Paper 182-2011

Taking Disease and Health Management Analytics into the Next Generation

George S. Habek, M.S., SAS Institute Inc.

ABSTRACT

Do you struggle to visualize results of predictive modeling and segmentation within your healthcare member network? Would you like to leverage new opportunities within specific patient/member segments? A streamlined data mining approach that uses PROC GTILE to answer these questions has arrived.

This approach combines the results of a segmentation process through cluster analysis for the patient/member population with predictive model results formed through decision trees and visualized through interactive tree maps.

This document outlines a five-phase best-practice strategy for conducting disease management analytics:

1. Data preparation
2. Segmentation analysis
3. Predictive modeling
4. Linking of segmentation analysis and predictive modeling
5. Use of PROC GTILE to visualize results of Phase 4 via tree maps.

INTRODUCTION

Suppose that healthcare customer ABC is interested in enhancing its portfolio of disease management analytic capabilities. As with most health plans today, current disease management analytic efforts at ABC focus on a limited number of chronic conditions (for example, coronary artery disease, diabetes, and asthma). The cost associated with patients/members who suffer from chronic conditions makes it imperative that health plans implement disease management programs to ensure early identification and cost-effective treatment of affected individuals.

For any given employer account, however, chronic conditions affect only a small percentage of the commercial insurance population. The result is continued pressure from the employer community to interact with a much larger cross-section of the employee population. To do so requires the expansion of disease and health management activities into populations that, unlike populations with chronic conditions, have not been defined by clinical research studies. Equally important is the ability to efficiently contact eligible patients/members, enroll them in appropriate disease management programs, and monitor compliance with intervention strategies.

DATA PREPARATION

Before advanced analytics can be performed, data must be prepared in such a manner that makes it easier to analyze and more likely to generate useful results. Figure 1 color-codes the information contained in two tables received from ABC: a claims table at the claim and year levels and a membership table at the patient/member and year levels. Information from the claims table contains amounts charged, amounts actually paid, and diagnosis and disease conditions. The membership table contains information such as patient/member genders and ages.

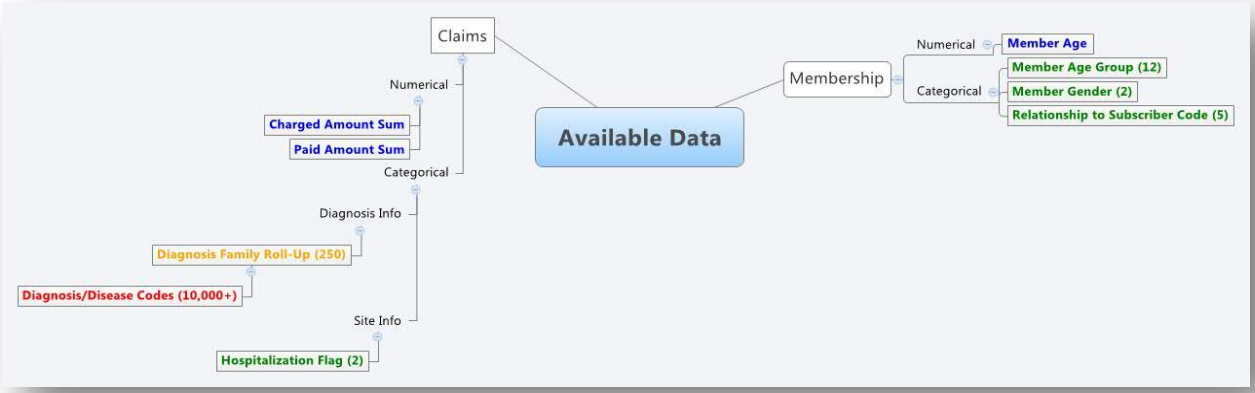

Figure {1}.

The colors indicate different types of data:

- Blue represents numerical information.
- Orange represents categorical variables with many levels, which result in a cardinality issue.
- Green represents categorical attributes deemed to be acceptable for analysis.
- Red represents categorical information with too many levels, such as diagnosis codes.

  From a predictive modeling perspective, the issue of having several hundred or thousands of levels for categorical variables does not lend itself to proper analysis and therefore must be addressed. When predictive models are being developed, a matrix of the number of observations (vertical) and number of variables (horizontal) is created. These variables can be made up of either nominal (discrete) or interval (continuous) sets of values, and it is the former that may cause an issue. When there is a nominal variable containing several hundred or thousands of discrete levels, the matrix for predictive models exponentially increases, thereby causing a computing issue for processing time and memory and resulting in a very inefficient way to build predictive models. Thus, information must be captured in nominal variables while also solving the matrix issue. The variables depicted in red are important and should not be discarded from the analysis, but the yellow variables represent at a high level the same information found in the red variables. Careful preparation of the yellow variables can yield “almost-as-good” results as the red variables would reveal. While granularity is important in the analysis, the purpose of this solution is to achieve a much broader spectrum of disease management analytics.

One preparation method to safeguard against cardinality problems is to take each variable that has many levels and create binary variables (0/1) from them, meaning that the condition is either present or it is not. This is sometimes referred to as “exploding” variables into several other ones, thus creating a more horizontal structure rather than a vertical one. From a data preparation perspective, the following steps are required to perform analytical processing:

1. Summarize each of the claims and membership tables by patient/member for each year and remove duplicate patients/members from the patient/member file for each year.
2. Due to the cardinality of the data, transpose the analysis variable values to an individual binary variable (0/1). Some analysis variables have up to 75 unique values, and the transposition process results in the creation of more than 250 new variables. A decision was made to transpose with the SAS DATA step instead of by using PROC TRANSPOSE because the SAS DATA step is more efficient for large volumes of data.
3. The claims table is rolled up to the patient/member level. Once both tables are at the patient/member level, the claims and membership tables are merged by patient/member.

Figure 2 illustrates the final data structure. Figure 2 is different from Figure 1 because it depicts the explosion of the 250 diagnosis codes into separate variables representing the 10,000+ codes depicted in Figure 1.

Figure {2}.

SEGMENTATION ANALYSIS

An important beginning to understanding the data provided involves performing a cluster analysis against a random sample (usually 10 percent) of the entire population, thus yielding smaller sub-populations possessing similar characteristics within each segment or cluster and different characteristics among clusters. These results form a book-of-business clinical segmentation, which is summarized for ABC so as to display the collection of clinical sub-populations identified by SAS® software. Every patient/member in this clinical segmentation is assigned to a sub-population, and sub-populations associated with existing ABC disease management programs are supplemented by all other sub-populations existing in the clinical segmentation. With respect to the actual clustering process within SAS® Enterprise Miner™, clustering can either be set to automatically select the “best” number of segments, or the exact number to be produced can be specified.

It is important to discuss in more detail how the clustering process is executed. The process involves applying the observation clustering node from SEMMA (sample, explore, modify, model, and assess), which is the data mining process within SAS Enterprise Miner. The main goal here is to group observations (patients/members in this case) that are similar within segments or clusters and observations that are dissimilar across segments or clusters. Simply put, everyone in a given cluster should look alike and should also be different from patients/members in other clusters. The statistical technique chosen to assign patients/members to clusters is the Ward Method. Here, the distance between two clusters is the ANOVA (analysis of variance) sum of squares between those two clusters summed over all of the variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from previous generations.

Results

There are four primary outputs that are recommended for exploration and further assessment of clustering, as follows:

1. Variable importance table
2. Segment size graph
3. Segment plot
4. Cluster profiling.

The first three are output from the cluster node in SAS Enterprise Miner, while cluster profiling is produced via SAS coding.

Variable Importance Table

The variable importance table displays, from an overall perspective, the relative significance (similar to a correlation statistic) of each of the variables in driving the clusters, measured on a scale from 0 to 1. It displays all variables with their respective importance measures (Figure 3 depicts only a portion of the complete table results). The variables and their statistics are sorted in descending order down to 0. It is safe to assume that the variables having the most impact in driving the clusters to be those that have an importance greater than 0. For example, variables such as disorders of lipid metabolism and hyperplasia of the prostate overall have fairly high significance in driving the clusters.

Variable Importance

Variable Name	Label ▲	Number of Splitting Rules	Number of Surrogate Rules	Importance
Administrative_social_admission		1	0	0.896642
Otitis_media_and_related_conditi		0	2	0.852854
Other_perinatal_conditions		0	1	0.801733
Liveborn		0	1	0.801568
Fever_of_unknown_origin		0	1	0.801513
Disorders_of_lipid_metabolism		0	6	0.73828
Other_screening_for_suspected_co		2	2	0.723012
Nonmalignant_breast_conditions		0	2	0.678949
Medical_examination_evaluation		1	1	0.67883
Hyperplasia_of_prostate		0	3	0.627194
MEMBER_GENDER		3	1	0.595883
Other_female_genital_disorders		0	1	0.567331
Menstrual_disorders		0	1	0.564605
Immunizations_and_screening_for		2	1	0.518495
Other_connective_tissue_disease		2	0	0.462305
Other_male_genital_disorders		0	2	0.448083
Gout_and_other_crystal_arthropat		0	2	0.446784
Cancer_of_colon		0	1	0.440249
Immunity_disorders		0	1	0.440195
Inflammatory_conditions_of_male		0	1	0.418394
Cancer_of_prostate		0	1	0.417891
Other_non_traumatic_joint_disord		0	2	0.402326
Other_nervous_system_disorders		0	2	0.397277
Menopausal_disorders		0	1	0.379367
Osteoporosis		0	1	0.379093
Spontaneous_abortion		0	1	0.378887
Endometriosis		0	1	0.378853
Spondylosis__intervertebral_disc		1	1	0.361687
Osteoarthritis		0	2	0.361618
Other_complications_of_birth__pu		1	1	0.358482
Normal_pregnancy_and_or_delivery		1	0	0.35748
Coma__stupor__and_brain_damage		0	1	0.354543
Other_complications_of_pregnancy		0	1	0.353499
Hemorrhage_during_pregnancy__abr		0	1	0.353219
OB_related_trauma_to_perineum_an		0	1	0.353039
Early_or_threatened_labor		0	1	0.352705
RELATIONSHIP_TO_SUBSCRIBER_CODE		2	0	0.343986
Acute_and_chronic_tonsillitis		0	1	0.264447
Asthma		0	1	0.264433
Essential_hypertension		0	4	0.212276
Sprains_and_strains		0	2	0.186219
Other_injuries_and_conditions_du		0	1	0.174478

Figure {3}.

Segment Size Graph

The segment size graph takes the form of a pie chart representing the frequency distribution of the clusters produced from the 10 percent population random sample. Six clusters were created, as shown in Figure 4. In each pie slice, the first value is the cluster number, the second value is the frequency count in that segment, and the third value is the percentage of the sample population represented in that segment.

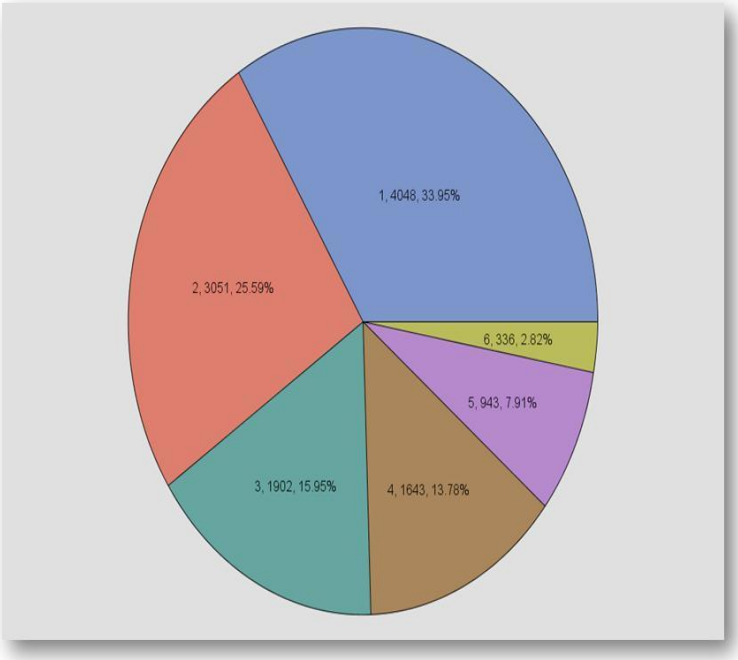

Figure {4}.

Segment Plot

The segment plot represents a stacked bar percentage frequency distribution of each of the variables and their respective dominance in forming the various clusters (because there are over 250 variables, only a 12-variable sample of the plot is displayed in Figure 5, with four on each row). The legend at the bottom indicates the content of the variables. The majority of the variables are categorical and binary in structure, where red represents variables having that event ("1") and blue represents variables not having that particular event ("0"). For example, the last variable on the bottom row depicts whether or not patients/members have menstrual disorders. Out of the six clusters, Clusters 3, 5, and 6 tend to have that condition, while Clusters 1, 2, and 4 do not. Another example to point out is the first variable on the bottom row, the age group of patients/members. Here Cluster 2 is driven by 10-year-old children or younger and Cluster 6 is driven by patients/members 31 – 35 years of age. *It is impossible to have each cluster exhibit exclusively unique characteristics; therefore, overlap is a side effect of the process.* However, certain themes that profile patients/members adequately should emerge from the segmentation process.

Figure {5}.

Cluster Profiling

The next piece of output can be deemed to be the most important, as it delves deeper into the segments in order to determine which variables contribute to the cluster formations. There are two main aspects to assess: 1) descriptive statistics (basic statistics such as means for intervals or continuous variables) and 2) cluster graphs (graphs for each cluster illustrating driving variables for each respective segment). For the descriptive statistics, means for the interval variables are observed by each segment, and there are six segments produced. Table 1 illustrates with two clusters as an example. These two clusters display broad ranges of the spectrum with respect to ages and paid amounts. Cluster 2 has an average age of about 7 or 8 years with a paid amount of slightly more than $2,200 for the year, while Cluster 6 yields an average age of about 33 years with a paid amount of almost $11,000 for the year.

Segment Id	N Obs	Variable	Mean
2	3051	member_age paid_amount_sum	7.56 2228.09
6	336	member_age paid_amount_sum	33.31 10764.86

Table {1}.

Figures 6 and 7 depict horizontal bar graphs of the driving variables for each cluster shown in Table 1. Figure 6 displays the most significant variables that drive the formation of Cluster 2, and the measurements (deviations) for these drivers are calculated. Two values are created for each cluster: 1) the average for the cluster and 2) the average for the overall population; the deviation is simply the average for a cluster minus the average for the overall population, which provides the true drivers for each of the six clusters. The “Deviation2 SUMs” to the right of the bar graph in Figure 6 measures the amount of variability being explained by that variable within that cluster. Based on the driving variables shown, a theme around health check-ups for youths emerges. Figure 7 also offers a fair amount of driving variables, and a theme around young married females with some issues around pregnancy is apparent.

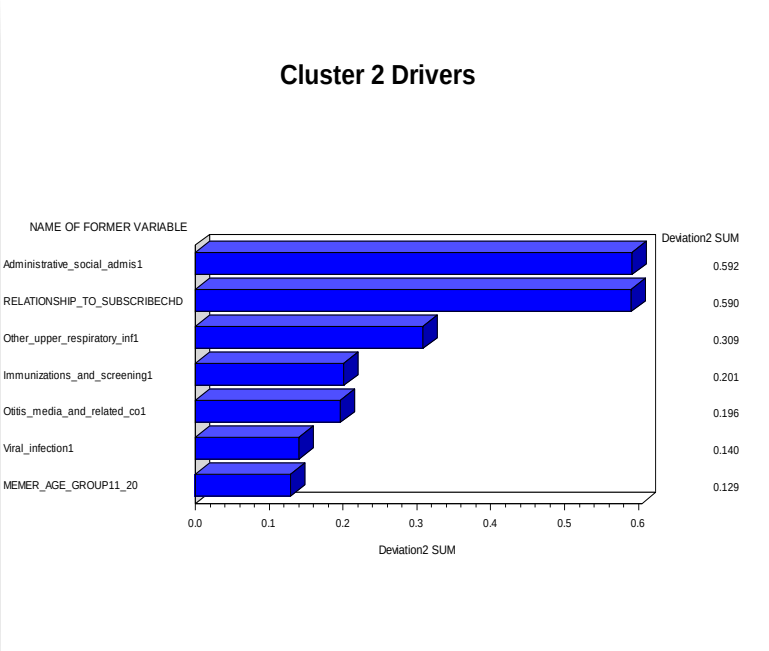

Figure {6}.

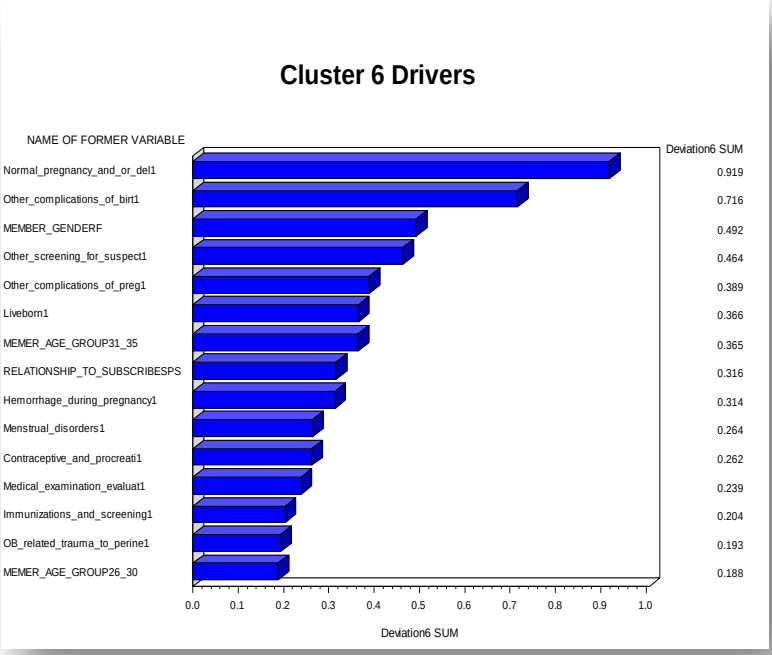

Figure {7}.

The next common task is to assign the entire population of interest to one of the six clusters developed in this example. Figure 8 displays a portion of the SAS Enterprise Miner flow with the initial SAS data set for the segmentation process followed by the cluster node using the "specify" property to produce six clusters. The score node is where the same SAS data set is connected in order to assign the entire sample into one of the six clusters. *The SAS data set shown towards the top of Figure 8 must be of the "score" role type, thus informing SAS Enterprise Miner to score all patients/members into one of the produced clusters.* The cluster assignment process does *not* work otherwise.

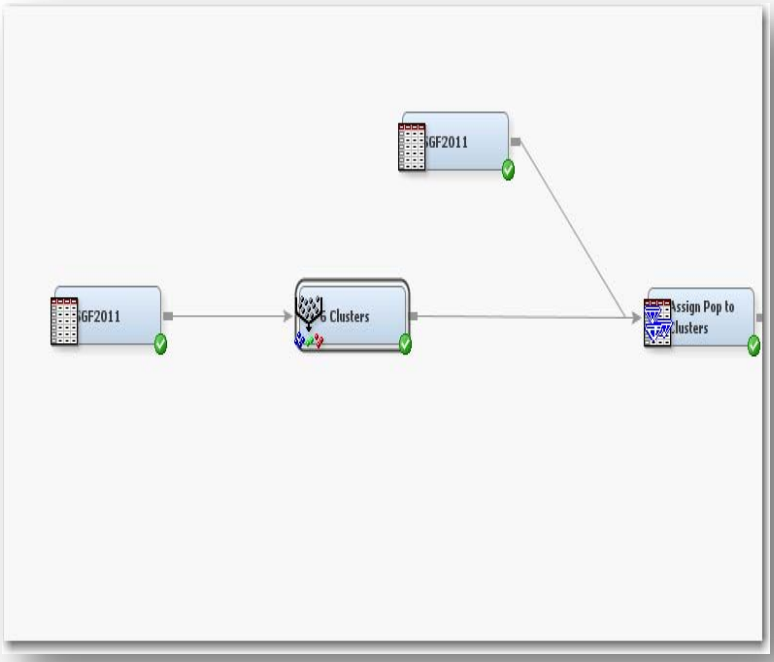

Figure {8}.

PREDICTIVE MODELING

Another important phase in the analysis incorporates more proactive (rather than reactive) thinking. While this stratification/segmentation/cluster analysis process (unsupervised learning) is occurring, there may be the need to execute some predictive modeling/analytics (supervised learning) to that process in parallel. Two models are deemed ready for execution, as follows:

- *Model I* – establishes drivers for patients/members being hospitalized, thus developing the likelihood of hospitalization
- *Model II* – establishes drivers for patients/members having thyroid disorders, thus developing the likelihood of being diagnosed with a thyroid disorder.

It is usually a good idea to sample the input SAS data set unless there are less than approximately 100,000 observations. One might also want to perform some visual exploration and descriptive statistics, as follows:

- Visual exploration executed through the "MultiPlot" node offers two important results:
  - Assess the distribution of the target (0/1) against various inputs (*X*s)
  - Determine normality of the inputs to see if transformations may be required

- Descriptive statistics executed through the “StatExplore” node offer three important results:
    - Produce the most correlated inputs graphically against the target prior to modeling
    - Display basic statistics for all variables
    - Determine whether or not any variables contain missing values and therefore whether or not imputation may be warranted.

The next portion of the flow deals with partitioning the data (or sample). Best practices suggest designating 70 percent of the data for model development (training) and 30 percent for model validation. Comparing the various techniques to determine which model emerges as the champion or winner is also a best practice. Two common models are utilized: 1) decision trees and 2) stepwise logistic regression.

Model I Results

The champion algorithm is the decision tree, although stepwise logistic regression is a very close second. In fact, both algorithms are outstanding. Significant drivers for predicting the likelihood of a patient/member being hospitalized include attributes such as paid amounts in excess of $8,000 for the year, pregnancy complications, and having one’s appendix removed. When the results of a decision tree are very large, the technique turns into an excellent segmentation and exploration mechanism. For the regression output, the key piece is the summary of final drivers selected from the algorithm. The drivers for predicting the likelihood of a patient/member being hospitalized are as follows:

- Paid amount (medical risk)
- Whether or not a patient/member had a pregnancy where the baby was stillborn (and whether a patient/member had a normal pregnancy or delivery, in general)
- Whether or not a patient/member had appendicitis.

The overall misclassification rate for both techniques is approximately 3 percent, meaning that Model I accurately predicts who would likely be hospitalized approximately 97 percent of the time.

Model II Results

Both algorithms are very close once again, but stepwise logistic regression is deemed to be the champion over the decision tree. Significant drivers for predicting the likelihood of a patient/member being diagnosed with a thyroid disorder include attributes such as whether or not a patient/member had a medical examination and/or evaluation, having malaise or fatigue, having nutritional deficiencies, paid amounts in excess of $1,400 for the year, and whether or not a patient/member is female. For the regression output, the top five drivers for predicting the likelihood of a patient/member being diagnosed with a thyroid disorder are as follows:

1. Medical examination and/or evaluation for a patient/member
2. Whether or not a patient/member had malaise or fatigue
3. Whether or not a patient/member had nutritional deficiencies
4. Paid amount (medical risk)
5. Patient/member gender.

The overall misclassification rate for both techniques is approximately 3 percent, meaning that Model II accurately predicts who would likely be diagnosed with a thyroid disorder approximately 97 percent of the time.

Cost Analysis

In addition to misclassification as an accuracy measure for Models I and II, there are costs associated with both false positives and false negatives, defined as follows:

- False positives:
    - Model I predicting that a patient/member would likely be hospitalized when in fact they are not
    - Model II predicting that a patient/member would likely be diagnosed with a thyroid disorder when in fact they are not

- False negatives
    - Model I predicting that a patient/member would not likely be hospitalized when in fact they are
    - Model II predicting that a patient/member would not likely be diagnosed with a thyroid disorder when in fact they are.

The median was chosen instead of the mean, since the median is not as sensitive to outliers within the data. In a cost analysis of Model I, the median paid amount sum (cost) for a patient/member being hospitalized is $14,078. Model I yielded a total of 201 false positives (an associated total cost of $2,829,678) and 175 false negatives (an associated total cost of $2,463,650). Although false negatives may be more of a concern because health initiative efforts must be increased in those cases, false positives and false negatives are both misclassifications from the model.

In a cost analysis of Model II, the median paid amount (cost) for a patient/member being diagnosed with a thyroid disorder is $2,129. Model II yielded a total of 13 false positives (an associated total cost of $27,677) and 438 false negatives (an associated total cost of $932,502). This is a high cost despite the fact that Models I and II both have very good accuracy, so it might be useful to be rather strict when deeming what a "good" misclassification rate is (such as less than 1 percent). *Both techniques for Models I and II show essentially the same drivers.* The main value of decision trees is that the various splits for each driver and how they segment the population can be seen.

Scoring

Now that solid predictive models have been established for the business questions, a common next step is to take the winning algorithms and score 2010 data for a 2011 prediction for each model. *The decision tree is a discrete algorithm, so the probability scores are grouped. The regression algorithm is continuous, so the probability scores are more linear.* Figure 9 illustrates the model scores from the hospitalization model and shows that the model is very useful not only in assessing the distribution of likelihoods but more importantly to decide on an appropriate cut-off for assigning which patient/members exhibit the likelihood of a given event (for example, being hospitalized or being diagnosed with a thyroid disorder). Thus, any patient/member possessing a score greater than the cut-off score is deemed to be a "1"; otherwise, they are tagged as a "0". In this example, the likelihood of a patient/member being hospitalized is calculated. It would be desirable for a majority of patients/members to be at the low end of the score spectrum, as that would mean lower risk of being hospitalized and as a result would imply lower medical risk. In this case, almost 90 percent of the population has approximately a 3 percent risk of being hospitalized.

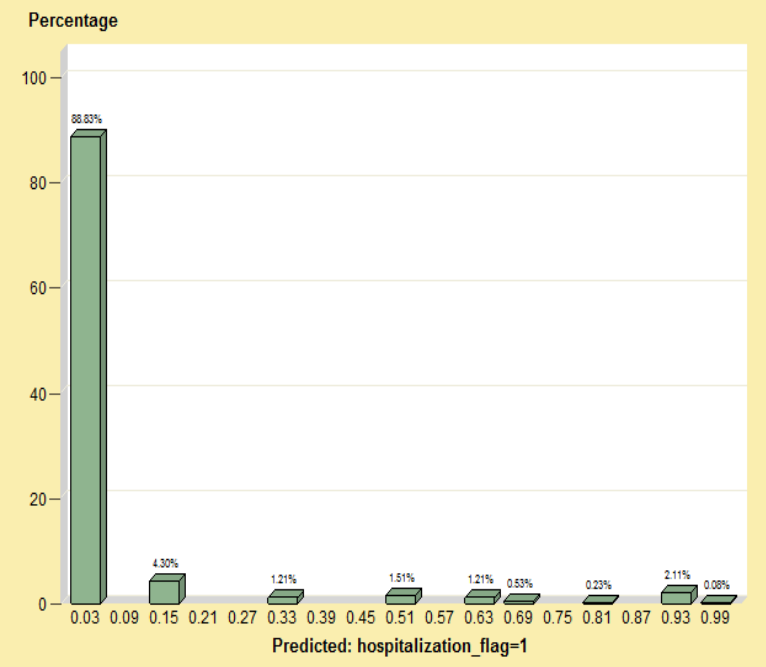

Figure {9}.

LINKING OF SEGMENTATION ANALYSIS AND PREDICTIVE MODELING

Thus far, exploratory analysis using a stratification/segmentation/cluster analysis process (unsupervised learning) has been discussed. Predictive modeling/analytics (supervised learning) is used as a specific target to build models. It is usually a good idea to try and link the two types of analytical techniques together. The two models developed previously are independent of the stratification/segmentation/cluster analysis process. In addition to the main objective previously set forth, another goal is to showcase some different applications for stratification/segmentation/cluster analysis and predictive modeling. From a business perspective, there may be several ideas or questions that one may want answered in conducting disease management analytics. In order to create a link between the two concepts of stratification/segmentation/cluster analysis and predictive modeling, the six clusters already developed are used to associate a driver (in this case, second-year cost [paid amount]) with the target and find the predicted value of that target. This establishes the drivers for medical risk of the segmented population. The SAS Enterprise Miner flow utilized incorporates the following steps:

1. Establishes six clusters from a 10 percent random sample that may or may not have been developed inside SAS Enterprise Miner
2. Assigns the population into one of the six clusters
3. Filters each cluster to produce a predictive model for second-year cost (paid amount)
4. Sets up the necessary data to produce visualization (tree maps) of the results.

TREE MAPS

Tree maps are a very helpful technique in visualizing predicted results for clusters, and do so by using nested rectangles as shown in Figure 10.

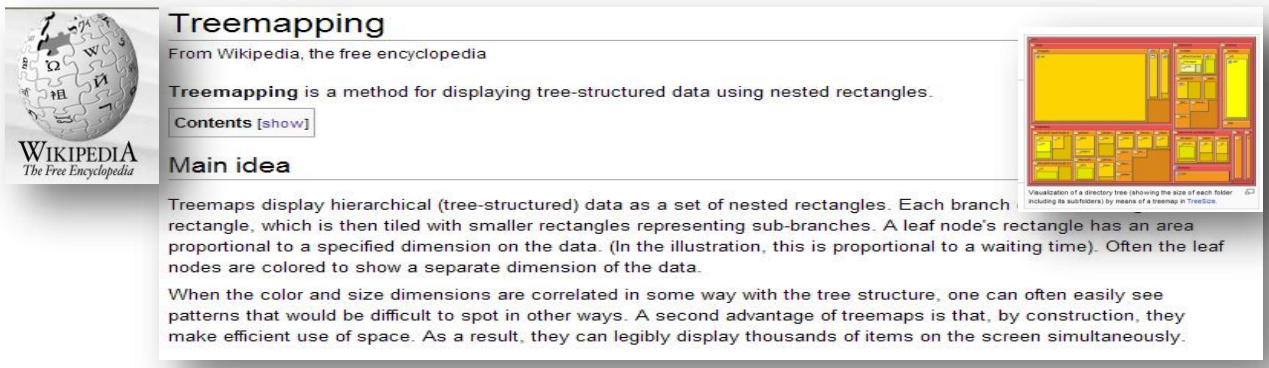

Figure {10}.

What Do You Need?

Data should be set up in a hierarchical manner. For example, tree maps can have two layers, with Layer 1 being the creation of segments from the population of interest and Layer 2 being a more specific sub-group within a segment. In this discussion, Layer 1 represents the six clusters from the patient/member population sample, while Layer 2 represents the top drivers for predicting second-year cost for one of those clusters.

Let's Talk PROC!

PROC GTILE is shown in two parts below, the first part being some administrative code, as follows:

```
ODS LISTING CLOSE;
ODS HTML FILE = 'C:\SGF2011\Tree Maps\clusterseg.html' GPATH = 'C:\'
GOPTIONS RESET = ALL DEVICE = JAVA HSIZE = 8.42 VSIZE = 5.31;
```

The following is a summary of the meaning of each line of code:

- The first line closes any open output delivery system that may exist
- The second line indicates where the HTML file containing the tree map is to be located
- The third line dictates sizing for the tree map.

The heart of the code is as follows:

```
PROC GTILE DATA = TREEMAPIN;
 TILE Member_Count TILEBY = (Population_Cluster, Top_Clinical_Driver
 / COLORVAR = Predicted_Medical_Risk COLORRAMP = (green orange red);
RUN;
QUIT;
ODS HTML CLOSE;
ODS LISTING;
```

The following is a summary of the meaning of each line of code:

- The first line indicates the input SAS data set to be used for the tree map
- The second line specifies the size dimensions of the tiles in the tree map; layers 1 and 2 of the tree map are also indicated
- The third line indicates the variable to be used for the color gradient across the tree map, along with the color scale
- The next-to-last line closes the Java™ map
- The last line specifies the status of the output delivery system.

Figure 11 illustrates a sample of the input SAS data set, as follows:

- The first column represents Layer 1 of the tree map (in this case, the population cluster that emerged from the segmentation process from SAS Enterprise Miner)
- The second column represents Layer 2 of the tree map (in this case, the top clinical drivers for the specific sub-segments within the clusters)
- The third column represents the response from the predictive modeling process within SAS Enterprise Miner using decision trees; it also becomes the color gradient across the tree map for the predicted medical risk of the second-year cost (paid amount) (low risk = green, medium risk = orange, and high risk = red)
- The fourth column represents the count of patients/members within the sub-segments.

	Population_Cluster	Top_Clinical_Driver	Predicted: paid_amount_sum	Member_Count
1	Back-to-School Youths	Abdominal_pain Otitis_media_and_related_conditi	3401.7451724	29
2	Back-to-School Youths	Acute_and_chronic_tonsillitis Otitis_media_and_related_conditi	4534.66625	32
3	Back-to-School Youths	Joint_disorders_and_dislocations MEMBER_AGE	2585.1891304	23
4	Back-to-School Youths	MEMBER_AGE Acute_and_chronic_tonsillitis	2517.856	30
5	Back-to-School Youths	MEMBER_AGE Fracture_of_upper_limb Otitis_media_and_related_conditi	6192.0366667	12
6	Back-to-School Youths	Medical_examination_evaluation Otitis_media_and_related_conditi	3817.7322222	9
7	Back-to-School Youths	Other_aftercare	3636.8996	25
8	Back-to-School Youths	Other_gastrointestinal_disorders	1994.3222222	36
9	Back-to-School Youths	Other_nervous_system_disorders	2536.6757895	38
10	Back-to-School Youths	Otitis_media_and_related_conditi	1655.3469748	357
11	Female Routine Utilizers	Disorders_of_lipid_metabolism	2094.172973	37
12	Female Routine Utilizers	Disorders_of_lipid_metabolism Nonmalignant_breast_conditions	7762.9966667	6
13	Female Routine Utilizers	Headache__including_migraine	3749.7111765	17
14	Female Routine Utilizers	MEMBER_AGE	1361.1447059	442
15	Female Routine Utilizers	Nonspecific_chest_pain	3618.1985185	27
16	Female Routine Utilizers	Nutritional_deficiencies Disorders_of_lipid_metabolism	8054.3033333	6
17	Female Routine Utilizers	Spondylosis__intervertebral_disc	2488.1781579	38
18	High Risk Pregnancy	Hemorrhage_during_pregnancy__abr Other_complications_of_pregnancy	3957.4416667	6
19	High Risk Pregnancy	Other_complications_of_pregnancy	3598.0828571	7
20	Male Accidental Youths	Joint_disorders_and_dislocations	2295.14	19
21	Male Accidental Youths	Other_male_genital_disorders	3046.4546154	13
22	Male Accidental Youths	Other_non_traumatic_joint_disord	1772.7396774	62
23	Male Accidental Youths	Other_non_traumatic_joint_disord Joint_disorders_and_dislocations	4553.4786364	22
24	Male Accidental Youths	Other_upper_respiratory_disease	1418.7476471	51
25	Male Accidental Youths	Spondylosis__intervertebral_disc	1712.4979518	83
26	Middle-Aged Female Acutes	Menstrual_disorders	9620.8778571	14
27	Middle-Aged Female Acutes	Osteoarthritis	8025.95	7
28	Middle-Aged Female Acutes	Other_nervous_system_disorders	16268.052	10
29	Middle-Aged Female Acutes	Viral_infection	8627.916	10

Figure {11}.

Figure 12 illustrates the predictive model results using the clinical drivers of the clusters for second-year cost. The size of the grids represents the number of patients/members and the color gradient represents the predicted medical risk for those patients/members. This is a drillable map where a second level exists when one of the clusters is selected. Figure 13 shows the selection of middle-aged male acutes drilled down to the next level. The top drivers in predicting medical risk include heart valve disorders, osteoarthritis, and other upper respiratory diseases.

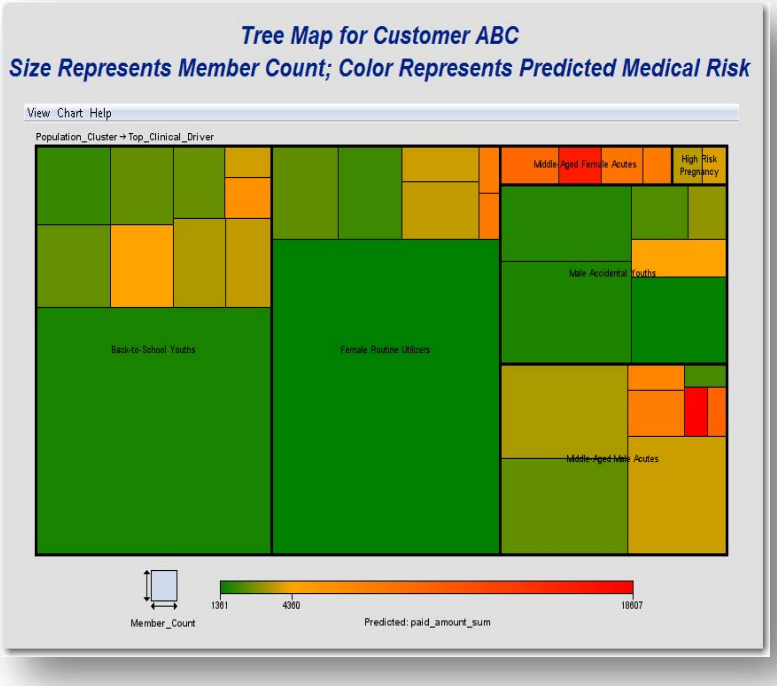

Figure {12}.

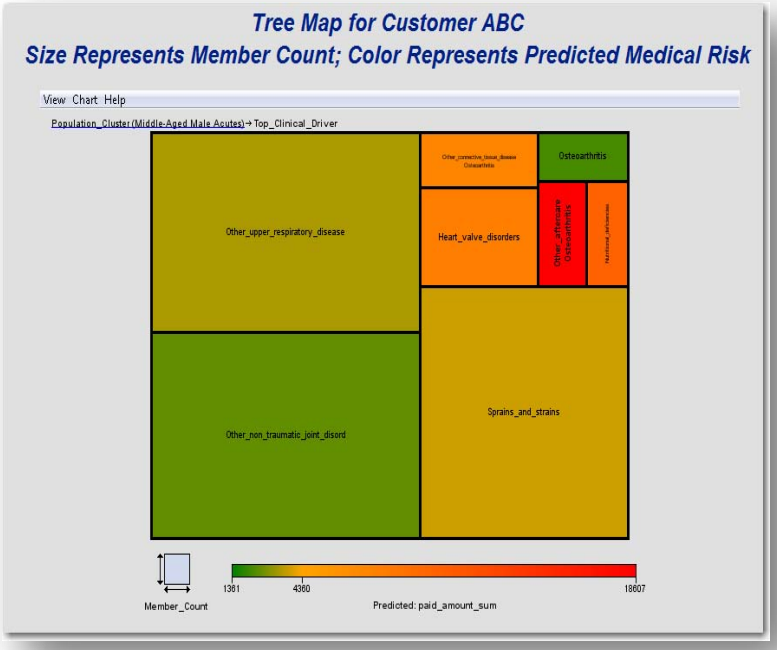

Figure {13}.

CONCLUSION

ABC and SAS have a unique opportunity to advance the state of disease management analytics. Through a development partnership that offers access to the power of SAS software's predictive modeling and automated campaign management capabilities, ABC can demonstrate the value of such tools for the identification of new disease management analytic opportunities. At the same time, the partnership offers SAS software insight into health plan data so as to validate disease management analytics currently under development. By working together, SAS can enable ABC to emerge at the forefront of health plan disease management analytic methodologies and demonstrate greatly improved return on investment for disease management programs to ABC's employees. Some key take-aways from the disease management analytics presented are as follows:

- Data preparation activities are vital for the successful application of disease management analytics
- Stratification/segmentation/cluster analysis process (unsupervised learning) allows for an entire patient/member population within a healthcare network to be executed
    - Profiling a patient/member population is vital to better understanding behaviors with the goal of establishing more cost-effective treatment plans
- Predictive modeling/analytics (supervised learning) allows for the association of several attributes in order to establish drivers for predicting certain events or quantitative values
    - Enables proactive rather than reactive thinking
- Joining the two types of learning greatly enhances the expansion of disease management analytics
- Tree maps paint a solid picture of the stratification/segmentation/cluster analysis and predictive modeling results.

ACKNOWLEDGMENTS

My thanks for the excellent feedback and suggestions from the following:

- Anne Baxter, Technical Editor, SAS Institute Inc.
- Dave Caira, Software Developer, SAS Institute Inc.
- Doug Grossman, Technical Editor, SAS Institute Inc.
- David Ogden, Principal Analytic Consultant, SAS Institute Inc.
- Chris Scheib, Health Business Solutions Manager, SAS Institute Inc.
- John Shipway, Solutions Architect, SAS Institute Inc.

CONTACT INFORMATION

Comments and questions are valued and encouraged. Contact the author at George.Habek@SAS.com.