Paper 1285-2014

Customer Profiling for Marketing Strategies in a Healthcare Environment

MaryAnne DePesquo, Phoenix, Arizona

ABSTRACT

In this new era of healthcare reform, health insurance companies have heightened their efforts to pinpoint who their customers are, what their characteristics are, what they look like today, and how this impacts business in today's and tomorrow's healthcare environment. The passing of the Healthcare Reform policies led insurance companies to focus and prioritize their projects on understanding who the members in their current population were. The goal was to provide an integrated single view of the customer that could be used for retention, increased market share, balancing population risk, improving customer relations, and providing programs to meet the members' needs. By understanding the customer, a marketing strategy could be built for each customer segment classification, as predefined by specific attributes. This paper describes how SAS® was used to perform the analytics that were used to characterize their insured population. The high-level discussion of the project includes regression modeling, customer segmentation, variable selection, and propensity scoring using claims, enrollment, and third-party psychographic data.

INTRODUCTION

The enactment of the Affordable Care Act (ACA) impacted the health insurance industry in most facets of its business model. In the previous environment, the insurer knew and determined who their member population were based on actuarial practices and product development. Starting in October 2014 the members would look very different. Now the insurer must look at other ways to increase profitability and enhance customer service not using the customary actuarial model. As ACA continues, insurers will need to maintain a balance of high and low risk members within their population and develop a portfolio of products to maintain business objectives.

This paper will focus on the approach developed to profile the pre-ACA population and apply the outcomes to marketing campaigns in the years leading up to the 2014 reform implementation. The following discussion will outline the process and methodologies applied in the process with limited information on statistical formulas or specific business results.

PROJECT OVERVIEW

The first questions asked in strategic planning meetings were 'Who are our customers?', 'What are their characteristics?', 'How will this impact business today and tomorrow?'. To answer these questions it was agreed that the primary objective of the research project was to gain insight to our current customers and identify their characteristics for the upcoming marketing campaign strategies. Of those characteristics, which members were the most desirable for our business model?

The discussion in this paper will describe gathering requirements, identifying data sources and conducting data transformations, selecting variables statistical methodology, implementation, monitoring the action plan and summarizing the outcomes.

PLANNING AND REQUIREMENTS GATHERING

Requirements Gathering

A most important factor when beginning to plan a project is to involve the stakeholders. Collaboration with the Business owner is key to conducting an analysis that will be used by those owners as it provides a benefit to the specific department as well as to the company. A valuable exercise in planning is to research similar projects in the industry to learn what worked in similar business models. The focus of the team discussions was the purpose of the customer profiling: current marketing campaigns, future campaigns, and qualified product development. Overall, the agreed upon Business ROI is interpreted in terms of a financial outcome or market share, as examples.

With a Business owner involvement, requirements are gathered to understand and agree upon the scope of the project. Examples of requirements:

- Project scope
- Timeline
- Target population - line of business, market share, timeframe for member enrollment and specific products
- Data sources- enrollment, medical claims, pharmacy claims, revenue, survey
- Allocation of analyst resources
- Business owner time commitment
- Actionable outcomes

The documentation and review of these items will determine the feasibility of the project, most importantly the availability of data and resources.

Collecting and Preparing Data

The next task was to identify the available data, locate the data, and assess the aging of the data. The preparation and cleaning of the data usually takes more time than the time to build and run models.

Data: In-house (medical claims, enrollment revenue), third-party survey is selected by using company business knowledge and articles from similar analytics.

Sample size: The data collected was a large sample knowing that many records will be eliminated through field selection techniques due to sparseness of values, correlated content, lack of statistical significance, and partitioning the data into training and validation subsets. Approximately, 150K records were selected in the initial extraction and 42k remained in the final analysis. An important fact is that the collection of data needs to be representative of the member population in order to make inferences and result in actionable outcomes.

Fields: To enrich the in-house data, 2500 fields were purchased from a Claritas database which included information on behavior such as opinions, hobbies, lifestyles, purchasing attitudes. There are thousands of available fields, therefore the fields selected should be carefully chosen as related to the market population and include a request on the statistics for completeness of the fields. Due to the nature of survey data, responses can be sparse and more so as related to your company business or location. In this analysis, fields with missing values of 80% or above 80% were removed.

Exploration: After collecting the data, the next important step is to explore the data. Exploration begins by looking at the distribution of all the fields using PROC MEANS for continuous variables and PROC FREQ for categorical variables. These procedures will evaluate the identify patterns and discover new relationships. Both SAS procedures are very powerful tools used to identify outliers and assist in the decisions to keep or drop the fields.

Normality and Transformation: The next look is to evaluate the normalcy of the distribution. A review of the skewness of data is useful to identify outliers and determine if a transformation is required. Several approaches to address outliers are to remove those extreme values, regroup the values into deciles or replace the values with its logarithm. A widely accepted practice is to apply a log transformation to those variables with non-normal distribution to create a better symmetrical distribution and as a result control the variance. This transformation will improve interpretability, provide easier visualization and construct 95% confidence intervals.

Re-categorize: Another exercise is to evaluate the levels within a categorical field. The sparseness of levels may lead to combining levels however the business implication should be examined before masking possible important outcomes. A regrouping may not provide the actionable outcome originally intended or not reveal important stratum. An additional consideration is to keep missing values as a valid level. For specific variables this can be interpreted as ‘none’ or ‘not applicable’ and provide revealing information. There is also the option to ‘explode’ categories. For this transformation, each unique level is created as a new variable. A caution in transforming such a field is that a five level variable adds five new variables to your data, repeatedly doing this will add many more fields to the data.

STATISTICAL METHODOLOGY

The next processes implemented were to statistically analyze the fields for consideration in the analysis, conduct the Logistic regression model, perform Cluster analyses, and apply Propensity scoring.

Variable Reduction

In order to further reduce the number of fields in the analysis, the SAS procedure PROC VARCLUS method was run. The PROC VARCLUS procedure identifies redundant variables (dependent and independent) which otherwise could degrade the model by undermining the parameter estimates and confounding the interpretation. Basically, the procedure identifies groups of variables, resulting in clusters that are highly correlated and uncorrelated with other variable clusters, similar to a Principal Component methodology. The algorithm uses binary and divisive methodology

where all variables start in one cluster and the second eigenvalue is evaluated to the current threshold. If higher, the cluster is split. The repetitive process is conducted evaluating the correlation of each variable in the cluster and determining if the variance is better explained with that computation. An eigenvalue of .7 is commonly used which will result in fewer clusters.

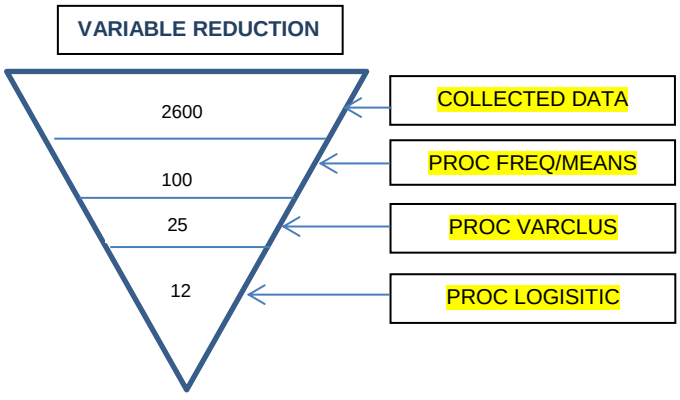
Sample PROC VARCLUS code:

```
PROC VARLCUS MAXEIGEN =0.7 data=mydat.clustervars;

VAR age gender tenure education ethnicity income dwelling num_persons;

RUN ;
```

Table 1. PROC VARCLUS example

4 Clusters		R-squared with		1-R**2	Variable
Cluster	Variable	Own Cluster	Next Closest	Ratio	Label
Cluster 1	Age	0.6376	0.0073	0.3651	
	Tenure	0.5176	0.0143	0.5942	
Cluster 2	Gender	0.7754	0.0338	0.4108	
	Education	0.0329	0.0038	0.1936	
Cluster 3	Dwelling	0.5977	0.02145	0.2091	
	Ethnicity	0.8909	0.0837	0.9164	
Cluster 4	Num_persons	0.8252	0.0338	0. 2385	
	Income	0.7500	0.0056	0.4385	

A small 1-R**2 value indicates that the variable has a strong correlation with variables within its own cluster. These variables are chosen as a driver for that cluster. Here in this example age, education, dwelling and number of persons were selected. This process was repeated using various combinations of fields.



After applying these several different approaches for variable reduction, the number of variables was reduced from 2600 to 100 to 25. A majority of the fields were eliminated due to sparseness and multi-collinearity. The next reduction of fields occurred when the logistic regression analysis was conducted. From the final assessment some of

the selected fields were age, marital status, ethnicity, education dwelling status (own/rent), urbanicity and employment at the household level.

Logistic Regression

A stepwise logistic regression was conducted using backwards elimination. This method is used because it is less inclined to exclude important inputs or spurious inputs when compared to forward computation. The downside is the heavy computational factor. The binary target was the Medical Loss Ratio (MLR) which was set at a specific threshold to establish a 0/1 value. A series of models initially using 20 - 25 variables was reduced to 8 - 12 variables in the final model in a sample of approximately 70,000 records. The c-stat and odd-ratios statistics were examined to determine what predictors contributed to modeling what the 'best' customer looked like. The c-stat is a good summary measure of model accuracy. A c-stat of .70 was considered acceptable and in this final model .71 was the strongest probability attained.

Sample regression code:

```
PROC LOGISTIC DATA= mydat.master_segmentation_analytic DESCENDING
    CLASS Income_Indicator (PARAM=REF REF=FIRST)
          Employment (PARAM=REF REF=LAST) ;
    MODEL Indicator = Num_Persons Gender Dwelling Previous_Insur Deductible Tenure Age Ethnicity
                      Marital Status Education /
    SELECTION=BACKWARD  SLS=0.05  CORRB  LINK=LOGIT  RSQ  LACKFIT  CTABLE
                PPROB= (.05 to 1.0 by .05)
                OUTPUT  OUT=mydat.logistic_out
                P=pred_prob
 RUN;
```

The results of the regression analysis using odd ratios were discussed with the Marketing department in 'business' language. Their Marketing input was very important in this phase of the analysis. Interpretation of the odds ratio was presented in a business-related manner such as the likelihood of a being 'best' customer increases 6% in the younger (25- 35) age ranges. The regression analysis showed what information defined the best customer in the member population however these results now needed to be translated to the non-member population for campaigns.

***C*luster Analysis**

Following the discussions with the Marketing department, it was decided the next steps were to conduct the segmentation analysis with current health insurance data representative of the state population. The regression analysis findings would need to be applied to non-members to act on the goal of targeting non-members who fit our 'best' customer. Nielson Health Insurance survey data was purchased for 1.6 million households in Arizona. This survey data included information such as demographics, insurance attitudes, channel preference (magazine, radio, TV ads, etc.), and utilization of services. A third-party company was engaged to match the members identified in the final regression model to those in the newly acquired state population data. Full name and address were used as the matching fields, thereby now identifying members and non-members in the 1.6 million households.

Cluster analysis was then performed to develop homogenous groups using key elements from the previous analysis: demographics and health insurance elements. The variable contribution of each cluster analysis was reviewed to determine the significant drivers specific to each cluster. Several hierarchal and non-hierarchal clustering methods were run: Average, Ward's, Centroid and K-Means.

For the Hierarchal methodology one needs to know the underlying distribution in order to determine which divisive method to use. All three methods (Average, Centroid, Ward's) were run taking into account the advantages and limitations of each. This methodology is more commonly used with survey data due to the ability to capture non-spherical clusters.

Generally, average linkage is more distinguishing, resulting in smaller within-cluster variation and less affected by outliers. Centroid linkage measures the distance between cluster centroids and is also less affected by outliers. Ward's method is variance based within clusters and is easily distorted by outliers. In this analysis, Ward's and Centroid methods were found to have the closest results identifying the smallest number of clusters with a small

number of variables, low correlations among the clusters and kurtosis and skewness close to zero. Some of the statistics generated from these procedures such as pseudo-F and Cubic Clustering Criterion (CCC) were used to help decide the best approach. The objective for this analysis was to find 5 – 8 clusters in order to easily draw conclusions and interpret for a direct mail campaign. The process was repeated using the different clustering techniques, splitting the data into test and validation subsets for replication reliability, and deleting variables one at a time.

Sample Ward's method code:

```
PROC CLUSTER data=mydat.azpop_all
        METHOD=WARD
        SIMPLE
        CCC
        PSUEDO;
VAR Insurvar1 Insurvar2 Channel Age Income Dwelling Tenure;
RUN;
```

An advantage of using the hierarchal method is determining the number of clusters that best fits this data and thus eliminates the need to guess at the number when using the non-hierarchal method k-means. A subset of the data was used in the cluster analysis because hierarchal methods are not as efficient with large datasets. When comparing various scenarios, the three statistics, Pseudo F-statistic, Overall R-squared and CCC should be of the highest values. Generally, a CCC over 2 is a good indicator.

The non-hierarchal K-means method was run on the full dataset using the cluster results from Ward's method indicating six to eight clusters statistical fit this data the best. The SAS procedure PROC FASTCLUS is used for this analysis.

Sample K-means method code:

```
PROC FASTCLUS data=mydat.azpop_all
        OUT = kmeans_out
        MAXC=7
        MAXITER=100
        MEANS=cluster_mns out =cluster_outall;
VAR Insurvar1 Insurvar2 Channel Age Income Dwelling Tenure;
RUN;
```

The final analysis of the clustering resulted in seven clusters with 10 primary characteristics. As a final exercise to validate the differences between the clusters, PROC UNIVARIATE was used to compare the attributes across clusters. This analysis also assists in describing the strength of each characteristic and how much it contributes to the profiling the group. Segment profiles were built on each cluster describing the predominate attributes such as age ranges, income bands, positive/negative attitudes toward health insurance, ethnicity and number of people in the household.

Propensity Scoring

The final step in the analysis was to assign the probability of being the 'best' member to the non-members within each cluster. In the logistic regression, the dependent variable was the value 0/1 as non-member / member and the covariates were the drivers from the cluster analysis. The analysis computed the propensity to be in one group or the other by using the predicted probability from the regression model. The data was then split by member/non-member,

sorted and rejoined using the probability score as the by variable. Each regression model was unique to a cluster. The covariates in each model were based on the strength of the driver when building the cluster with only a slight variable variation across models.

Sample Propensity Regression code:

```
PROC LOGISITC data=mydata.cluster1 descend noprint;
  Model indicator= age education dwelling num_persons ;
  Output out = mydata.cluster1_propens prob=PROB1;
Run;
```

After completing the propensity analysis for each cluster, every record had a MLR probability that was now considered a primary attribute of the each profile. The probability distributions ranged from .35 to .94 and used as input to a PROC UNIVARIATE procedure categorizing the data into deciles.

IMPLEMENTATION AND MONITORING

Campaign Design

Throughout the analysis frequent meetings with Marketing were held to discuss the variables eliminated, the drivers of the clusters and the final descriptions of the segments. Segment names were assigned depicting the strongest characteristics. Out of the seven clusters, four were identified as the most desirable for a direct mail marketing campaign. In each of the four clusters the top 10% MLR probability scores were selected.

The strategy developed was to attract more like those members tagged as 'best'. The campaign was designed with a Control group (usual direct mail list characteristics) and a Profile group (characteristics from the segmentation analysis). Specific mail list sources were aligned to the predominate attributes of the cluster. A timeline was established for the marketing campaign cycle within an eight month cycle.

Final Outcome Analysis

Leads from the campaigns were monitored over three campaign cycles followed with the tracking of application statuses to conversion rates defined as a new member enrollment. For these new members, enrollment demographics and medical claim data over an eight month period were collected for each group. The demographic characteristics and utilization of services were compared between the two groups. An MLR was calculated for each new member in both groups and percentiles compared as well as average MLR's by group. Simple statistical comparisons were conducted between the two groups using Chi-SQ tests and T-tests. The expectation was the P-values on each of the primary attributes including the MLR were significantly different between the Control and Profile groups.

CONCLUSION

This paper describes one approach to clustering and propensity scoring knowing there are many statistical methods as well as many additional steps to select, clean and standardize your data not mentioned here.

The customer profiling was able to target our 'best' customer for the campaigns prior to the implementation of the Healthcare Reform mandate. SAS software procedures provided the necessary tools to conduct an integrated view of the member population. However, going forward these profiles will most likely not represent the 2014 customers. The expectation is new members will have a wide variety of different characteristics as compared to the previously targeted members. The methodology applied pre-reform can now be applied to the 2014 member population.

REFERENCES

SAS Business Knowledge Series. "Customer Segmentation Using SAS Enterprise Miner. Course Notes 2009.

Nelson, Bryan. “Variable Reduction for Modeling using PROC VARLCUS.” *Proceedings of the SAS Global 2010 Conference*>. Minnetonka, MN: Fingerhut Companies Incorporated.

Collica, Randall. “Customer Segmentation and Clustering”. Second Edition 2011. SAS Institute Inc., Cary, NC

ACKNOWLEDGMENTS

I would like to acknowledge Ryan Sandhaus for the contributing analytic work on the Customer Segmentation project.

CONTACT INFORMATION

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

Name: MaryAnne DePesquo
Email: mdepesquo@cox.net