Data Exploration in SAS EM

Dr. Goutam Chakraborty

Data Exploration

- GIGO principle is very important in data mining projects
- It is imperative that as an analyst, you spend **substantial time** (before running any modeling) on checking, exploring and understanding data.
- At the minimum explore following:
  - Summary statistics (including minimum and maximum) for numeric variables
    - Missing values, extreme values etc.
  - Categories (classes) for discrete variables
  - Plots
    - Univariate (histograms)
    - Bivariate (relationship with target variable and other input variables)

Demonstration of SAS EM Nodes for Data Exploration

- DMDB
- Multiplot
- Stat Explore
- Graph Explore

DMDB Tool (Explore Tab)

- Use Data Mining Data Base (DMDB) to get a **quick** view of:
  - Summary statistics for numerical variables
  - Number of categories for class variables
  - Extent of missing values in data
- DMDB node can be used anywhere in a process flow diagram

Multiplot Node (Explore Tab)

- Use MultiPlot to **quickly** visualize your data from a wide range of perspectives.
- The MultiPlot node creates the following types of charts:
- **Bar Charts**:
  - Histogram of each input and target.
  - Bar chart of each input versus each class target.
  - Bar chart of each input grouped by each interval target.
- **Scatter Plots**:
  - Plot of each interval input versus the target.
  - Plot of each class input versus the target.

StatExplore Node (Explore Tab)

- It is a multipurpose tool that you use to examine variable distributions and statistics in your data sets.
  - 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.
- In predictive modeling applications, the number of variables may be quite large. A challenge in using graphics is to reduce the number of displayed attributes so that the plot is readable, meaningful, and less resource intensive.
  - As a result, the StatExplore node works to select only the most important variables for automated display.

Graph Explore Node (Explore Tab)

- Use Graph Explore tool to **interactively** visualize your data from a wide range of perspectives.
  - If the Graph Explore node follows a node that exports a data set in the process flow, then it uses either a sample (default) or the entire data set as input.
  - The resulting plot is fully interactive - you can rotate a chart to different angles and move it anywhere on the screen to obtain different perspectives on the data.
    - You can also probe the data by positioning the cursor over a particular bar within the chart. A text window displays the values that correspond to that bar.
  - Your exploratory graphs are persisted when the Graph Explore Results window is closed. When you re-open the Graph Explore Results window the persisted graphs are recreated.

Charity Direct Mail Demonstration

Analysis goal:

A veterans' organization seeks continued contributions from lapsing donors. Use lapsing-donor responses from an earlier campaign to predict future lapsing-donor responses.

...

Charity Direct Mail Demonstration

Analysis goal:

A veterans' organization seeks continued contributions from lapsing donors. Use lapsing-donor responses from an earlier campaign to predict future lapsing-donor responses.

Analysis data:

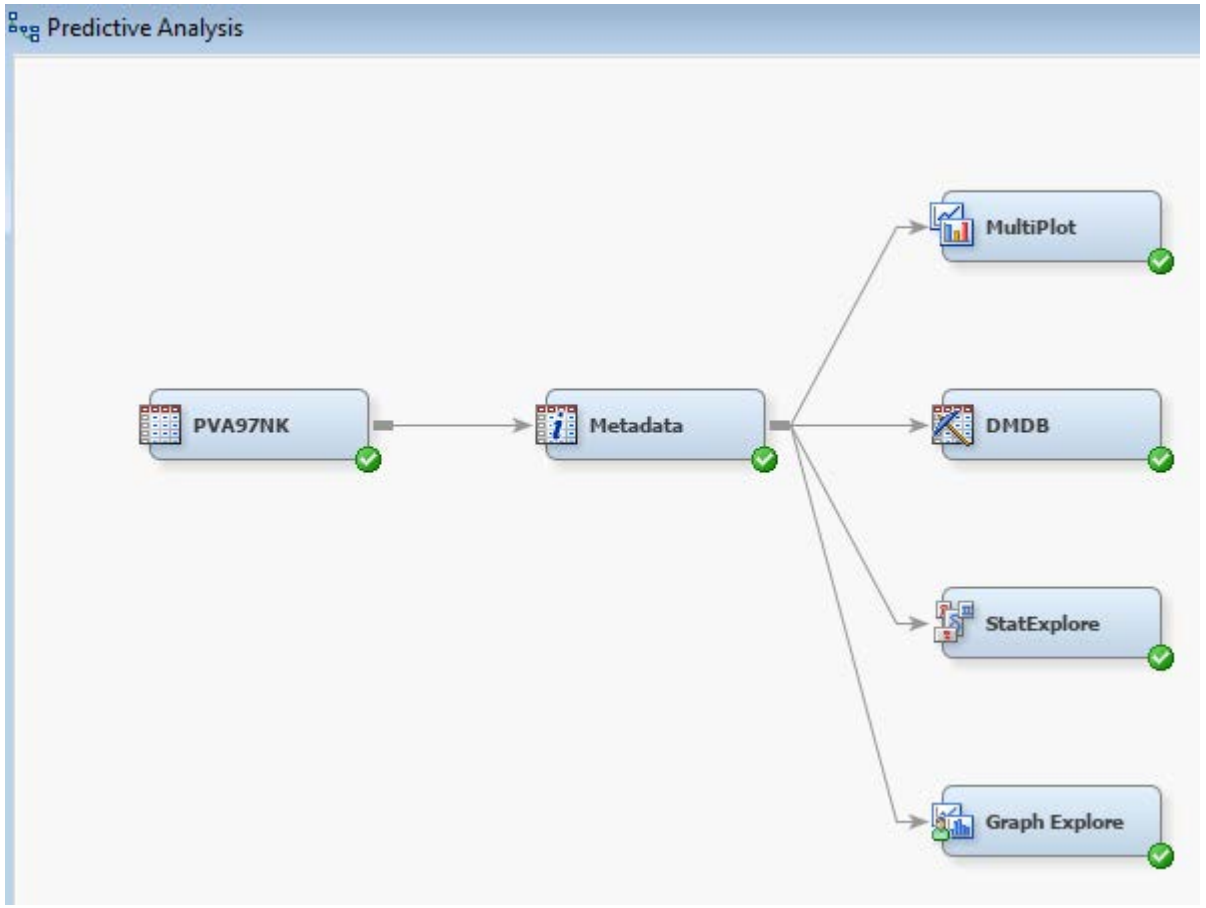
- **extracted from previous year's campaign**
- **sample balances response and non-response rate**
- **actual response rate of approximately 5%**

Data Set Name: PVA97NK

Variables and Their Roles

Name	Model Role	Measurement Level	Description
DemAge	Input	Interval	Age
DemCluster	Input	Nominal	Demographic Cluster
DemGender	Input	Nominal	Gender
DemHomeOwner	Input	Binary	Home Owner
DemMedHomeValue	Input	Interval	Median Home Value Region
DemMedIncome	Input	Interval	Median Income Region
DemPctVeterans	Input	Interval	Percent Veterans Region
GiftAvg36	Input	Interval	Gift Amount Average 36 Months
GiftAvgAll	Input	Interval	Gift Amount Average All Months
GiftAvgCard36	Input	Interval	Gift Amount Average Card 36 Months
GiftAvgLast	Input	Interval	Gift Amount Last
GiftCnt36	Input	Interval	Gift Count 36 Months
GiftCntAll	Input	Interval	Gift Count All Months
GiftCntCard36	Input	Interval	Gift Count Card 36 Months
GiftCntCardAll	Input	Interval	Gift Count Card All Months
GiftTimeFirst	Input	Interval	Time Since First Gift
GiftTimeLast	Input	Interval	Time Since Last Gift
ID	ID	Nominal	Control Number
PromCnt12	Input	Interval	Promotion Count 12 Months
PromCnt36	Input	Interval	Promotion Count 36 Months
PromCntAll	Input	Interval	Promotion Count All Months
PromCntCard12	Input	Interval	Promotion Count Card 12 Months
PromCntCard36	Input	Interval	Promotion Count Card 36 Months
PromCntCardAll	Input	Interval	Promotion Count Card All Months
StatusCat96NK	Input	Nominal	Status Category 96NK
StatusCatStarAll	Input	Binary	Status Category Star All Months
TARGET_B	Target	Binary	Target Gift Flag
TARGET_D	Rejected	Interval	Target Gift Amount

Create a Project, Add Data Source (PVA97NK)…



What if I Want to Use TargetD instead of TargetB?

- Once TargetD has been assigned to a rejected role, it is not possible to change that role within a node such as StatExplore.
- Use Meta Data node to make such role changes and then analyze the data.
- Drag another **Meta Data** Node (under **Utility** tab) and place it to the right of the PVA97NK data node.
- Connect data node to metadata node
  - Click on the **ellipsis button** for **Train** (under **Variables**) on Meta Data node property panel.
  - Click and change New Role of **TargetD** to **Target** and **TargetB** to **Rejected**. Click OK.
- Add another **Multiplot**, **StatExplore** and **GraphExplore** (under **Explore** tab) node and connect these to the Metadata Node.
- In the Multiplot, change type of charts to **Both**. Run the Multiplot node.
- Run the StatExplore and GraphExplore node.