Paper 154-2008

Understanding Your Customer: Segmentation Techniques for Gaining Customer Insight and Predicting Risk in the Telecom Industry

Glendon Cross, AT&T Corporation
Wayne Thompson, SAS Institute, Inc.

ABSTRACT

The explosion of customer data in the last twenty years has increased the need for data mining aimed at customer relationship management (CRM) and understanding the customer. It is well known that the telecom sector consists of customers with a wide array of customer behaviors. These customers pose different risks, making it imperative to implement different treatment strategies to maximize shareholder profit and improve revenue generation. Segmentation is the process of developing meaningful customer groups that are similar based on individual account characteristics and behaviors. The goal of segmentation is to know your customer better and to apply that knowledge to increase profitability, reduce operational cost, and enhance customer service. Segmentation can provide a multidimensional view of the customer for better treatment targeting. An improved understanding of customer risk and behaviors enables more effective portfolio management and the proactive application of targeted treatments to lift account profitability. In this paper we outline several segmentation techniques using SAS® Enterprise Miner™.

INTRODUCTION

Rapid advances in computer technology and an explosion of data collection systems over the last thirty years make it more critical for business to understand their customers. Companies employing data driven analytical strategies often enjoy a competitive advantage.

Many organizations across several industries widely employ analytical models to gain a better understanding of their customers. They use these models to predict a wide array of events such as behavioral risk, fraud, or the likelihood of response. Regardless of the predictive variable, a single model may not perform optimally across the target population because there may be distinct segments with different characteristics inherent in the population. Segmentation may be done judgmentally based on experience, but such segmentation schema is limited to the use of only a few variables at best. True multivariate segmentation with the goal of identifying the different segments in your population is best achieved through the use of cluster analysis. Clustering and profiling of the customer base can answer the following questions:

- Who are my customers?
- How profitable are my customers?
- Who are my least profitable customers?
- Why are my customers leaving?
- What do my best customers look like?

This paper discusses the use of SAS Enterprise Miner to segment a population of customers using cluster analysis and decision trees. Focus is placed on the methodology rather than the results to ensure the integrity and confidentiality of customer data. Other statistical strategies are presented in this paper which could be employed in the pursuit of further customer intelligence. It has been said that statistical cluster analysis is as much art as it is science because it is performed without the benefit of well established statistical criteria.

DATA DESCRIPTION and PREPARATION

In clustering or unsupervised learning it is important to exclude variables which are unnecessary or irrelevant to the stated objectives. The basis variables used in the cluster detection algorithm should represent characteristics of customers which are easily measured and understood. Potential basis variables include demographic, delinquency, usage, billing, and payment variables. A large number of variables were available for clustering but these were reduced to a smaller set which had the potential to be analytically and strategically useful. The basis variables used in the analysis where chosen based on the following characteristics:

- are meaningful to the analysis objective
- have low correlation between input variables
- are predominantly interval variables
- have low kurtosis and skewness

Choosing meaningful inputs makes cluster interpretation easier. Low correlation between input variables produces more stable clusters. Class input variables have a tendency to dominate cluster formation. Low kurtosis and skewness reduces the chance of creating small outlier clusters.

The data used in this analysis is a random sample of residential customers. The original data contained 250,000 observations and 276 variables. The response (or target) variable (BAD) indicates whether a customer became delinquent during the performance period of six months. The sample contained a 4.0% bad rate. Variables with little relationship to the target were excluded from the analysis. For the cluster solution, focus is placed on customer attributes which are driven by the customers and not by AT&T. This self selection avoids the problem of internal policy changes impacting cluster membership. For example, the timing of account treatment is influenced by AT&T policy and therefore would not be included in the cluster algorithm.

DATA OVERSAMPLING

Since the data is dominated by records with target value of good (BAD=0), the models built will be biased towards predicting good for the target. To compensate for the rare proportion of BADs in the raw sample, over-sampling of the data was done to produce a more balanced data set. Over-sampling rare classes often leads to more accurate predictions. The data was over-sampled using the Sample node to retain all BAD observations and a random sample of good observations. The final BAD proportion was increased to 25% from the original 4% BAD rate.

Prior to developing predictive models, it is important to specify the correct priors using the Decision node to correctly adjust model predictions regardless of what the proportions in the training set are. If no prior probabilities are used, the estimated posterior probability for the bad event class will be too high.

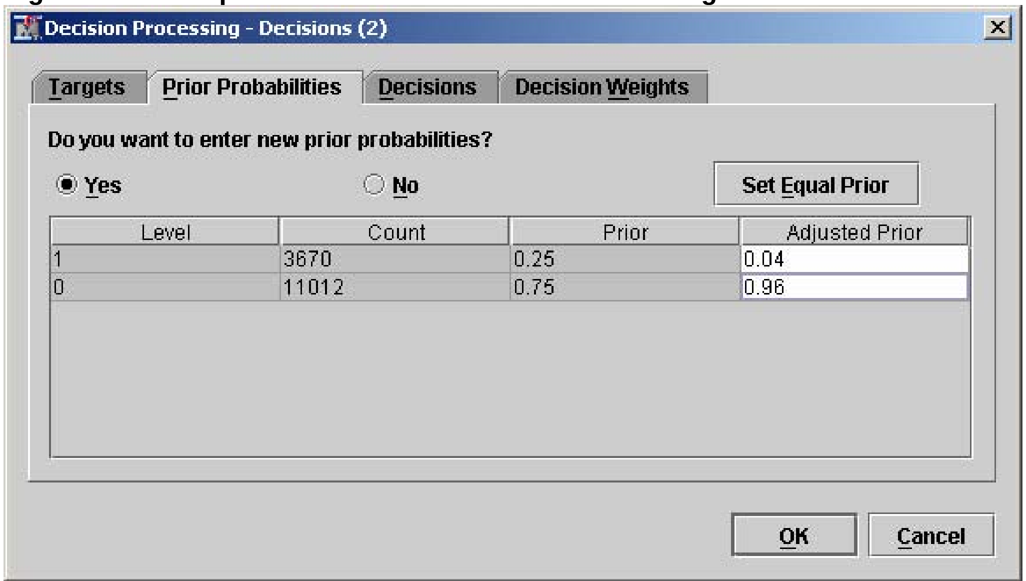
It is also common to tune and assess predictive models based on the profit or loss consequence of a model-based decision. Unfortunately, accurate specification of this profit or loss consequence is a difficult--if not impossible--task. In lieu of a traditional profit matrix, SAS Enterprise Miner allows the analyst to specify a diagonal profit matrix with elements equal to the inverse of the prior distribution for each outcome.

Let π_i = prior distribution for target outcome i.

The inverse prior profit matrix for a binary target has the form:

```
Outcome         Decision
                  1         0
             -------------------
    1      |    1/π1        0
    0      |     0        1/π0
```

Figure 1 - Prior Specifications for Each Level of the Target BAD.



For a binary target, the inverse prior profit matrix assigns decision 1 to each case with a posterior probability in excess of π1. Intuitively, this means that cases predicted more likely than average to have the primary outcome will receive the primary decision. An interesting side note to the inverse prior profit matrix is that the overall average profit calculated using the profit matrix equals, on the average, the KS statistic plus one. In this way, the model with the highest inverse prior profit will also, on the average, have the highest KS statistic.

The data was partitioned in a proportion of 60 to 40 for the training and validation for predictive modeling purposes. The Data Partition node automatically uses the target BAD as a stratification variable to ensure equal distribution in both the training and validation data. The training data is used for learning while the validation data is used to prevent overfitting and for model selection. SAS Enterprise Miner also supports a test partitioned data set to help determine how well the model generalizes on true hold out data.

PRELIMINARY ANALYSIS

Before embarking upon a statistical cluster analysis, there are certain data cleaning and data preparation steps that should be performed. These steps include:

- Handle invalid and missing values
- Remove extreme outliers
- Standardize variables
- Transform variables
- Decide what type of cluster analysis to use

Typically, the majority of the analyst time is spent on data preprocessing. For this analysis the Stat Explore node was used to examine the distributions and statistics of the variables in the raw dataset. Invalid and missing values were identified for each variable for further processing. If an attribute contains a high proportion of missing values that is not fully understood, then it should probably be removed from the analysis. Attributes with a lower proportion of missing values should be verified and handled in some manner. Missing values should be replaced with values that make sense. Fortunately, the data used in this analysis contains a small number of variables with missing values that should have been zero. These missing values were recoded to zero using the Replacement node.

Extreme values in the data can lead to small isolated clusters. SAS Enterprise Miner provides several methods to remove outliers. In this analysis outliers outside a user defined cutoff were removed from the analysis. It is important to check the number of observations removed from the analysis since you do not want to remove more than 10 percent of the raw data. Eight percent of the raw observations were removed from this analysis.

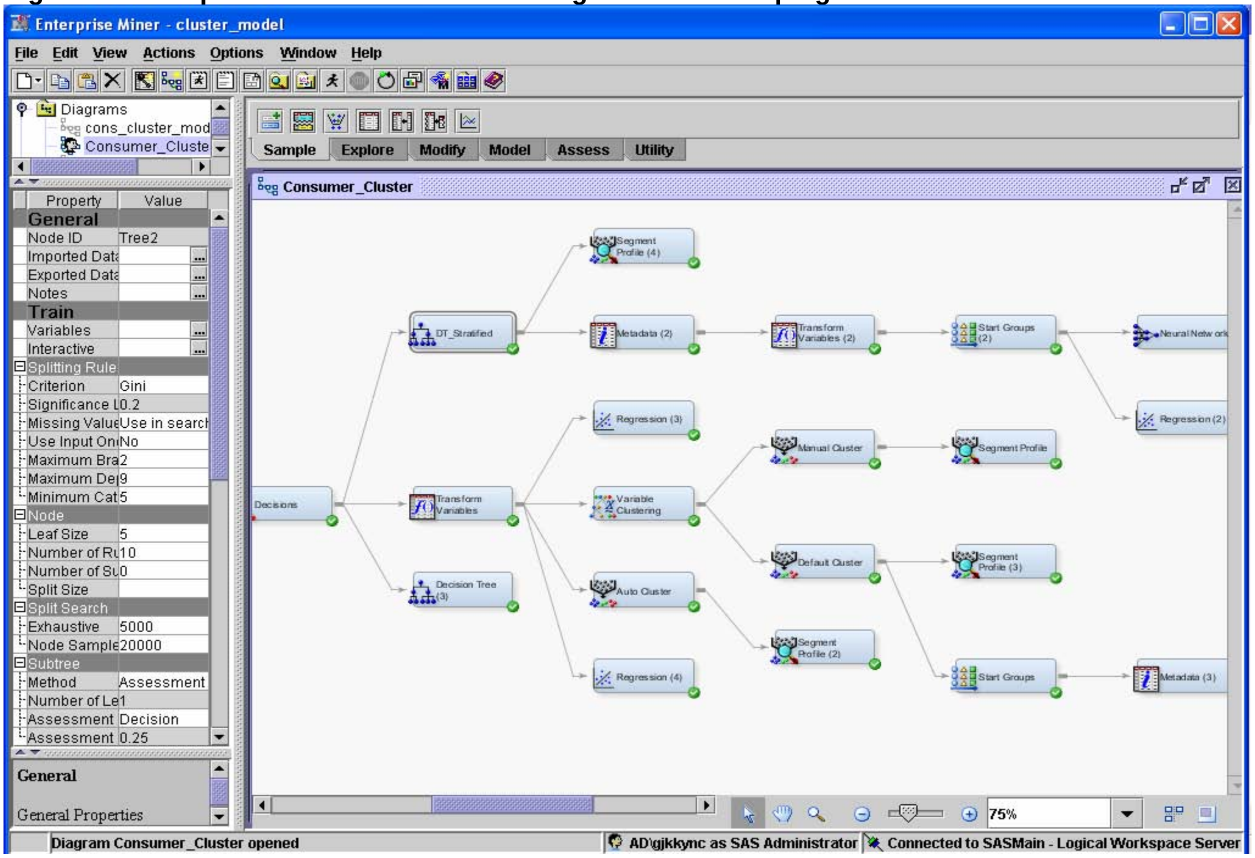
It is very also important to standardize the variables in your analysis. Similarity measures are very sensitive to different scales of measurement of the input variables. The K-means clustering method may produce unexpected results if the variables are measured on different scales. Input variables with large variances tend to have more influence on the cluster results than do variables with smaller variances. All variables used in the clustering algorithm were standardized to a mean of 0 and standardization of 1 using the internal standardization option of the SAS Enterprise Miner Cluster node.

For variables having highly skewed distributions, a small number of values will have greater influence on the results. To remedy this situation, it is recommended that the data be transformed to produce a more normal distribution prior to clustering. This will result in models that will produce a better fit.

CLUSTERING

Figure 2 shows a partial view of the Enterprise Miner process flow diagram used to develop the clusters. Primary emphasis is placed on clusters developed using the Cluster node and segments from the Decision Tree node. Benchmark logistic regression and decision tree models were also developed using all of the data. In the final analysis each of these modeling methods should be compared using common hold out data to determine which provides the best classifier. Further exploration and profiling of the results will reveal how effective each method is at identifying customer segments.

Figure 2 - Enterprise Miner Process Flow Diagram for Developing Candidate Models.



Variable Selection

A common strategy prior to developing clusters or predictive models is to reduce the number of variables to a more manageable uncorrelated, non-redundant set. Since the data contains variables with a high degree of collinearity, a variable reduction method is necessary for selecting a subset of orthogonal variables to be used in the cluster solution. The original data contained 250,000 observations and 276 variables. Many of these variables such as total bill balances over the last 1, 3, and 6 month are expected to be highly correlated. Variable clustering is one dimension reduction approach that finds groups of inputs which are correlated with their own cluster and also are uncorrelated with variables in other clusters. An advantage of variable clustering is an actual subset of the variable inputs can be output rather than transformations of the inputs. SAS/STAT includes the Varclus procedure, which groups variables together into clusters by their relative values. SAS Enterprise Miner now includes a Variable Clustering node that uses Proc Varclus with an optional fast two-stage algorithm for extremely large problems. Irrelevant variables were excluded from the input variables used in the Varclus procedure. The Variable Clustering node attempts to produce non-overlapping clusters with a minimum loss of information. Each cluster is reduced to essentially one dimension where one variable can act as a representative of the group.

Forty one potential variables were selected as basis variables based on the business objective and the potential to create interpretable stable clusters. The Variable Clustering node was run with ten clusters requested. The best variable from each cluster was chosen as an input for subsequent observational clustering using the Cluster node of SAS Enterprise Miner. Since the best variable setting was used for the variable selection property of the Variable Clustering node, the best variable in each cluster is the variable that has the lowest 1-R^2 ratio.

Table 1 lists the input variables selected within each cluster. These variables will be passed as inputs to the successor nodes in the process flow diagram. The third column shows how the variable correlates with its own cluster while the fifth column shows the correlation of that variable with the next closest cluster.

Table 1 - Cluster Plot of the Proximity of the Ten Variable Clusters.

Cluster	Variable	R-Square With Own Cluster Component	Next Closest Cluster	R-Sqaure with Next Cluster Component	1-R2 Ratio	Variable Selected
CLUS1	Past Due Bal	0.883674997	CLUS7	0.420419745	0.2007056	YES
CLUS10	Total Pmts Adjustments	1	CLUS2	0.226861707	0	YES
CLUS2	Toll Charges Last 3 Months	0.89224877	CLUS8	0.320516491	0.158578138	YES
CLUS3	Number of Broken Promise	0.793499267	CLUS1	0.155372392	0.244487312	YES
CLUS4	Months in Service	1	CLUS7	0.106836788	0	YES
CLUS5	Local Service Amt	0.885541721	CLUS8	0.186924541	0.140772025	YES
CLUS6	Value Score	1	CLUS3	0.174427992	0	YES
CLUS7	Percent Paid Attribute	0.966733496	CLUS1	0.345500412	0.050827388	YES
CLUS8	Unreg Balance	0.974952327	CLUS2	0.374892536	0.040069387	YES
CLUS9	Number of Days Since Last Pmt	1	CLUS1	0.261060161	0	YES

The cluster constellation plot is also useful to understand how the clusters are related. The triangles represent the input variables and circles represent the cluster centers. The clusters are separated based on the R^2 with next cluster component statistic. The variables which have been grouped together show obvious similarity relationships.

Figure 3 - Cluster Plot of the Proximity of the Ten Variable Clusters.

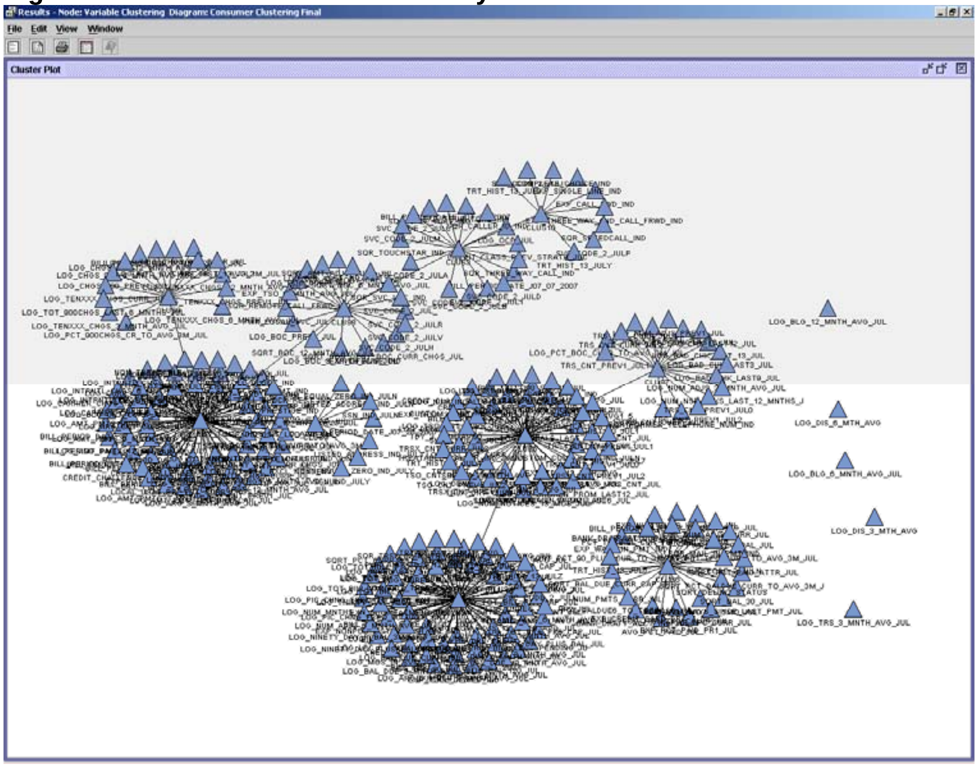
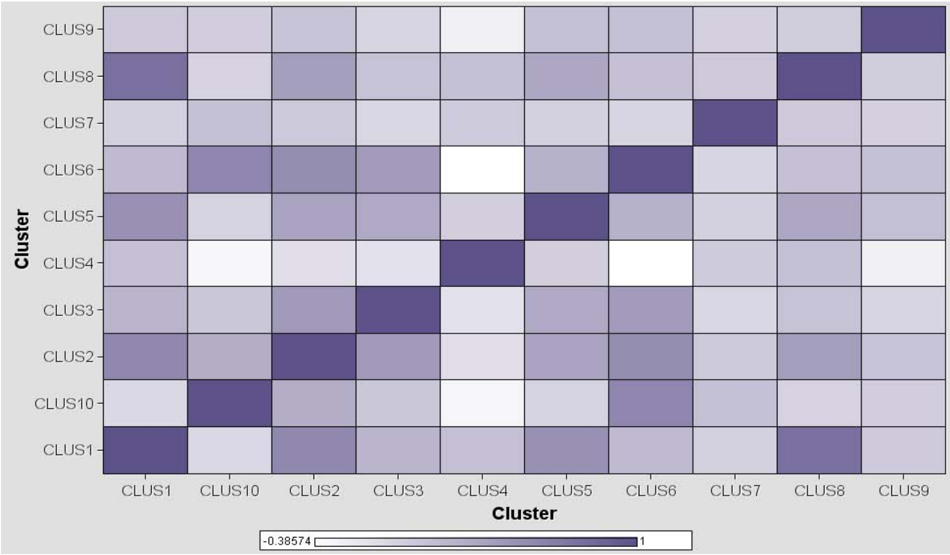


Figure 4 shows the relationship of the clusters identified in checkerboard matrix form. The shading indicates the degree of correlation between the clusters. The results below indicate there is a high degree of correlation between cluster 1 and cluster 8. This is probably due to the relationship of the basis variables that form these clusters.

Enterprise Miner provides many options for customizing the Varclus node. Instead of selecting the variables based on the lowest $1\text{-}R^2$, variables can be selected manually based on business knowledge or perhaps to produce more stable clusters.

Figure 4 - Cluster Plot of the Proximity of the Ten Variable Clusters.



Observation Clustering

The Cluster node of Enterprise Miner implements K-means clustering using the DMVQ procedure. The K-means algorithm works well for large datasets and is designed to find good clusters with only a few iterations of the data. A default Cluster node was run to quickly determine if there were any natural clusters in the data. One of the key questions the analyst must answer is, “How many questions are there?”. While there is no perfect answer to this question, there are a number of statistics that will aid in determining an optimal number of clusters. These statistics include the Cubic Clustering Criterion (CCC), the Pseudo-F statistic, Pseudo-T statistic, and the overall R^2. Graphical tools, such as, dendrograms and the CCC plot are also helpful in evaluating cluster formations.

The Cluster node selected eight clusters in the data based on the Cubic Clustering Criterion as seen in the segment size pie chart of Figure 5. Theoretically, the number of clusters is determined by the peak of the CCC. CCC values greater than 3 are considered good. It should be noted that the clusters derived are highly dependent on the basis variables used in the algorithm.

A plot of the CCC as shown in the upper left quadrant of figure 5 shows a peak at 10 clusters. Although the CCC is a useful tool for evaluating cluster results, the utility of the CCC statistic is questionable for high dimension and large scale clustering which are/is highly elongated or irregular shaped.

Figure 6 shows the cluster proximities based on Multidimensional Scaling (MDS). The MDS procedure is a dimension reduction technique which scales the large number of variables down to two dimensions while preserving the rank-order distance between the cluster means. The asterisks denote the cluster means and the circles indicate the cluster radius. Although it appears that the clusters overlap, each observation can be in only one cluster. A small radius for the clusters may indicate that there are too many clusters while a large radius may indicate too few clusters. Further analysis should be performed on these cluster solutions to determine if they are significantly different. Clusters that are not significantly different could be combined.

Figure 5 – K-means Cluster Results Obtained using a Default Cluster Node.

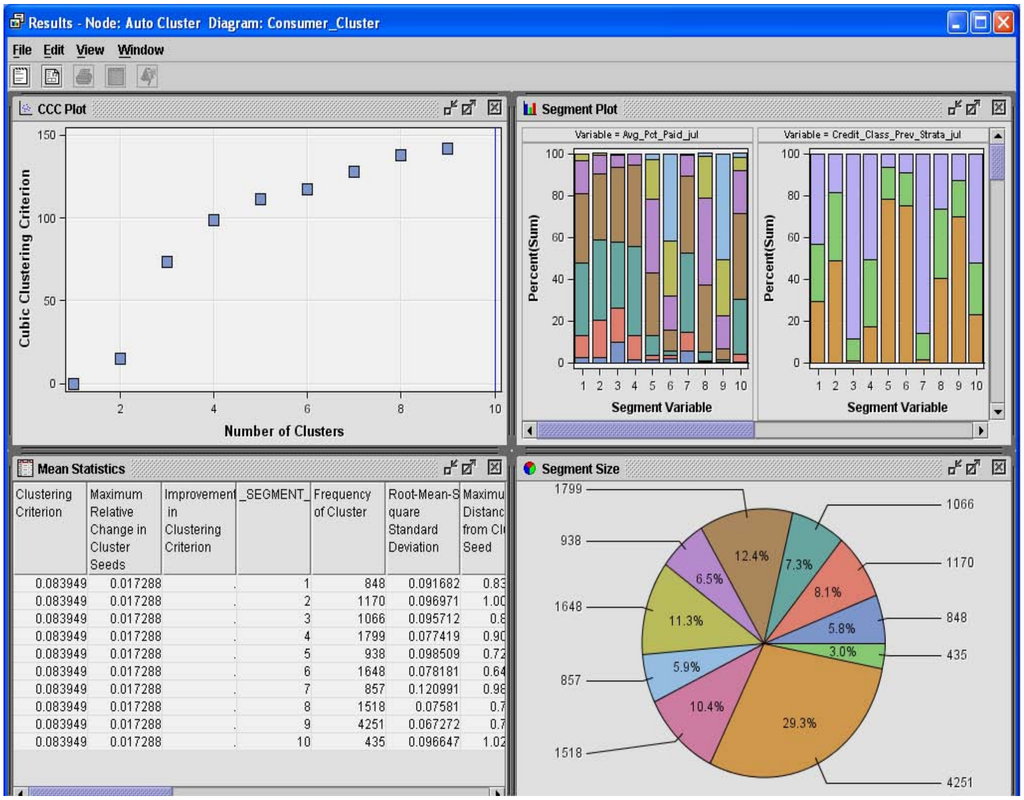
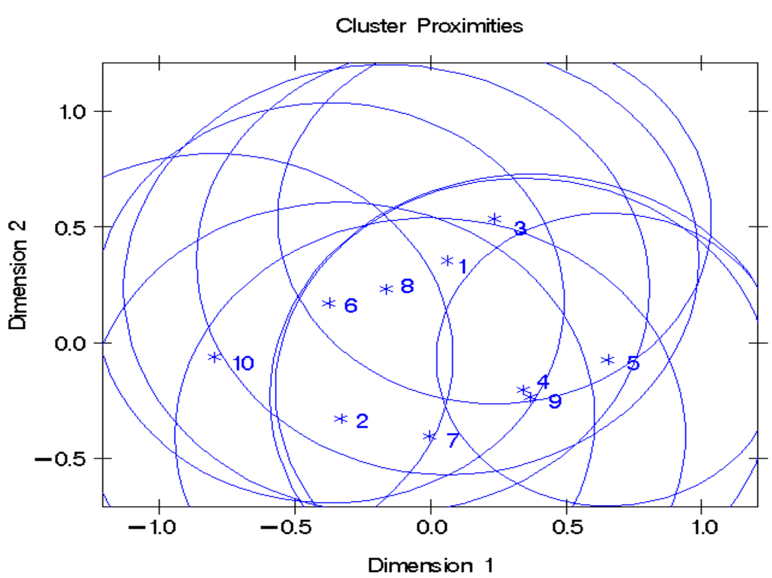


Figure 6 – Cluster Proximity Plot for Default Cluster Node.



The Segment Profile node enables you to examine clustered data and identify factors that differentiate data segments from the population. This node generates various reports that aid in exploring and comparing the distribution of these factors within the segments and population (Figure 8). Key input variables that best differentiate each cluster are ranked by variable worth. The lattice graph displays class variables as pie charts and interval variables as histograms. The blue shaded region of the histograms represents the distribution of the given-segment while the red outline represents the population. The inner ring of the pie chart displays the distribution of the population while the outer ring displays the distribution of the given segment. In the results below, the BAD rate in the first two segments is less than the bad rate in the overall population. The bad rate in third segment is much higher than the overall population. There are clear differences in the distribution of some input variables in each cluster versus the overall population. However, details about the cluster profiles are not provided due to confidentiality.

Figure 8 – Segment Profile Output for Default Cluster Node.

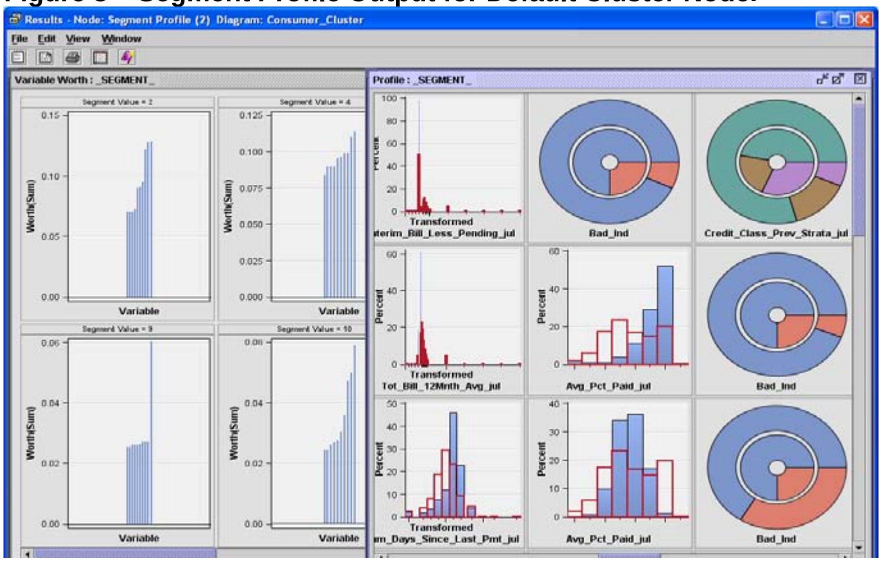


Table 2 shows the bad rate for the overall population and the individual segments. Segments with bad rates greater than the overall population are highlighted in red while segments with bad rates lower than the overall population are highlighted in green.

Table 2 – Bad Rate for Default Cluster Node.

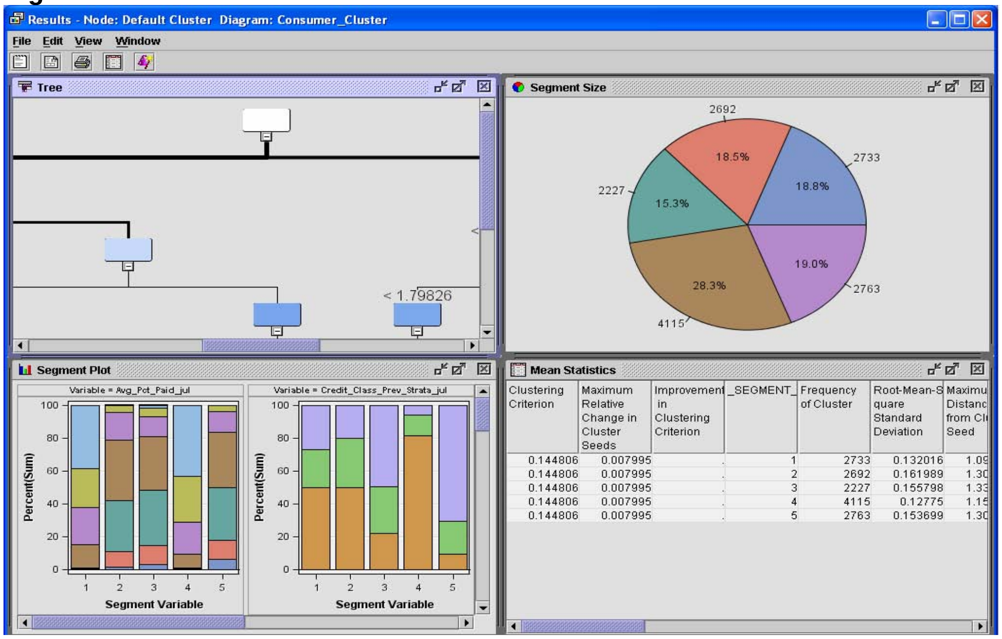
SEGMENTVALUE	Value	Count	Percent
Overall	0	10,898	75.00%
Overall	1	3,632	25.00%
1	0	1,457	76.89%
1	1	438	**23.11%**
2	0	621	69.62%
2	1	271	**30.38%**
3	0	926	81.51%
3	1	210	**18.49%**
4	0	1,047	87.76%
4	1	146	**12.24%**
5	0	1,481	96.86%
5	1	48	**3.14%**
6	0	888	49.25%
6	1	915	**50.75%**
7	0	1,022	79.84%
7	1	258	**20.16%**
8	0	813	54.13%
8	1	689	**45.87%**
9	0	2,403	95.70%
9	1	108	**4.30%**
10	0	240	30.42%
10	1	549	**69.58%**

Selection of Final Cluster Model

Once it was determined that clusters are inherent in the data, a second customized cluster analysis was done with the goal of improving on the automated cluster results. Variables selected from the Variable Clustering node were replaced where necessary, and options in the cluster node properties were tested in an attempt to optimize the results. After each run of the Cluster node the results along with the Segment Profile node were used to validate and rank the clusters solutions derived. The results of the

Cluster node are shown in Figure 9. The Cluster node results also include a decision tree profile of the clusters. In this case, a decision tree is created using the input variables to predict cluster membership; the cluster IDs represents the target variable. The decision tree used to form the clusters is shown and the most important variables used in the formation can be determined by examining the tree. The segment plot shows the distribution of each variable for each cluster.

Figure 9 - Final Cluster Results.



The Segment Profiler node was also used to determine the characteristics of each cluster with particular attention paid to the BAD rate for each cluster. Figure 10 shows a partial view of the results of the segment profile for the final model selected. Segment 4 is the best cluster with a BAD rate of 3.67 percent as seen in the upper right quadrant. The percent paid for this segment is higher than the overall sample. By contrast segment 5 is the worst cluster with a bad rate of 52.67 percent (lower right quadrant). The percent paid for this cluster is also lower than the overall sample. The other clusters contain customers with bad rates between these two extremes. Any variable in the data set can be used to profile the customers in the dataset. This includes variables that were excluded from the cluster analysis.

Figure 10 – Segment Profile Output for Final Cluster Node.

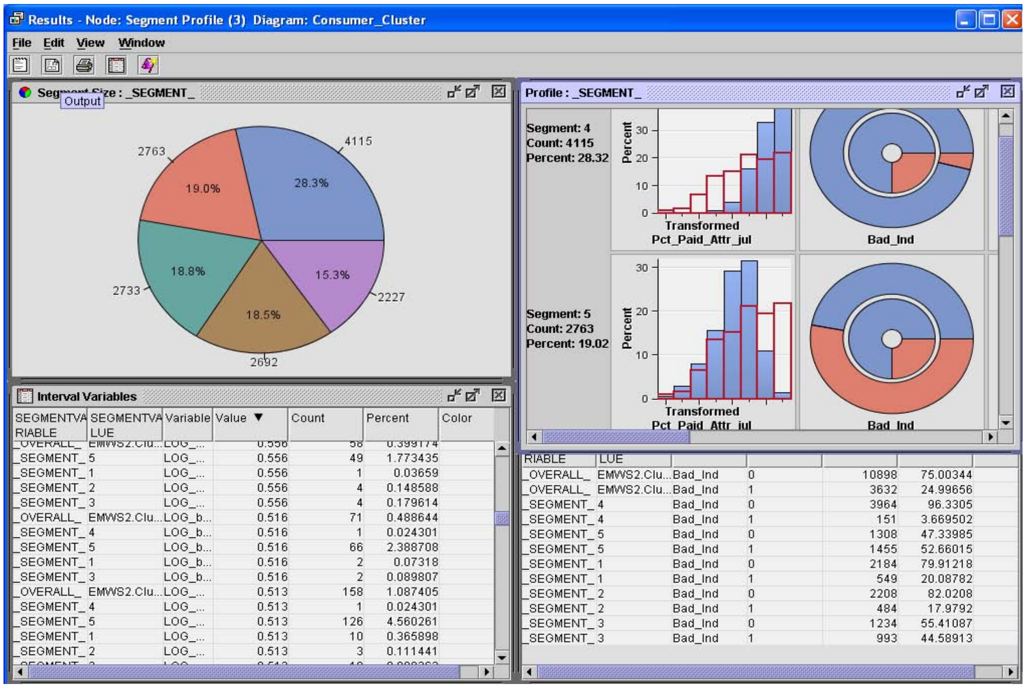
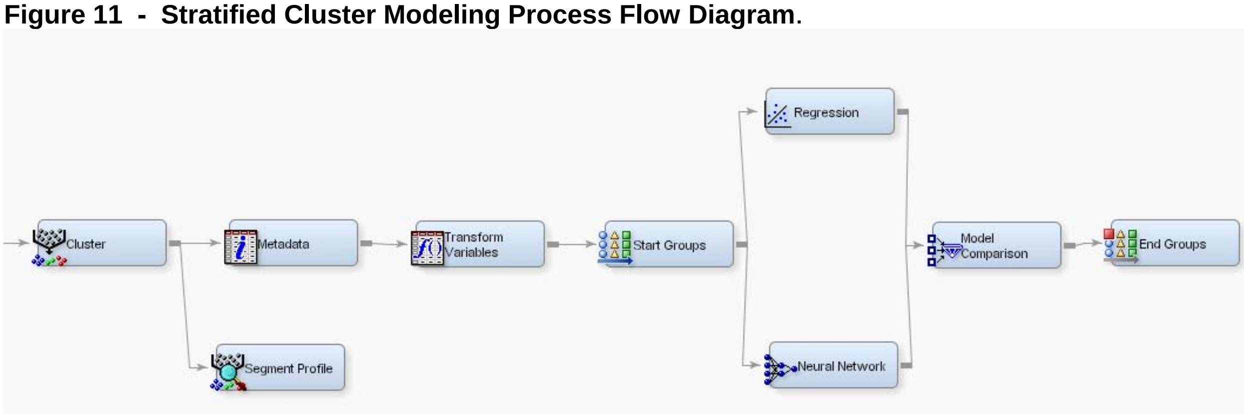


Figure 11 - Stratified Cluster Modeling Process Flow Diagram.



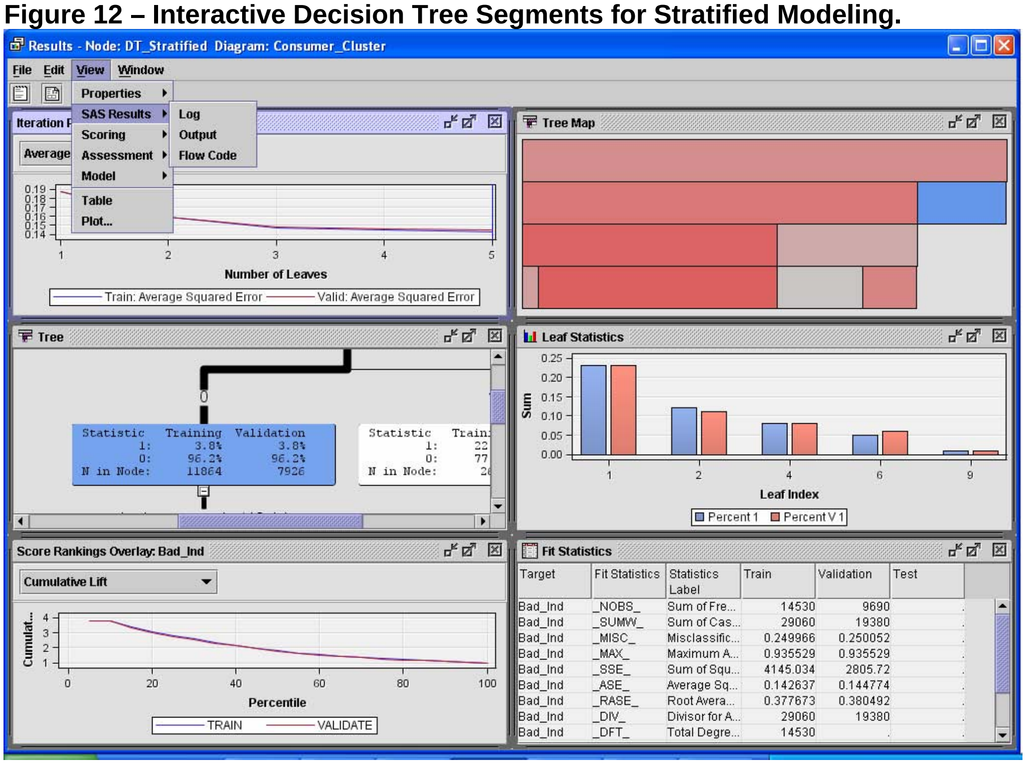
For data with multiple natural BY groups, analytically computed BY groups, or multiple target variables, a single modeling process can be repeated many times. Automating this task is a key productivity feature in Enterprise Miner. In this example, the Start Groups and End Groups nodes define a diagram segment in which a regression and neural network model are trained for each cluster. The Model Comparison node chooses the best predictive model for each cluster based on maximizing profit in the validation data. The End Group node appends the models together into a single model having independent trained functions for each cluster. Stratified modeling in each cluster can potentially tune the BAD predictions to local structures found in each cluster.

Stratified Modeling Using Decision Trees

A common strategy for potentially enhancing the predictive ability of a regression model is to first stratify the data using a decision tree and then fit a separate regression model in each leaf. This practice is often called stratified modeling with the resulting model being a hybrid model. The stratified model sometimes solves the problem of non-additivity. Typically, a small decision tree is grown interactively by the user. Unlike K-means clustering, the decision tree is a classification tool which develops segments that best differentiate the response BAD. A decision tree is often easy to interpret and evaluate both in terms of overall goodness of fit using validation data and also if the splitting rules make sense to the business user with domain knowledge of the problem. Other candidate splitting variables are also identified via surrogate or back up rules.

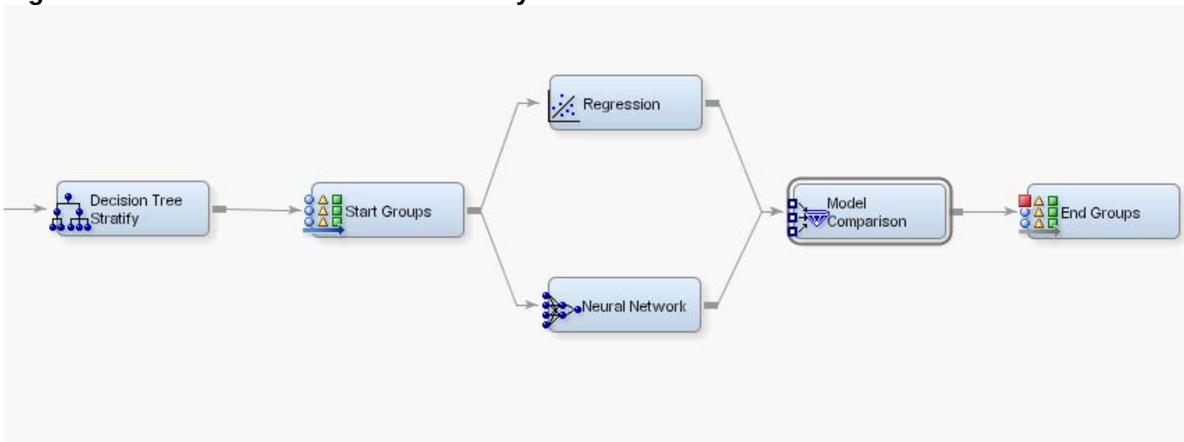
The SAS Enterprise Miner Tree Desktop Application was used to build the decision tree. Default options were used, with the exception of setting the number of rules saved in each node to 10 to ensure a larger number of candidate variables to select from. The priors were also incorporated into the split search along with profit as the model assessment criterion.

Figure 12 – Interactive Decision Tree Segments for Stratified Modeling.



This simple decision tree provides an excellent overall prediction of the target BAD, implying there is a large set of customer behavioral variable inputs (credit worthiness; number of times the customer was treated or number of times the customer broke a promise to pay) that classify well the accounts likely to go bad. The candidate variables are rank ordered based on their purity value with respect to the target variable BAD. At each split, there are several inputs that could be employed. An input with a lower Logworth value can be chosen based on business knowledge. The Segment Profile node can then be used to profile the clusters identified by this methodology. Separate predictive models were then developed in each leaf (Figure 13).

Figure 13 – Stratified Decision Tree Analysis.



Modeling

Once the clusters are developed and validated, separate predictive models can be built based on each cluster provided there are enough of each type of the performance group. If a cluster has a low proportion of BADs, the analyst has to make a decision whether a model is necessary for this cluster. If a cluster is too small to build a model, then that cluster can be combined with the next closest cluster. Another option is to hold out the small cluster and validate its sample on all the models to choose the best model. It also is a good idea to build a single model using all of the cases for comparison to the cluster solution.

CONCLUSION

This paper has provided some techniques for developing clusters to get a better understanding of your customers using SAS Enterprise Miner. It should be noted that there are other segmentation techniques such as recency, frequency, and monetary variable (RFM) that also provide customer intelligence for decision making. Since clustering is part science and part art, there is no one single correct cluster. It is critical that the clusters derived be validated to determine if the business objectives is met.

REFERENCES

Advanced Predictive Modeling Using SAS® Enterprise Miner™ 5.1 Course Notes. 2004. SAS Institute Inc. Cary, NC.

Collica, Randall S. 2007. “CRM Segmentation and Clustering Using SAS® Enterprise Miner”. Cary, NC: SAS Institute Inc.

Duling, D.; W. Thompson, S. Schubert. 2007. “Cool New Features in SAS Enterprise Miner 5.3”. *Proceedings of the SAS Global Forum 2008 Conference*. Cary, NC: SAS Institute Inc.

Lee, T.; D. Duling; S. Liu; D. Latour. 2007. “Two-stage Variable Clustering for Large Data Sets.” *Proceedings of the SAS Global Forum 2008 Conference*. Cary, NC: SAS Institute Inc.

Neville, P. 1999. "Growing Trees for Stratified Modeling," *Computing Science and Statistics, 30*.

SAS Enterprise Miner 5.3 Reference Help. 2008. SAS Institute Inc. Cary, NC.

What’s New in SAS Enterprise Miner 5.3. 2007. SAS Institute, Inc. Cary, NC. http://support.sas.com/documentation/whatsnew/91x/emgui53whatsnew.htm

Sarle, W.S., .Cubic Clustering Criterion., SAS Technical report A-108. Cary, NC: SAS Institute Inc., 1983.

SAS Enterprise Miner 5.3 Fact Sheet. 2007. SAS Institute Inc., Cary, NC. http://www.sas.com/technologies/analytics/datamining/miner/factsheet.pdf

ACKNOWLEDGMENTS

The authors thank Jim Georges of SAS Education for his invaluable insight and council of our analysis. We would also like to thank Randy Collica insights drawn from his useful text on customer segmentation. Thanks is also extended to Kevin Hodge for reviewing this paper.

CONTACT INFORMATION

Glendon Cross
AT&T Corporation
1025 Lenox Park Blvd 9B73
Atlanta, GA 30039

Wayne Thompson
SAS Institute, Inc.
SAS Campus Drive
Cary, North Carolina, 277513
Wayne.Thompson@sas.com