

Segmentation of Data Prior to Modeling

Gregg Weldon, Merkle, Inc., Atlanta, GA
Zidun He, Merkle, Inc., Atlanta, GA

ABSTRACT

In the early years of statistical modeling, a “one model fits all” for clients was commonplace. Thanks to increasing efficiencies in data storage and greater computing power, it is now economically feasible to build, code, and manage multiple models for a single client. This allows for a much greater level of accuracy in prediction and portfolio management. The question becomes how many models are optimal, and which models will give a client the best business results. This paper reviews a SAS® program designed for detailed segmentation analysis. The results are easily transferred into a final report for review with the client.

BACKGROUND OF SEGMENTATION

Segmentation is merely the act of deciding which models a client needs for conducting their business. For example, a small credit union may make direct loans (for autos, boats, vacations, etc.) to its members. It may also issue a VISA card. A single risk model for this client would work relatively well. However, because of different delinquency rates, approval requirements, and application variables, we may find that two models would work better. A model for direct loans may have very different variables in it than a model for just VISA customers. By building two models, we're able to capture specific attributes for that portfolio that an overall model would miss. This way, the credit union can set different score cut-offs, policies, and procedures for approval.

Another example that is very common is when a client engages in both direct mail and telemarketing in order to attract new customers. A single response model would work, but separate models may be better.

HOW MANY MODELS?

In the above examples, it's fairly easy to arrive at the optimal number of models for a client. The credit union client will want two models, one for direct loans and one for VISA applicants. The marketing client will want a telemarketing model and a direct mail model. Problem solved!

Many times, the client isn't sure how many models they will need, or what these models are going to be. For example, assume the credit union makes loans to industry-specific members (airline employees) as well as to the general public. Should the two models be direct loans vs. VISA, as before, or should they be industry-specific vs. everyone else?

Many marketing firms will need different models based on specific offers rather than by marketing channel (direct mail vs. telemarketing). These specific offers may be based on products, demographics, geographic regions, and varying incomes.

EXAMPLE OF THE SEGMENTATION PROBLEM

Let's take a typical client to illustrate the problem with segmentation. We assume the financing arm of an auto manufacturer needs a series of risk models to better control loan delinquencies in its portfolio. After meeting with the client, a list of potential model splits is listed below:

New cars vs. Used cars
Individual risk models for all 5 (or more) auto models
Models based on low risk, medium risk, and high risk (using a generic risk score)
Models based on young vs. mature buyers
Models based on existing customers vs. new customers
Models based on loan vs. lease

As you can see, the number of models can get long very quickly. Which group is best?

Segmentation is a combination of art and science. The art comes from sitting with the client and deciding on the best business solution. The science is where SAS® comes into play. By utilizing SAS®, the Analyst can arrive at all the numbers and statistics necessary to make an educated recommendation to the client. Clients need and want feedback, and the program we'll discuss here will help this occur.

THE SEGMENTATION PROGRAM

The process of segmentation starts by researching the client's needs. Our client in this case is a financial company whose major concern is being able to distinguish the higher risk group of applicants from the rest of the database. With that in mind, the Analyst proceeds to the next step of gathering needed data.

Two types of variables are used for this analysis – Segment variables and Analysis variables. **Segment variables** are variables that are candidates for splitting into separate models, and often come from statistical analysis and client feedback. The credit union that believes direct loans vs. VISA card customers would want to look at "Type of Loan". The marketing client may want to examine "Type of Contact" (direct mail vs. telemarketing). In many segmentation analyses of credit data, "Age of Oldest Trade" is a nice potential split variable.

Analysis variables are predictive variables that are used in segmentation for means testing. These are variables that the Analyst knows, from prior experience, are good possibilities for coming into any potential models that are eventually developed. Analysis variables provide a profile of the characteristics of the observations in the client's data. Means testing in this way is also called profile analysis.

For example, risk models, in general, look at five overall types of variables: inquiries, age of trades, types of trades, balances, and delinquencies. In segmentation analysis for a risk model, then, we may take 10-12 credit variables, preferably at least one from each of the above groups. In this way, we get an accurate profile of the client's overall data. For marketing models, we would do the same. In these cases, we would want a variety of variables that would give us a snapshot of the people on whom our models will be built. Examples could include education, income, time at job, household makeup, as well as a variety of aggregated variables at the zip+4 level.

In our project example, the result of decision tree analysis showed that Risk Score is a good candidate for segmentation with a split at 570. Risk Score measures risk from a 0 to 1000 scale, where high scores mean high credit quality (low risk of default) and low scores mean low credit quality (high risk of default). The client also suggests that it makes business sense to segment the universe with Income at \$4000. Figure1 shows the details of the Segment variables selected.

Var	Splits
Risk_Score	0-569 570 or Up
INCOME	0 - 4000 4001 or Up

Figure1.

At this point, the Analyst is ready to utilize SAS® for further analysis. In the SAS® program, two macro variables are created to store the Analysis and Segment variables. To keep the dataset to a manageable size, only the variables needed are kept. For this analysis, a performance variable called BG (BAD = 0, GOOD = 1) is created. Since data is never 100% populated, it is very important to implement logic to handle

missing values. The Analyst could delete the missing variables, populate the missing variables with value 0, or artificially give the missing variables some other value for analysis.

One of the major SAS® tasks is to calculate the Bad Rates and the mean values of the analysis variables within each segment. Since the same set of statements is run repeatedly here, it is easier to package all the pieces in a macro like the code shown below.

```
%MACRO SEGMENT(VAR, LOW, HIGH);  
ods html body = "C:\client\&var&low&high..xls";  
  PROC MEANS DATA=SEGMENT1 mean;  
    VAR %SEGLST;  
    class bg;  
    WHERE (&LOW <= &VAR <= &HIGH );  
    TITLE "&VAR : &LOW - &HIGH BADS";  
  RUN;  
  
  PROC FREQ DATA=SEGMENT1;  
    TABLE BAD/MISSING;  
    WHERE (&LOW <= &VAR <= &HIGH );  
    TITLE "&VAR : &LOW - &HIGH";  
  RUN;  
ods html close;  
%MEND SEGMENT;
```

The macro variable %SEGLST in the VAR statement of the means procedure contains all the Analysis Variables for which SAS® will return mean values. The SAS® code creates this macro variable is as follow:

```
%MACRO SEGLST;  
Time_current_job  
Avg_acct_age  
Num_acct_pastdue  
Num_acct_60dpd  
%MEND SEGLST;
```

The Class statement produces separate output for each stage of the performance variable – BG. The three parameters, &var, &low, and &high, in the Where statement efficiently define the subset of data for the procedure. The macro is invoked by this statement:

```
%SEGMENT( INCOME, 0, 4000 )
```

The macro variable &var is replaced with value INCOME, &low with 0, and &high with 4000. For every possible split the data might have, there is a macro invoke statement to pass over the parameter values. Through the ODS statement in SAS®, all the procedure outputs are written to Excel reports in a desired folder.

Figure2 below summarizes all the analyses done so far. For each of the candidate splits and the total universe, the report shows their Bad Rate distribution and average mean values for all the analysis variables.

Segmentation Analysis

Variable		Total	INCOME		Risk_Score	
			0-4000	4001+	0-569	570+
Pop. %		100.00%	51.70%	48.30%	42.40%	56.90%
Bad Rate		12.81%	14.31%	11.15%	16.32%	10.11%
num_acct_pastdue	Bad	1.48	1.5	1.45	2.09	0.77
	Good	1.15	1.18	1.12	1.85	0.73
num_acct_60dpd	Bad	2.03	2.04	2.02	2.84	1.08
	Good	1.61	1.64	1.57	2.6	1
time_current_job	Bad	61.53	55.97	69.19	62.14	59.85
	Good	65.58	59.52	71.79	66	65.44
avg_acct_age	Bad	35.07	32.78	38.23	36.27	33.69
	Good	40.96	38.05	43.94	40.47	41.29

Figure2.

THE ANALYTICAL PROCESS

Step1

The first thing to do is compare the Bad Rate distributions of the segments. The most impressive finding here is the Bad Rate for the lower Risk Score group (low credit quality). It is 27% higher than the Bad Rate of the entire population (16.32% vs. 12.81%), and 61% higher than the higher Risk Score segment (16.32% vs. 10.11%) The higher Risk Score group (high credit quality) has a 10.11% Bad Rate, which is 21% lower than the Bad Rate of the total data.

On the other hand, the lower Income group has Bad Rate 11% higher and the high Income group has Bad Rate 13% lower than total population. As the report shows, splits on Income are not as impressive.

Step2

Next, the Analyst will go through several rounds of analysis to examine every analysis variable in each of the segments (see Figure 3).

The first analysis variable is the number of accounts currently past due (NUM_ACCT_PASTDUE). The Bad population in Total (entire universe) has 1.48 average past due accounts. In the Income segments, Seg1 (lower income) has 1.5 average past due accounts, and Seg2 (higher income) has 1.45. These three numbers are not much different from each other. For the Good population, the Total column has 1.15 average past due. Seg1 of Income segments has 1.18 and Seg2 has 1.12. This again is very flat for both of the segments.

Segmentation Analysis

Variable		Total	INCOME		Risk_Score	
			seg1	seg2	seg1	seg2
Pop. %		100.00%	51.70%	48.30%	42.40%	56.90%
Bad Rate		12.81%	14.31%	11.15%	16.32%	10.11%
num_acct_pastdue	Bad	1.48	1.5	1.45	2.09	0.77
	Good	1.15	1.18	1.12	1.85	0.73

Figure 3

The Risk Score segments, however, show a better separation. For the Bad population, Risk Score Seg1 has 41% more past due accounts than total (2.09 vs 1.48) and Seg2 is 47% less than total (0.77 vs. 1.48). Furthermore, Seg1 has almost 3 times more past due accounts than Seg2 (2.09 vs. 0.77). The same trend is found in the Good population for the Risk Score segments. Compared with the Total, Seg1 has 60% more past due accounts (1.85 vs. 1.15) and Seg2 is 36% less (0.73 vs 1.15). Seg1 is also found with past due accounts 2.5 times more than Seg2.

The second analysis variable is for number of accounts 60 days past due (NUM_ACCT_60DPD) (see Figure 4). The variable should show a trend that is similar to the first variable since both of them carry delinquency information. The fact that the Total Bad Population has averaged 2.03 accounts 60 days past due, Seg1 has 2.04 accounts, and Seg2 has 2.02 accounts shows that there is little difference for these segments. The analysis for the Good population shows same result since the Total has 1.61 account 60 days past due, Seg1 is 1.64 and Seg2 is 1.57.

Variable		Total	INCOME		Risk_Score	
			seg1	seg2	seg1	seg2
Pop. %		100.00%	51.70%	48.30%	42.40%	56.90%
Bad Rate		12.81%	14.31%	11.15%	16.32%	10.11%
num_acct_60dpd	Bad	2.03	2.04	2.02	2.84	1.08
	Good	1.61	1.64	1.57	2.6	1

Figure 4

As expected, the Risk Score segments have much better trends. Seg1 is much riskier than the entire universe because it has 40% (in Bad population) and 60% (in Good Population) more accounts 60 days past due than the Total population. As the data shows, Seg2 in Risk Score segment is 46% (in Bad population) and 37% (in Good population) lower than the universe.

The next variable in the list is for duration of time at current job (time_current_job) (see Figure 5). The Income segments show a strong trend that indicates longer employment awards higher compensation, which makes sense. The lower income group for Seg1 here has much shorter job duration for both the Good and the Bad universes.

Segmentation Analysis

Variable		Total	INCOME		Risk_Score	
			seg1	seg2	seg1	seg2
Pop. %		100.00%	51.70%	48.30%	42.40%	56.90%
Bad Rate		12.81%	14.31%	11.15%	16.32%	10.11%
time_current_job	Bad	61.53	55.97	69.19	62.14	59.85
	Good	65.58	59.52	71.79	66	65.44

Figure 5

According to time_current_job, Risk Score is not a very good candidate split. The average mean values of time_current_job in the Risk Score segments are almost the same for the Total, the Good, and the Bad. As the data clearly points out, the difference between the Bad population and the Total is only around 1% for Seg1, and 2% for Seg2. For the Good population, the difference is even less than 1% for both segments.

The last variable is average age of trades (see Figure 6). The Analyst will again examine the average mean value for the variable in great detail for each of the segments to determine which segments show the best trends. This time, the data suggests that Income is a better indicator. The average age of trades for Seg1 is about 15% shorter than the Seg2 for both the Good and Bad universes.

Segmentation Analysis

Variable		Total	INCOME		Risk_Score	
			seg1	seg2	seg1	seg2
Pop. %		100.00%	51.70%	48.30%	42.40%	56.90%
Bad Rate		12.81%	14.31%	11.15%	16.32%	10.11%
avg_acct_age	Bad	35.07	32.78	38.23	36.27	33.69
	Good	40.96	38.05	43.94	40.47	41.29

Figure 6

Compared with the Income segments, the Risk Score segments are not showing impressive separation power. The trend for the Good population is especially weak with lesser than 1% of variance between Total and Seg1, Total and Seg2, or Seg1 and Seg2.

Step3

The records in each segment will be somewhat homogenous after segmentation. The level of homogeneity is examined to make sure that it's not so high as to cause problems later in the modeling stage. When all the records in a modeling universe are acting alike, it is difficult to build a model with high performance. In the third step, the Analyst implements some quick checks to see how the trends of the analysis variable changed after the split by examining the column differences (Good vs. Bad) for each analysis variable, rather than row differences (Good vs. Good and Bad vs. Bad).

In this case study, there are significant differences between Good and Bad for the Income splits (num_acct_pastdue and num_acct_60dpd). However, time_current_job and avg_acct_age don't look very different for Income. This is the opposite from what we saw in Step 2, when we examined row differences. That's because by splitting on Income, we've now created two fairly homogenous groups, where time at job and account age are very similar for each observation within the splits. One thing to be careful of is splitting the data into groups that negate your best potential model variables.

For the Risk Score splits, we see that the risk variables (pastdue and 60dpd) are somewhat flat, although not so much for Seg1 as for Seg2.

Also note that in this case study, only 4 analysis variables are being shown. In reality, 10 to 12 variables are the norm. Since this is a risk project, most of the analysis variables are risk-related.

FINAL RESULTS

Segmentation is a unique aspect in modeling in that it blends art and science in almost equal measures. It's entirely possible to build a segmentation routine that would calculate the absolute best statistical split (Income or Risk Score, in this case). Doing so would be wrong. The numbers presented in the segmentation analysis chart are very important, but they only serve as a GUIDE for making a decision on modeling splits. Segmentation is ONLY effective when these numbers are coupled with business needs, common sense, and client buy-in.

Models built on Income would work, and the client would probably be happy with them. The same can be said for models built on the Risk Score. In this situation, there is no wrong answer. (Many times, however, you'll find splits that are so pathetic they recommending them WOULD be the wrong answer!). Which split do you recommend to the client? It depends on the client's wants and needs for the project. In this example, we recommended the Risk Score split, as that variable had more of a direct relationship on how the client ran his business. It was a split that made sense to the client and was a split that he could explain to his superiors and his lenders.

Segmentation is a very interactive assignment, as frequent discussion with the client is paramount to the success of a project. The program used in this paper serves as a fast, efficient way to churn through a large quantity of data in order to give the client the information needed for a final decision. It is hoped that this paper will assist the reader in making this phase of a project more user-friendly and manageable.

CONTACT INFORMATION:

Gregg Weldon
Merkle, Inc.
Senior Director of Analytical Services
7000 Central Parkway, Suite 330
Atlanta, GA 30328
(770)804-1088
gweldon@merklenet.com

Zidun He
Merkle, Inc.
Data Analyst II
7000 Central Parkway, Suite 330
Atlanta, GA 30328
(770)551-1884
zhe@merklenet.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand or product names are trademarks of their respective companies.