MWSUG 2016 - Paper AA12

An Innovative Method of Customer Clustering

Brian Borchers, Ph.D., Direct Options Inc. West Chester, OH

ABSTRACT

This paper describes an innovative way to identify groupings of customer offerings using SAS® software. The authors investigated the customer enrollments in nine different programs offered by a large energy utility. These programs included levelized billing plans, electronic payment options, renewable energy, energy efficiency programs, a home protection plan, and a home energy report for managing usage. Of the 640,788 residential customers, 374,441 had been solicited for a program and had adequate data for analysis. Nearly half of these eligible customers (49.8%) enrolled in some type of program. To examine the commonality among programs based on characteristics of customers who enroll, cluster analysis procedures and correlation matrices are often used. However, the value of these procedures was greatly limited by the binary nature of enrollments (enroll or no enroll), as well as the fact that some programs are mutually exclusive (limiting cross-enrollments for correlation measures). To overcome these limitations, the PROC LOGISTIC procedure was used to generate predicted scores for each customer for a given program. Using the same predictor variables, PROC LOGISTIC was used on each program to generate predictive scores for all customers. This provided a broad range of scores for each program, under the assumption that customers who are likely to join similar programs would have similar predicted scores for these programs. The PROC FASTCLUS procedure was used to build *k*-means cluster models based on these predicted logistic scores. Two distinct clusters were identified from the nine programs. These clusters not only aligned with the hypothesized model, but were generally supported by correlations (using PROC CORR) among program predicted scores as well as program enrollments.

INTRODUCTION

Many energy utilities offer a range of products, services, and energy-saving methods for their customers. These include levelized billing programs, payment convenience options, incentives for shifting electric usage to off-peak hours, whole home surge protection, service plans for home wiring and water heaters, home energy usage reports, home weatherization services, and various others. Utilities offer these programs for a number of reasons. Providing convenient payment options can enhance customer satisfaction and possibly improve on-time payments. Protection plans can give customers peace-of-mind while providing revenue for the utility. Energy efficiency programs can help customers save money and energy while enabling utilities to meet energy reduction goals set by regulators.

While offering valuable programs and services, many utilities have limited understanding of their customers' behavior, needs, and preferences. It is often useful to create demographic and behavioral profiles of existing profile participants for marketing purposes. This can allow the utility to engage new customers who do not currently participate in any programs beyond basic energy delivery. Another way the utility can increase their customer engagement is to help their current program customers enroll in additional programs and services, thereby increasing the enrollments per customer. In order to achieve this, utilities can benefit from information on which programs attract similar customers. Essentially this calls for a customer segmentation where products and services are grouped based on similar customers, and then align the customers their most appropriate programs. This enables the utility to cross-sell its programs with greater success and reduced cost.

SAS/STAT software provides a number of options for cluster analysis, which can help identify groupings based on similarities. This paper will demonstrate a method of using the FASTCLUS procedure of cluster analysis based on each customer's probability of enrolling in each program or service. Predicted probabilities are computed with logistic regression using the LOGISTIC procedure.

HYPOTHESIZED MODEL

As a basis for understanding how customers align to utility programs, Direct Options has developed a theoretical model based on over 20 years of observation in the industry. According to this hypothesized model, utilities offer programs which fall into two general categories and attract two different types of customers (see Figure 1). The first category is called Products and Services, which provide convenient options for billing and payment, as well as protection plans to cover certain home repairs. The second category is called Energy Efficiency, which includes programs that help customers reduce their household electric usage or provide energy from renewable sources (e.g., wind power). According to the hypothesis, one group of customers enrolls in Products and Services while another group enrolls in Energy Efficiency, while the remaining customers are 'Disengaged' and do not enroll in any programs.

Products and Services Customers	**Disengaged Customers**	**Energy Efficiency Customers**
Enrolled in programs which offer billing options, convenience, or protection	Not enrolled in any utility programs	Enrolled in programs which enable the customer to reduce energy consumption or use renewable energy sources

Figure 1. Hypothesized Model of Utility Programs and Associated Customers

CUSTOMERS AND ENROLLMENTS

This analysis uses data from 640,788 residential customers of a mid-sized investor-owned electric utility. Customer variables include monthly electric usage (kWh) from the previous year, PRIZM demographic segments (Nielsen Company), and enrollments in any of nine programs and services offered by the utility. To qualify for analysis, customers must have 12 months of usage data, a PRIZM segment code, and must have been solicited for at least one program or service, resulting in 374,441 customers in the analysis group. For each customer, general categories of age and income are estimated using PRIZM Social Groups and Lifestage Groups.

Due to the nature of the data, traditional methods of grouping programs and services by customer similarities have limited success. In the data set, each program is represented in a column, and each customer is scored 1 for each program enrolled and 0 for each program not enrolled. One option is to use Spearman (i.e., nonparametric) correlation to generate a correlation matrix of dual-enrollments between each program. This strategy is not used because only a small portion of customers (16%) had enrolled in more than one program. Also, some programs are mutually exclusive where customers can enroll in only one or the other, making correlation methods impractical.

Cluster analysis is another commonly used solution, where variables like utility programs are grouped based on meaningful customer behaviors such as enrollments. SAS/STAT provides a non-hierarchical procedure called FASTCLUS which is well-suited to large data sets such this. The FASTCLUS procedure is commonly programmed as such:

```
PROC FASTCLUS DATA=Dataset MAXCLUS=2 OUT=Output;
  VAR Variable1 Variable2 Variable3 Variable4 Variable5;
RUN;
```

In this example code, PROC FASTCLUS is run with the variables to be clustered listed on the VAR line. Specify the number of clusters in the MAXCLUS option. A two-cluster model is specified in this example. Unfortunately this method is also limited with our data because clustering procedures require variability of data within variables, and we have only binary (1/0) data to indicate enrollments. Another option that is often used is to begin with a principal components factor analysis using the PRINCOMP procedure and

then using the factor loadings to conduct a cluster analysis (Gibler & Tyvimaa, 2014). This method is also not practical with our analysis because PROC PRINCOMP is based on a correlation matrix and also requires variability within the data (Hair et al, 1992; SAS Institute, 2003). Although the above methods are not practical for our analysis, we can use these as points of validation.

THE LOGISTIC SOLUTION

Given the limitations of the traditional methods, we are testing a method of program grouping that is not commonly found in the literature. Rather than using actual enrollments, we generate predictive scores that indicate a customer's likelihood of enrolling in each program. The assumption is that by using a common set of predictors, programs with comparable customers will have more similar predictive scores between customers. Similarly, we assume programs with different customers will produce dissimilar predictive scores between customers.

The utility in our analysis has nine different programs offered to its customers. Each customer also has data on ten predictor variables, ranging from monthly household electric usage to demographics such as age and income. For each individual program we run a logistic regression using the LOGISTIC procedure to generate predictive scores, using each of the ten predictor variables and the binary enrollment value (1/0) as the criterion:

```
PROC LOGISTIC DATA=Dataset;
  MODEL Program1 (Event='1')= Predictor1 Predictor2 Predictor3
    Predictor4 Predictor5 Predictor6 Predictor7 Predictor8 Predictor9;
  OUTPUT OUT=WORK.Predict1 PREDPROBS=INDIVIDUAL;
RUN;
```

The procedure calculates predicted scores for Program1 into a new variable called IP_1 and saves this into the new data set WORK.Predict1. The procedure is then repeated Program2 and for each additional program until predicted scores are generated for all programs:

```
PROC LOGISTIC DATA=WORK.Predict1;
  MODEL Program2 (Event='1')= Predictor1 Predictor2 Predictor3
    Predictor4 Predictor5 Predictor6 Predictor7 Predictor8 Predictor9;
  OUTPUT OUT=WORK.Predict2 PREDPROBS=INDIVIDUAL;
RUN;
```

Running logistic models for all 9 programs produces predictive scores for each customer on each program using the same set of predictor variables. The ability to generate clusters from predictive scores depends on the predictive quality of each of the logistic models. As a guideline, a c-value of at least 0.6 is the standard for assuring that the predicted scores would be comparable across products. Logistic models achieved a c-value of 0.6 or greater for all but one of the programs, with only Program 6 falling slightly below this mark. The c-values for each of the logistic models are listed in Table 1.

Utility Program	Logistic model c-value
Program 1	0.628
Program 2	0.625
Program 3	0.724
Program 4	0.659
Program 5	0.679
Program 6	0.592
Program 7	0.652
Program 8	0.652
Program 9	0.696

Table 2. LOGISTIC Procedures: Predictive Quality of Logistic Models for Each Program

CLUSTER MODEL OF UTILITY PROGRAMS

Once predictive scores have been generated for each customer on each program, a cluster analysis can then be performed on the data. SAS offers a number of options that can be selected with PROC FASTCLUS. Different methods of calculating distance can be selected with the LEAST= option, where Euclidian Distance (LEAST=2) is the default method. Then specify the number of clusters in the model with the MAXCLUSTER= option. We decided to use the city-block distance method (LEAST=1) and build a model with two clusters:

```
PROC FASTCLUS DATA=Work.Predict9
  LEAST=1 MAXCLUSTERS=2;
VAR IP_1 IP_12 IP_13 IP_14 IP_15 IP_16 IP_17 IP_18 IP_19;
RUN;
```

This process is then repeated with 3-, 4-, and 5-cluster models by changing the MAXCLUSTERS option in the program. Only the 2-cluster model achieved convergence, while the other models failed to reach convergence after 20 iterations. The 2-cluster model produced clusters of programs that were more clear and distinct than those produced by the other models. Based on its successful convergence and parsimony, the 2-cluster model was selected as the final cluster model. The clusters and their associated programs are presented in Table 2:

Cluster A	Cluster B
Program 1	Program 2
Program 7	Program 3
Program 8	Program 4
	Program 5
	Program 6
	Program 9

Table 2. Initial Two-Cluster Model: Programs Grouped in Each Cluster

Cluster A consists of three billing, convenience, and protection programs that closely align with the Products and Services group in the hypothesized model. Cluster B consists of a number of energy conservation and renewable energy programs that align with the Energy Efficiency group in the hypothesized model. The only exception is Program 6, which is a billing convenience program that was expected to fall within Cluster A. Although it scored highly on Cluster A, Program 6 scored highest on Cluster B and therefore was assigned to the latter cluster in the initial model. There are two explanations why this may have occurred. The first may be due to simple marketing habits. Many utilities solicit customers in Program 6 for a wide range of other programs, and therefore may attract similar customers into these other programs. The second reason may be due to a statistical limitation of this method, as Program 6 generated the weakest prediction model and may have been miscategorized because the customer prediction data may not have a strong connection to this program.

As hypothesized, Program 2 is grouped in Cluster B with Energy Efficiency programs. However, this program scored on both clusters at nearly equivalent levels. This is interesting because this program, although commonly used for reduction of energy usage, also serves as a customer education program which would be associated with Cluster A. In the final cluster model, Program 2 is included in both clusters, as it serves as a bridge between Products and Services and Energy Efficiency. Program 6 is also moved to Cluster A because it is deemed to be a better fit in that group. The final cluster model is presented in Table 3:

Cluster A	Cluster B
Program 1	Program 2
Program 2	Program 3
Program 6	Program 4
Program 7	Program 5
Program 8	Program 9

Table 3. Final Two-Cluster Model: Programs Grouped in Each Cluster

VALIDATION OF MODEL

The final cluster model is quite similar to the hypothesized model and provides support for this theoretical grouping of programs based on similar customers. Although the theoretical model was used to guide the selection and interpretation of clusters, its similarity to the cluster model can serve as a validation of the method used to generate the model.

The current model and methodology can also be validated by correlations between programs. As noted previously, it is difficult to obtain correlations of actual program enrollments due data limitations and the fact that some programs are mutually exclusive to each other. Instead, programs are correlated based on predicted scores, similar to the clustering process. The correlation results are shown in Table 4. Using the final two-cluster model, programs within Cluster A (shaded in blue) have an average correlation of 0.58. Programs within Cluster B (shaded in green) have an average correlation of 0.68. The average correlation of programs between clusters fell lower at 0.34. Program 2 is included in both clusters as it appears in the final model. Although this analysis relies on predicted scores rather than direct customer data, the wide difference between within-cluster correlations and between-cluster correlations appears to support the current model.

	1	2	3	4	5	6	7	8
Program 1								
Program 2	0.25							
Program 3	-0.58	0.34						
Program 4	0.22	0.86	0.31					
Program 5	-0.07	0.83	0.71	0.74				
Program 6	0.12	0.77	0.41	0.49	0.75			
Program 7	0.53	0.78	-0.12	0.56	0.41	0.70		
Program 8	0.56	0.85	0.00	0.87	0.63	0.54	0.73	
Program 9	0.15	0.82	0.46	0.85	0.89	0.61	0.47	0.77

Table 4. Correlation Matrix of Predicted Scores for Each Program. All correlations statistically significant at .01 level.

CONCLUSION

The current project tested an alternative method to generating clusters of programs based on similar customers. Rather than clustering based on customer characteristics or purchases, the present analysis utilized customer probabilities to enroll in each program. These probabilities (or predicted scores) are generated using logistic regression (PROC LOGISTIC) with a common set of customer variables as predictors. These predicted scores for each program are then used in a nonhierarchical cluster analysis (PROC FASTCLUS). Based on the results and validation efforts, it appears that this method does work as a viable option in grouping programs by similar customers.

Cluster analysis in general is a useful procedure for marketers and other professionals who manage products or programs. Its most common use is in customer segmentation, where different groups with common behaviors or interests are identified within the customer base. This enables marketers to customize marketing messages and channels to best fit the interests and habits of the segment. When a company offers a large set of products or programs, cluster analysis can help identify groups of programs that share commonalities (such as common customers). This enables marketers to cross-sell more effectively to current customers and identify underserved populations that could be reached through new products or programs.

The clustering method used in this analysis is best suited for certain situations. When customers are being segmented and rich data is available, traditional methods of cluster analysis are preferred. This is because customer data links directly to customer behaviors and characteristics (e.g., number of visits, total dollars spent, age, income). However, when programs or products are being grouped by similar customers and only binary data (e.g., yes/no) are available, generating predictive scores for each program can provide a richer data set to conduct a cluster analysis.

This probability-based cluster analysis has the same limitations as any other clustering process. There is often no single correct answer, but a process of using available information to build a cluster model that fits the needs and goals of the organization. It helps to start with a theoretical model to give a framework for interpreting various cluster models that are generated through PROC FASTCLUS. The current method, however, does have a unique limitation in that it relies on indirect data. Because the predictive score data are generated through logistic regression, the usefulness of the data depends on the predictive quality of the logistic models. Any weakness in the logistic model extends into weaknesses in the clustering process. When strong logistic models can be built to predict a customer's involvement in each program, these prediction scores do have value as a foundation to conduct a cluster analysis.

REFERENCES

Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. 1992. *Multivariate Data Analysis*. 3rd ed. New York: Macmillan.

Gibler, K.M. & Tyvimaa, T. 2014. "The Potential for Customer Segmentation in the Finnish Housing Market." *The Journal of Consumer Affairs,* 48:351–379.

Nielsen Company. "Nielsen Prizm." Accessed February 19, 2015. Available at http://www.claritas.com/MyBestSegments/Default.jsp?ID=70&pageName=Learn%2BMore&menuOption=learnmore

SAS Institute. 2003. *Applied Clustering Techniques: Instructor-Based Training.* Cary, NC: SAS Institute.

ACKNOWLEDGMENTS

The original theoretical model was developed by Jan S. Moore, President/CEO of Direct Options, Inc. We thank our utility partners for the use of their data and we thank Direct Options for its support of this project.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Brian Borchers, PhD
Direct Options
513-779-4416
bborchers@directoptions.com
www.directoptions.com