Chi Square and V Kramer Stats calculation for continuous input 133 2.12 Exercise Notes 2.1 Introduction This chapter introduces you to THE SAS® Enterprise Miner™ 14.1 and some of the pre-processing and clearing tools (nodes) needed to develop data and project modeling forecasting. SAS Enterprise Miner modeling tools are not included in this chapter because they are widely used in chapters 4, 5 and 6. 2.2 Opening SAS Enterprise Miner 14.1 To launch SAS Enterprise Miner 14.1, click the SAS Enterprise Miner icon on your desktop.1 If you have a WorkStation configuration, the Welcome to Enterprise window opens, as shown on display 2.1. Display 2.1 2.3 Creating a new project in SAS Miner 14.1 When you select a new project, create New Project opens in the Enterprise Miner window. In this window, enter the name of the project and where you want to save the project. This example uses Chapters2 and C: The BookEM14.1EMProjects (the catalog where the project will be stored). Click on. A new window opens that shows information about the new project. Display 2.2 If you click Next, another window opens (not shown here). When you click on the finish, this window creates a new project and opens the SAS Enterprise Miner 14.1 interface window, showing the new project. 2.4 THE SAS Enterprise Miner Window Is a window in which you create a process flow diagram for a data mining project. The numbers on display 2.3 match the descriptions below the display. Display 2.3 (1) Menu Panel (2) Toolbar: This contains Enterprise Miner node icons. The icons displayed on the toolbar change according to the tab you select in the area specified by the group tabs (3). (3) Node (Tool): These tabs are for selecting different groups of nodes. The toolbar (2) changes depending on the group of the site you choose. If you select the Example tab on this line, you'll see icons for the app, data section, file import, filter, input, merging, and sample in (2). If you select the Explore tab, you'll see icons for Association, Cluster, DMDB, Count Explore, Link Analysis, Market Trash, Multiplot, Path Analysis, SOM/Kohonen, StatExplore, Variable Clustering and Variable Choice in (2). (4) Project Panel: It's for viewing, creating, deleting, and modifying data sources, diagrams, and package models. For example, if you want to create a data source (tell SAS Enterprise Miner where your data is and give information about variables, etc.), you'll click on the data sources and continue to work. To create a new chart, click on the Charts and continue to work. To open an existing chart, double-click on the chart you want. (5) Property Bar: In this panel, you can see project properties, data sources, diagrams, nodes, and package models by selecting them. In this example, nodes have not yet been created; hence you don't see them in display 2.3. You can view and edit the properties of any selected object. If you want to specify or change any settings in a node, such as a decision tree or a neural network, you should use a property bar. (6) Help panel: This displays a description of the property you choose in the Property panel. (7) State Bar: This indicates the state of the SAS Enterprise Miner task. (8) Toolbar Fast Access Buttons: These are quick access buttons to create a data source, create a diagram, run, etc. (9) Workspace Chart: This is used to create and run a process flow chart for a project with different nodes (tools) SAS Enterprise Miner. Project start-up code located in Project Start Code column in the panel (see Display 2.3). A window opens where you can enter the library path where the project's datasets are located. Project project start-up code 2.4 display. Display 2.4 The data for this project is in the C folder: TheBook EM14.1'Data-Chapter2. This is evidenced by the Libref TheBook. When you click Run Now, you create a library link to the path. You can check whether the library has been created successfully by opening the log window by clicking on the log tab. 2.5 Create a data source sas you need to create a data source before you start working on the SAS Enterprise Miner project. Once you've created a data source, it contains all the information related to your data: the catalog path to the file that contains the data file name, the names and scales of the variable measurements in the dataset, as well as the cost and solution matrix and the target profiles you specify. The Profit Matrix, also known as the weight of the solution in SAS Enterprise Miner 14.1, is used in solutions such as assigning a target class to monitor and evaluate models. This section shows how a data source is created that covers the main steps. For more information and information, check out the SAS Enterprise Miner Help menu. SAS Enterprise Miner retains all this information or metadata because of the different datasets in a folder called Data Sources in the project directory. To create a data source, click on the toolbar's quick access button or press the data sources in the project bar right, as shown in display 2.5. Display 2.5 When you press the Create a Data Source button, the Data Source Master window opens, and SAS Enterprise Miner encourages you to log into the data source. If you're using the SAS dataset in the project, use the default SAS table in the Source box and click Next. Then another window opens, which prompts you to know the location of the SAS dataset. When you click on the View, you open a window that shows a list of library links. This window appears on display 2.6. Display 2.6 Since the data for this project is in the Thebook library, double tap Thebook. The window opens with a list of all the datasets in this library. This window appears on display 2.7. Display 2.7 Select a dataset called NN_RESP_DATA and click OK. The Data Source Master window opens, as shown in display 2.8. Display 2.8 This display displays the libref and the name of the dataset. Click on. Another window opens, displaying the properties of the table. This is shown on display 2.9. Show 2.9 Click Next to show metadata Advisor options. This is shown on the 2.10 display. Display 2.10 Use the Metadata Advisor Options window to determine metadata. Metadata is data on data sets. It shows how each variable is used in the simulation process. Metadata contains information about the role of each variable, its measurement scale2, etc. This means that if the variable is numerical, its Measurements are marked as interval, no matter how many different values can have. For example, the numerical binary variable will also be initially based on the interval scale. If your target variable is binary in numerical form, it will be treated as a variable at scale intervals, and it will be treated as such in subsequent nodes. If the subsequent node is a regression node, SAS Enterprise Miner automatically uses the usual smallest square regression rather than logistical regression, which usually fits with a double target variable. In the basic version, all variable characters are assigned a nominal measurement scale, and all numerical variables are assigned an interval measurement scale. If you choose Advanced, SAS Enterprise Miner applies a little more logic because it automatically sets variable roles and measurement scales. If the variable is numerical and has more than 20 different values, SAS Enterprise Miner sets the scale of measurement (level) at intervals. You can also set up measurement scales if you choose an extended version. For example, by default, the Advanced option sets a scale for measuring any numerical variable to a nominal if it takes less than 20 unique values, but you can change that number by clicking to customize and setting the property to the class level score threshold (see Display 2.11) on the number it is than the default 20. For example, consider a numerical variable such as X, where X can be the number of times a credit card holder has been more than 60 days after payment in the last 24 months. In the dataset, modeling X takes 0, 1, 2, 3, 4, and 5 only. With the Advanced Advisor option, SAS Enterprise Miner will assign the X-rated measurement scale by default. But if you change the X level counting threshold from 20 to 3, 4 or 5, SAS Enterprise Miner will set the X measurement scale at intervals. Later this chapter gives a detailed discussion of the scale of measurements assigned when selecting the main, extended EA options with default property values, as well as the extended EA options with customizable properties. The 2.11 display shows the default settings for the EA's extended options. Display 2.11 One of the advantages of selecting an extended option is that SAS Enterprise Miner automatically sets the role of each non-emergency variable for the rejected. If any of the settings don't fit, you can change them later in the window shown on the 2.12 display. In this example, the class level threshold property has been changed to 10. I closed the Advanced Advisor Options box by clicking OK and then clicked on. This opens the window displayed on the 2.12 display. Display 2.12 This window shows a variable list table with variable names, model roles, and levels of measurement of variables in the dataset. In this example, the role of the variable resp model is cited as the goal. If you check the Stats box at the top of the variable list table, Advanced Data Source Master functions calculates important stats stats like the number of levels, the percentage of missing persons, minimum, maximum, average, standard deviation, skew, and curtosis for each variable. If you check the base field, the variable list table also shows which type (symbol or numerical) belongs to each variable. Display 2.13 shows a partial representation of these additional statistics and variable types. Display 2.13 If predictive modeling with sas enterprise miner 3rd edition pdf. predictive modeling with sas enterprise miner 3rd edition. predictive modeling with sas enterprise miner practical solutions for business applications pdf. predictive modeling with sas enterprise miner free download. predictive modeling with sas enterprise miner certification. advanced predictive modeling with sas enterprise miner. predictive modeling with sas enterprise miner third edition pdf