

Population Segmentation in a Healthcare Environment

MaryAnne DePesquo, BlueCross BlueShield of Arizona, Phoenix, USA

ABSTRACT

In this new era of Healthcare Reform (HCR) in the United States, there was a heightened emphasis on profiling individuals in the health insurance industry. The importance of understanding the characteristics of members for population analytics, healthcare economics, and marketing was a necessity prior to the enactment of the Affordable Care Act (ACA) and became more essential in the post ACA environment due to HCR mandates. The goal to provide an integrated 360 view of the individual would be beneficial for disease and care management programs, balancing population risk, increasing market share, providing wellness programs to improve member's needs and assisting the consumer to make informed choices about their care through quality and cost transparency. By profiling the individuals, an approach for care can be built for each member segment classification as predefined by specific attributes. This paper describes how medical claim data, third-party psychographic data and SAS® were used to perform the segmentation analytics for an insured population. The discussion will include variable selection procedures such as PROC VARCLUS and cluster analyses using PROC FASTCLUS and PROC CLUSTER.

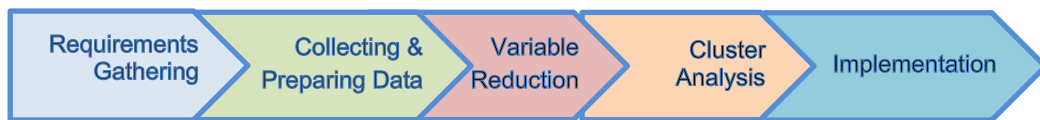
INTRODUCTION

The enactment of the Affordable Care Act (ACA) impacted the United States 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 when 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 support business objectives.

PROJECT OVERVIEW

The first questions asked in strategic planning are '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 product development and 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, and implementation.



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

- 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
- 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 an over-large sample knowing that many records would 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, and purchasing attitudes. There are thousands of available fields, therefore the fields selected should be carefully chosen as related to the market population in addition to reviewing 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 identify patterns and discover new variable 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. To review the asymmetry of the data, skewness is used 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 standard practice is to apply a log transformation to those variables with non-normal distributions to create a improved 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 nor 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 descriptive information. One should also consider 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, and conduct Cluster analyses.

Variable Reduction

In order to further reduce the number of fields in the analysis, the method used was the SAS procedure PROC VARCLUS. The PROC VARCLUS procedure identifies redundant variables (dependent and independent) which 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 within 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. A second approach is to specify the maximum number of clusters.

Sample PROC VARCLUS code:

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

VAR age gender tenure education ethnicity income dwelling num_persons;

RUN ;
```

PROC VARCLUS output example:

4 Clusters		R-squared with		1-R**2	Variable
Cluster	Variable	Own	Next	Ratio	Label
		Cluster	Closest		
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	

The small 1-R**2 value within a cluster indicates that the variable has a strong correlation in that cluster and weak correlation with variables in other clusters. 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.

Logistic Regression

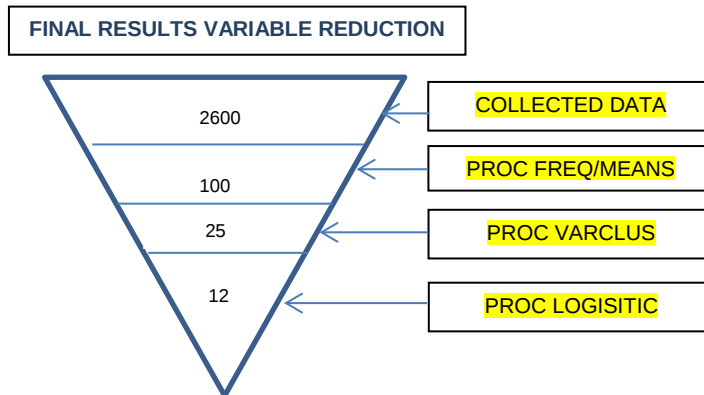
A stepwise logistic regression was conducted using backwards elimination. This method is used because it is less inclined to exclude important inputs or spurious inputs when compared to forward computation. The downside is the heavy computational factor. The binary target was the Medical Loss Ratio (MLR) which was set at a specific threshold to establish a 0/1 binary 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 which predictors contributed to modeling and 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. The Marketing leaders 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.



After applying these several different techniques 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 logistic regression analysis further examined the importance of variables contribution to the model. 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.

Cluster Analysis

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 definition. 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 performed: 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 executed 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 clearly interpret findings and plan 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 statistically fit this data the best. The SAS procedure PROC FASTCLUS was 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 the degree of contribution each characteristic has to the group profile. 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.

CONCLUSION

This paper describes one approach to clustering 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 most likely will not represent the 2014 customer population. The expectation is new members will have a variety of different characteristics as compared to the previously targeted members. However, the methodology applied pre-reform can now be applied to the 2015 member population.

REFERENCES

- SAS Business Knowledge Series. "Customer Segmentation Using SAS Enterprise Miner. Course Notes 2009.
- Nelson, Bryan. "Variable Reduction for Modeling using PROC VARCLUS." *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

CONTACT INFORMATION

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

Name: MaryAnne DePesquo
Email: mdepesquo@cox.net