Application of data mining techniques using SAS software

Prof.Ass Mit'hat MEMA [1], MSc. Edlira KALEMI [2], Alma KONDI [3], Blerina SUBASHI [4]

[1]Rector of "Aleksandër Moisiu" University, mithatmema@uamd.edu.al
[2] Computer Science Departament, "Aleksandër Moisiu" University of Durrës, edlirakalemi@uamd.edu.al
[3] Specialist of Survey Methodology Sector, INSTAT, akondi@instat.gov.al
[4] Specialist of Demography Sector, INSTAT, bsubashi@instat.gov.al

ABSTRACT

Data mining has captured the hearts and minds of business analysts seeking a solution for exploring and modeling vastly larger, more complex and less well-behaved datasets. Exploratory data analysis, typically consisting of activities like statistical visualization, hypothesis generation, and introductory model fitting is a vital first step in any successful data mining venture. Exploratory data analysis produces direct benefits for data miners in enhanced understanding of data, improved clarity and confidence of the modeling results, and avoidance of pitfalls early in the process. By using data mining techniques to analyze the data that is accumulating and filling vast data warehouses, organizations can harness more insight from their large data stores to drive proactive decision making. SAS data mining software can surface patterns and trends in your data that you may never have thought to look for. This paper will review the usefulness of SAS T software for exploratory data analysis, interactive regression modeling, and advanced multidimensional data visualization

Keywords: Data Mining Platform, SAS

I. Introduction

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets.[1] These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers),

[1] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999); Pieter Adriaans and Dolf Zantinge, *Data Mining* (New York: Addison Wesley, 1996).

classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases), clustering (finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes).[2]

As an application, compared to other data analysis applications, such as structured queries (used in many commercial databases) or statistical analysis software, data mining represents a *difference of kind rather than degree*. Many simpler analytical tools utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis. For example, a user might hypothesize that a customer, who buys a hammer, will also buy a box of nails. The effectiveness of this approach can be limited by the creativity of the user to develop various hypotheses, as well as the structure of the software being used. In contrast, data mining utilizes a discovery approach, in which algorithms can be used to examine several multidimensional data relationships simultaneously, identifying those that are unique or frequently represented. For example, a hardware store may compare their customers' tool purchases with home ownership, type of automobile driven, age, occupation, income, and/or distance between residence and the store. As a result of its complex capabilities, two precursors are important for a successful data mining exercise; a clear formulation of the problem to be solved, and access to the relevant data.[3]

Reflecting this conceptualization of data mining, some observers consider data mining to be just one step in a larger process known as knowledge discovery in databases (KDD). Other steps in the KDD process, in progressive order, include data cleaning, data integration, data selection, data transformation, (data mining), pattern evaluation, and knowledge presentation.[4]

A number of advances in technology and business processes have contributed to a growing interest in data mining in both the public and private sectors. Some of these changes include the growth of computer networks, which can be used to connect databases; the development of enhanced search-related techniques such as neural networks and advanced algorithms; the spread

---

[2] For a more technically-oriented definition of data mining, see [http://searchcrm.techtarget.com/gDefinition/0,294236,sid11_gci211901,00.html]

[3] John Makulowich, "Government Data Mining Systems Defy Definition," *Washington Technology*, 22 February 1999, [http://www.washingtontechnology.com/news/13_22/tech_features/393-3.html]. [3] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999); Pieter Adriaans and Dolf Zantinge, *Data Mining* (New York: Addison Wesley, 1996).

[3] For a more technically-oriented definition of data mining, see [http://searchcrm.techtarget.com/gDefinition/0,294236,sid11_gci211901,00.html]

[3] John Makulowich, "Government Data Mining Systems Defy Definition," *Washington Technology*, 22 February 1999, [http://www.washingtontechnology.com/news/13_22/tech_features/393-3.html].

[4] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques* (New York:Morgan Kaufmann Publishers, 2001), p. 7.

of the client/server computing model, allowing users to access centralized data resources from the desktop; and an increased ability to combine data from disparate sources into a single searchable source.[5]

In addition to these improved data management tools, the increased availability of information and the decreasing costs of storing it have also played a role. Over the past several years there has been a rapid increase in the volume of information collected and stored, with some observers suggesting that the quantity of the world's data approximately doubles every year.[6] At the same time, the costs of data storage have decreased significantly from dollars per megabyte to pennies per megabyte. Similarly, computing power has continued to double every 18-24 months, while the relative cost of computing power has continued to decrease.[7]

II. Application

In this part, we want to build models that predict the credit status of credit applicants. We will use champion model to determine whether to extend credit to new applicants. The aims are to anticipate and reduce charge-offs and defaults which management has deemed are too high.

This dataset consist of 1000 applications and their resulting credit rating ("Good "or "Bad"). The binary target (dependent, response variable) is named Good_Bad. The other 20 variables in the dataset will serve as model inputs (independent, explanatory variables).

Table 1. List of variable of the application used

Variable Name	Role	Level	Description
Control	input	ordinal	Checking account status
Duration	input	interval	Duration of credit in months
History	input	ordinal	Credit history
Reason	Input	nominal	Purpose of credit
Amount	Input	interval	Credit amount
Savings	Input	ordinal	Savings / BONDS
Employment	Input	ordinal	Present employment since
Install rate	Input	interval	Installment rate as % of disposable income

[5] Pieter Adriaans and Dolf Zantinge, *Data Mining* (New York: Addison Wesley, 1996), pp.5-6.
[6] Ibid., p. 2.
[7] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999), p. 4.

Status	Input	nominal	Personal Status
Guarantor	input	nominal	Applicant has a guarantor
Present Resident	Input	interval	Present resident since - years
Real_Estate	Input	nominal	Applicant owns real estate
Age	Input	interval	Age in years
Other	input	nominal	Applicant has other plans
House	input	nominal	House
Num_Credits	input	interval	Number of existing credits at this bank
Job	input	ordinal	Nature of job
Num_Dependents	input	interval	Number of people for whom liable to provide maintenance
Telephone	input	binary	Applicant has phone in his or her name
Foreign	input	binary	Foreign worker
Response	target	binary	Credit rating is good

60% of the data in the dataset will be use to build the models (the training data). The remainder of the data will be used to adjust the models for over fitting with regard to the training data and to compare the models. The models will be judged primarily on their assessed profitability and accuracy and secondarily on their interpretability.

Attribute description of the dataset:

Control: Status of existing checking account

- 1: 0 DM
- 2: 0 - 200 DM
- 3: >= 200 DM salary assignments for at least 1 year
- 4: 0, no checking account
-

Duration: Duration of credit in months

History: Credit history

- 0: no credits taken/ all credits paid back duly
- 1: all credits at this bank paid back duly
- 2: existing credits paid back duly till now
- 3: delay in paying off in the past

- 4: critical account/ other credits existing (not at this bank)

Reason: Purpose of credit

- 0: car (new)
- 1: car (used)
- 2: furniture/equipment
- 3: radio/television
- 4: domestic appliances
- 5: repairs
- 6: education
- 7: vacation - does not exist?
- 8: retraining
- 9: business
- X: others

Amount: Credit amount

Savings: Savings / BONDS

- 1: < 100 DM
- 2.: 100 – 500 DM
- 3: 500 – 1000 DM
- 4: >= 1000 DM
- 5: unknown/ no savings account
-

Employment: Present employment since

- 1: unemployed
- 2: < 1 year
- 3: 1 – 4 year
- 4: 4 – 7 year
- 5: >= 7 year

Install rate: Installment rate in percentage of disposable income

Status: Personal status and sex

- 1: male : divorced/separated
- 2: female : divorced/separated/married
- 3: male : single
- 4: male : married/widowed
- 5: female : single

Guarantor: Other debtors / guarantors

- 1: none
- 2: co-applicant
- 3: guarantor

Present Resident: Present residence since

Real_Estate: Property

- 1: real estate
- if not 1 : building society savings agreement/life insurance
- if not 1/2 : car or other, not in attribute 6
- unknown / no property

Age: Age in years

Other: Other plans

- 1: bank
- 2: stores
- 3: others

House:

- 1: rent
- 2: own
- 3: for free

Num_Credits: Number of existing credits at this bank

Job: Nature of Job

- 1: unemployed/ unskilled - non-resident
- 2: unskilled - resident
- 3: skilled employee / official
- 4: management/ self-employed/highly qualified employee/ officer

Num_Dependents: Number of people being liable to provide maintenance for

Telephone: Applicant has phone in his or her name

- 1: none
- 2: yes, registered under the customers name

Foreign: Foreign worker

- 1:Yes

- 2: No

Each of the modeling nodes can make a decision for each case in the data to be scored based on numerical consequences that we can specify via a decision matrix and cost variables or constant cost. In Enterprise Miner is defined as part of the target profile for the target. For this example flow, we want to define a loss matrix that adjusts the models for the expected losses for each decision.

Table 2: Target values for defining the decisions

	Decisions:	
Target Values	**Accept**	**Reject**
Good	$0	$1
Bad	$5	$0

The rows of the matrix represent the target values, and the columns represent the decisions. For the loss matrix, accepting a bad credit risk is five times worse than rejecting a good credit risk. However, this loss matrix also says that you cannot make any money no mater what you do, so the result may be difficult to interpret. In fact, if you accept a good credit risk, you will make money, that is, you will have a negative loss.

And if you reject an applicant (good or bad), there will be no profit or loss aside from the cost of processing the application, which will be ignored. Hence it would be more realistic to subtract one from the first row of the matrix to give a more realistic loss matrix.

Table 3: Adjusted target values for defining the decisions

	Decisions:	
Target Values	**Accept**	**Reject**
Good	$-1	$0
Bad	$5	$0

This loss matrix will yield the same decision and the same model selections as the first matrix, but the summary statistics for the second matrix will be easier to interpret.

For a categorical target, such as Good_Bad, each modeling node can estimate posterior probabilities for each class, which are defined as conditional probabilities of the classes, given the input variables.

Enterprise Miner computes the posterior probabilities under the assumption that the prior probabilities in the decision data set, because the sample proportions of the classes in the training data set differ substantially from the proportions in the operational data set to be scored. The training data set that we will

use for modeling contains respectively70% good credit application and 30% bad credit risk applicants. The actual assumed proportion of good-to-bad credit risk applicants in the score data set is 90% and 10 %.

If we specify the correct priors in the target profile for good_bad, the posterior probabilities will be correctly adjusted no mater what the proportions are in the training data set.

When the most appropriate model for the applications bad credit id determined, the scoring code will be developed to a fictitious score data set that is another data set that contains 75 new applicants. Scoring new data that does not contain the target is in the end the result of data mining applications.

To analyze this case we thought the application uses SAS Enterprise Miner for data ining analysis using the data flow diagram, nodes of which will describe the main stages of this case until we reach the final goal, which is building a model through which it will then identify applications that contain bad loans at a bank.

So the main stages of this analysis are:

1. Importing historical data on loan applications to be analyzed in the project.

```
libname paper 'D:\Paper\aplikimi' ;

run ;

PROC IMPORT OUT= paper.dhenat

        DATAFILE= " D:\Paper\aplikimi\te_dhenat.xls"

        DBMS=EXCEL REPLACE;

    SHEET="te_dhenat";

    GETNAMES=YES;

    MIXED=NO;

    SCANTEXT=YES;

    USEDATE=YES;

    SCANTIME=YES;

RUN;
```

We define the attribute Good_Bad like a target attribute for which we will do analyses. On the basis of the target attribute, we define the decision matrix for the loss minimization from bad credits specified above. This matrix will be used from the nodes of the model to calculate the excepted losses.

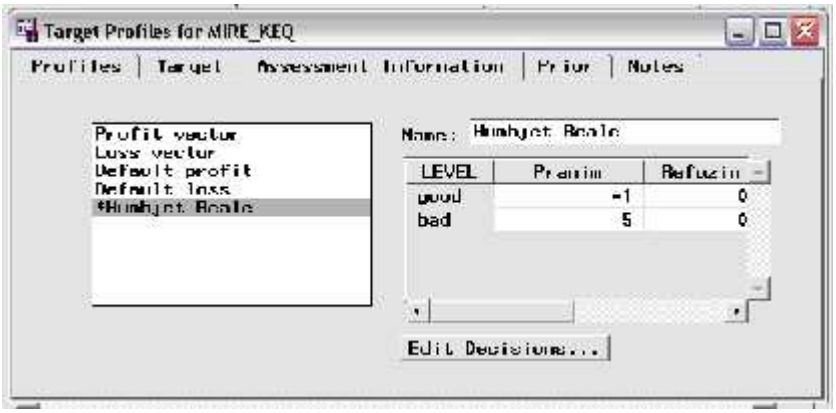

Figure 1: Defining the attribute Good_Bad is SAS software

Specify the following values in the Prior Probability cells of the Prior vector. The prior probabilities will be used to adjust the relative contribution of each class when computing the total and average loss.

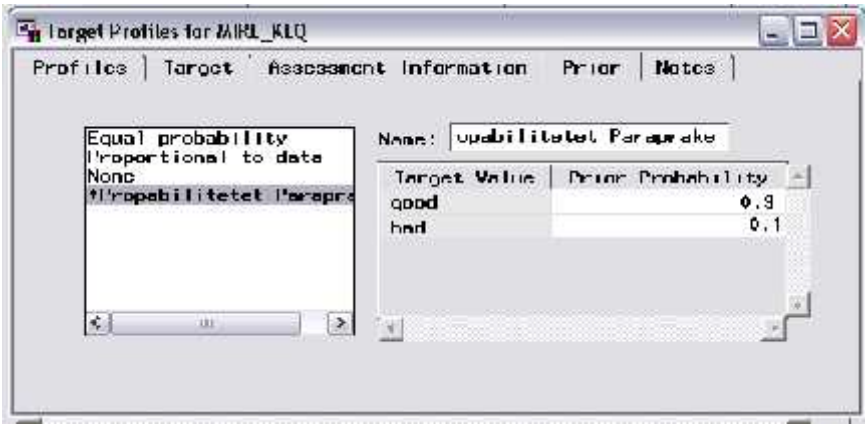

Figure 2: Specifying the following values in the Prior Probability cells

Examining Summary Statistics for the Interval and Class Variables. It is advisable to examine the summary statistics for the interval and class variables prior to modeling. Interval variables that are heavily skewed may need to be filtered and/or transformed prior to modeling. It is also important to identify the class and interval variables that have missing values. The entire customer case is excluded from the regression or neural network analysis when a variable attribute for a customer is missing. Although we tend to collect much information about our customers, missing values are invariably common in most data mining data marts. If there are variables that have missing values, you may want to consider imputing these variables with the Replacement node.

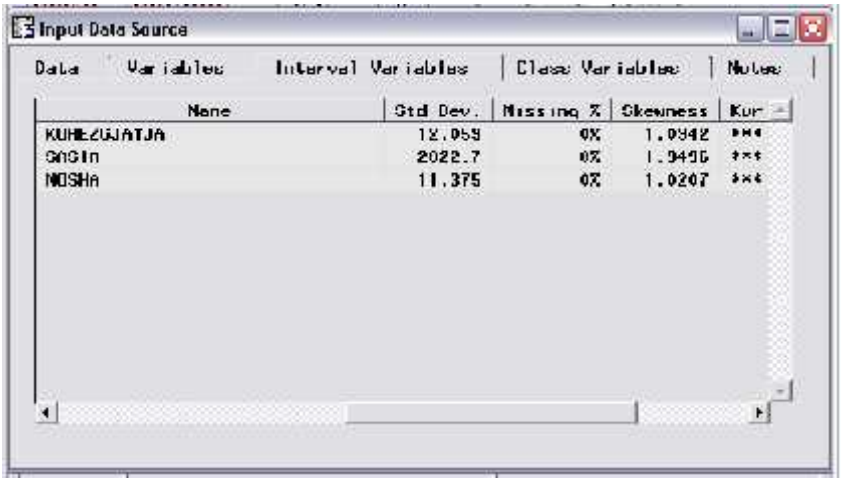

Figure 3: Examining Summary Statistics for the Interval and Class Variables.

The variable AMOUNT has the largest skewness statistic. The skewness statistic is 0 for a symmetric distribution. Also there is no missing value in interval variables.

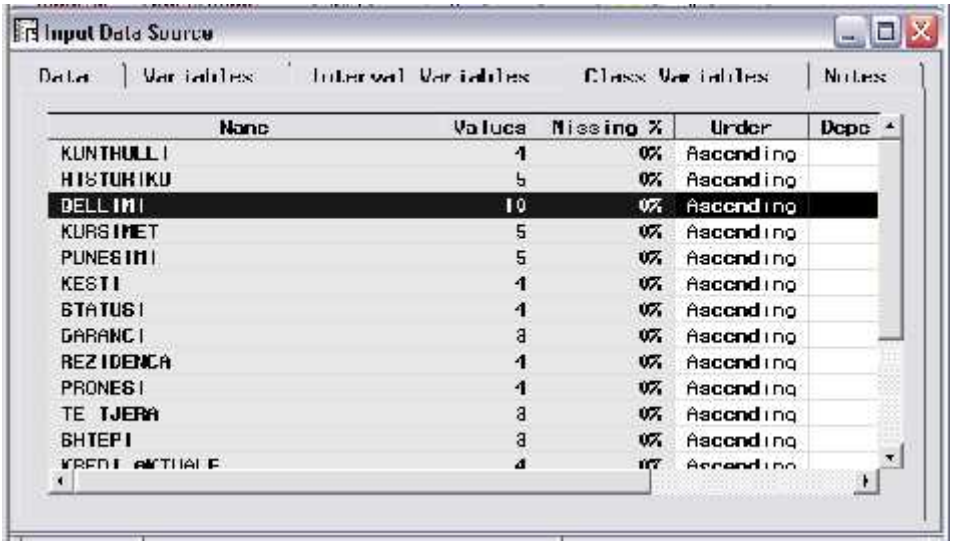

Figure 4: Consider imputing variables with the Replacement node

2. Creating Training and Validation Data Sets

In data mining, one strategy to assess model generalization is to partition the input data source. A portion of the data, called the training data, is used for model fitting. The rest is held out for empirical validation. The hold-out sample itself is often split into two parts: validation data and test data. The validation data is used to help prevent a modeling node from over fitting the training data (model fine-tuning), and to

compare prediction models. The test data set is used for a final assessment of the chosen model. Because there are only 1,000 customer cases in the input data source, only training and validation data sets will be created. The validation data set will be used to choose the champion model for screening new credit applicants based on the model that minimizes loss. Ideally, you would want to create a test data set to obtain a final, unbiased estimate of the generalization error of each model.

To create the training and validation data sets, follow these steps:

- Allocate 60% of the input data to the training data set and 40% of the input data to the validation data set.
- Create the partitioned data sets using a stratified sample by the target GOOD_BAD. A stratified sample helps to preserve the initial ratio of good to bad credit applicants in both the training and validation data sets. Stratification is often important when one of the target event levels is rare.

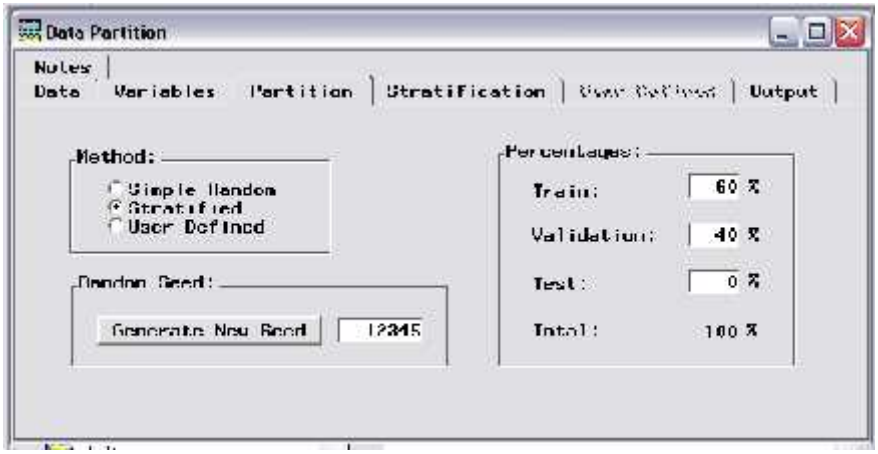

Figure 5**:** Allocate 60% of the input data to the training data set and 40% of the input data to the validation data set.

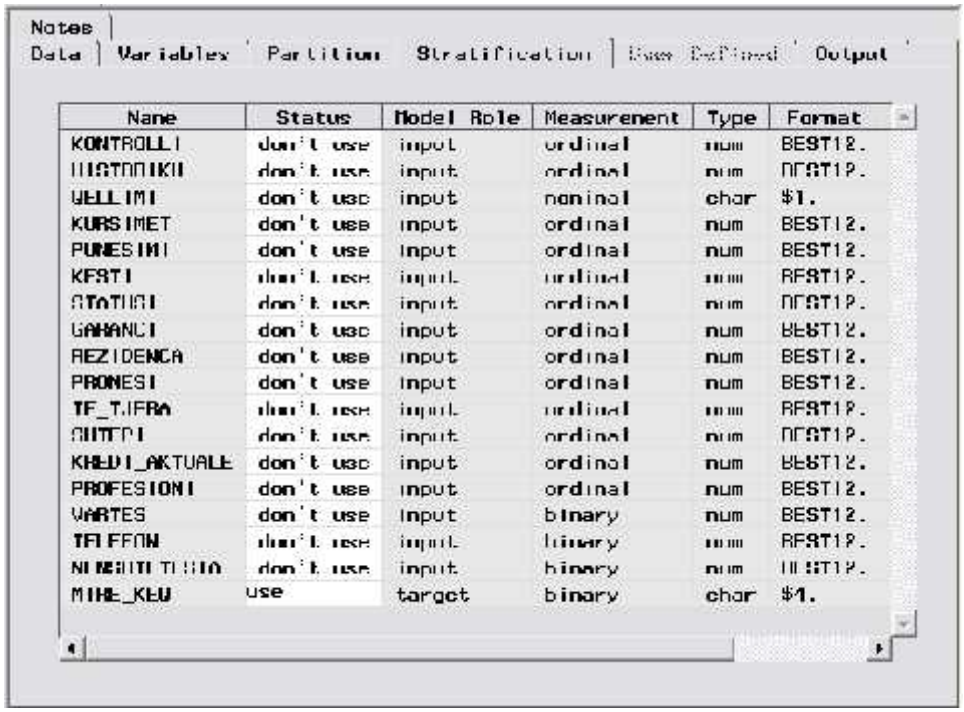

Figure 6: Create the partitioned data sets using a stratified sample by the target GOOD_BAD

3. Creating Variable Transformations

The data is often useful in its original form, but transformations may help to maximize the information content that you can retrieve. Transformations are useful when you want to improve the fit of a model to the data.

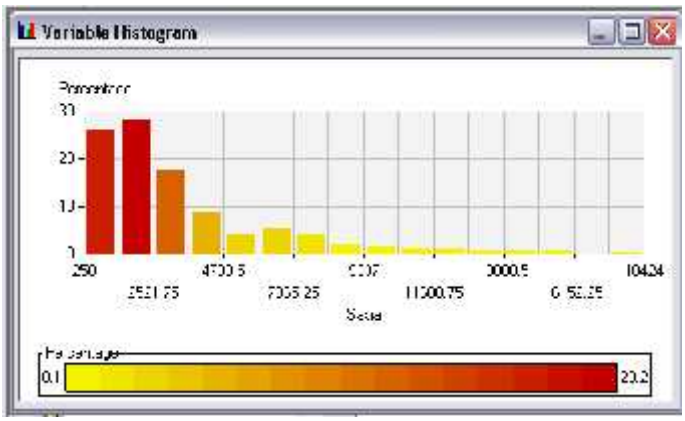

Figure 7: View the distribution of Amount:

Notice that the distribution for AMOUNT is skewed heavily to the right. The extreme values may cause imprecision in the parameter estimates. For this reasons we create a new input variable that maximizes the

normality of amount. The Maximize Normality transformation chooses the transformation from a set of best power transformations that yields sample quintiles that are closest to the theoretical quintiles of a normal distribution. We see that skewness statistic is reduced from 1.95 for the attribute "Sasia" in 0.13 in the transformed variable "Sasia_3S0". The Keep column identifies the variables that will be kept (Yes) and those that will not be kept (No). The Keep status for the original variable Quantity is set automatically into No when the transformation is applied.

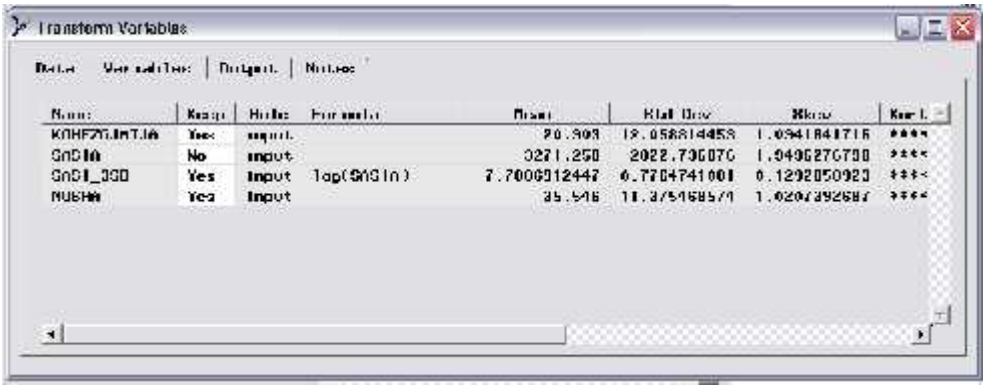

Figure 8: The Maximize Normality transformation

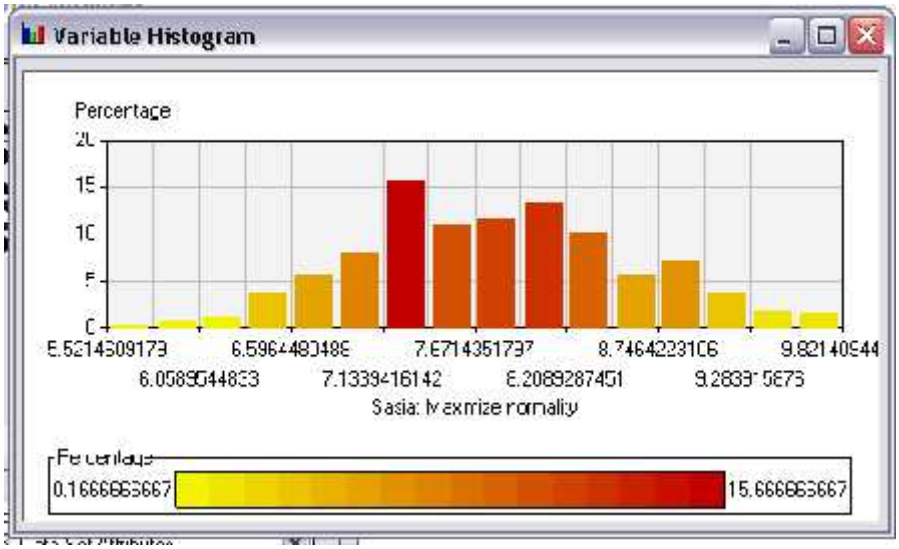

Figure 9: View the distribution for Sasi_3S0:

Notice that the distribution for the log of AMOUNT is fairly symmetrical.

Now create an ordinal grouping variable from an interval variable: With the Transformation node, you can easily transform an interval variable into a group variable. Because you are interested in the credit worthiness of certain age groups, create an ordinal grouping variable from the interval input AGE.

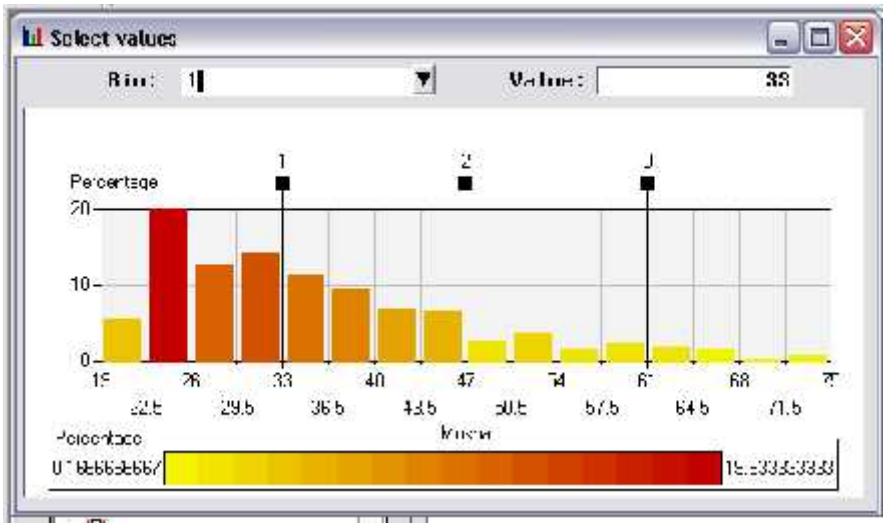

Figure 10: Grouping variable from an interval variable

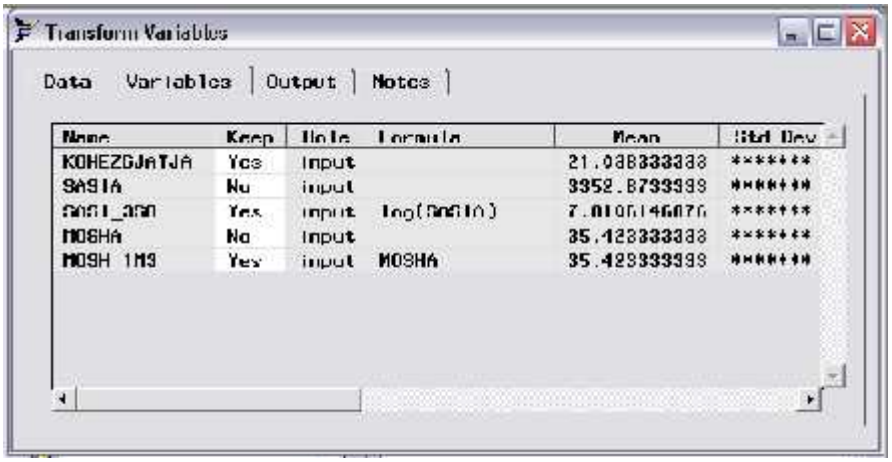

Figure 11: Transform an interval variable into a group variable

4. Assessing the Models

The *Assessment* node enables you to judge the generalization properties of each predictive model based on their predictive power, lift, sensitivity, profit or loss, and so on. Assessment statistics are automatically calculated by each modeling node during training. The *Assessment* node assembles these statistics, which enables you to compare the models with assessment charts. After modeling, each validation data can be assigned a degree of likelihood to be accepting (customer with good credit).

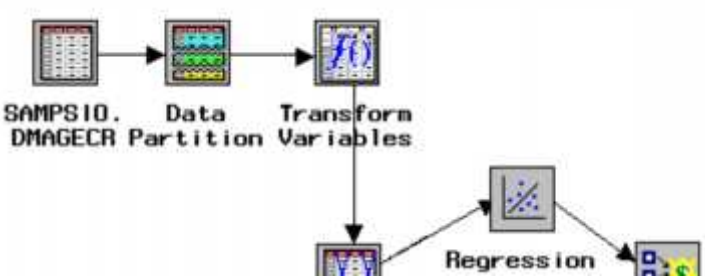

Figure 12: Assessment node

The purpose of predictive modeling is applying the model to new data. If you are satisfied with the performance of the Regression/Tree model, then you can use this model to screen (score) the credit applicants. If you are not satisfied with this model, cycle back through the desired components of the sample, explore, modify, model, and assess methodology to try to obtain a better predictive model.

III. Benefits using SAS in data mining

Every organization accumulates huge volumes of data from a variety of sources on a daily basis. Data Mining is an iterative process of creating predictive and descriptive models, by uncovering previously unknown trends and patterns in vast amounts of data from across the enterprise, in order to support decision making. Text mining applies the same analysis techniques to text-based documents. The knowledge gleaned from data and text mining can be used to fuel strategic decision making.

- **Support the entire data mining process with a broad set of tools.** Regardless of your data mining preference or skill level, SAS provides flexible software that addresses complex problems. Going from raw data to accurate, business-driven data mining models becomes a seamless process, enabling the statistical modeling group, business managers and the IT department to collaborate more efficiently.
- **Build more models faster with an easy-to-use GUI.** The process flow diagram environment of SAS dramatically shortens model development time for both business analysts and statisticians. SAS includes an intuitive user interface that incorporates common design principles established for SAS software and additional navigation tools for moving easily around the workspace. The GUI can be tailored for all analysts' needs via flexible, interactive property sheets, code editors and display settings.
- **Enhance accuracy of predictions and easily surface reliable business information.** Better performing models with new innovative algorithms enhance the stability and accuracy of predictions, which can be verified easily by visual model assessment and validation metrics. Both analytical and business users enjoy a common, easy-to-interpret visual view of the data mining process. Predictive results and assessment statistics from models built with different approaches can be displayed side by side for easy comparison. The created diagrams serve as self-documenting templates that can be updated easily or applied to new problems without starting over from scratch.

o **Ease the model deployment and scoring process.** Scoring – the process of applying a model to new data – is the end result of many data mining endeavors. SAS automates the tedious scoring process and supplies complete scoring code for all stages of model development in SAS, C, Java and PMML. The scoring code can be deployed in a variety of real-time or batch environments within SAS, on the Web or directly in relational databases. The outcome is faster implementation of data mining results.

IV. Conclusions

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Exploratory data analysis produces direct benefits for data miners in enhanced understanding of data, improved clarity and confidence of the modeling results, and avoidance of pitfalls early in the process.

Businesses and various enterprises can use SAS software to build models in decision making, in order to take better decisions and be more efficient in strategic actions. This software can allow to build models like tree, regressions etc. that can help businesses in this process. In order to work with these kinds of projects, team members have to have knowledge in decision making management and to have communication abilities in communicating with the stake holders of the enterprise. In order to have an efficient project, members of these teams can be from the field of statistics, computer science and management. This is the right combination of knowledge that works to build an appropriate SAS model.

There are a lot of benefits using SAS in data mining that a decision maker have to take into account, like SAS is able to build more models faster with an easy-to-use GUI, enhance accuracy of predictions and easily surface reliable business information, support the entire data mining process with a broad set of tools, ease the model deployment and scoring process. So we advice using SAS software in model like data mining where it is necessary to look into the data and to see if some hidden relationship or unknown links are there.

References

[1] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999); Pieter Adriaans and Dolf Zantinge, *Data Mining* (New York: Addison Wesley, 1996).

[2] For a more technically-oriented definition of data mining, see: http://searchcrm.techtarget.com/gDefinition/0,294236,sid11_gci211901,00.html

[3] John Makulowich, "Government Data Mining Systems Defy Definition," *Washington Technology*, 22 February 1999, [http://www.washingtontechnology.com/news/13_22/tech_features/393-3.html].

[4] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques* (New York:Morgan Kaufmann Publishers, 2001), p. 7.

[5] Pieter Adriaans and Dolf Zantinge, *Data Mining* (New York: Addison Wesley, 1996), pp.5-6.

[6] Ibid., p. 2.

[7] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999), p. 4.

[8] http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/benefits.html