

Customer Retention Predictive Modeling in HealthCare Insurance Industry

Jin Su, Kimberly Cooper, Tina Robinson, and Brad Jordan,
BlueCross BlueShield of Florida, Jacksonville, FL

ABSTRACT

Current economic volatility has increased the need in industries with limited resources to identify new opportunities to retain its current customer base. Customer retention is one of the key factors to sustain business growth. Developing retention models assist the company by identifying key drivers for increasing retention and focusing the communication to the "right" customers. This paper addressed the difficulties facing the healthcare industry in constructing a time-series data mart and developed a retention predictive model that can derive retention scores for each customer. Top factors affecting retention were identified through the modeling process; and some of their applications were discussed in this paper.

INTRODUCTION

The health insurance industry faces more competition than ever. The heated competition provides consumers with more choices, i.e., the acquisition cost increased and customer loyalty decreased. It costs at least 6-10 times more to acquire a new customer than it does to retain an existing one (Pogol, 2007), and acquisition costs are a large portion of an insurance company's administrative cost. Therefore, customer retention becomes one of the key factors in sustaining business growth. One of the benefits of a retention study is a better understanding of the customer, which can lead to improved customer service. The knowledge gained in this effort can therefore be applied in many areas of the company, and deserves the company's full attention. Given today's economic pressure, companies have limited resources to devote to every customer from every perspective. This requires the company to focus on what matters most in order to maximize its return on investment. The company is challenged to identify key factors associated with decreased retention and also identify the "right" customers with which to communicate. Fortunately, customer retention predictive modeling can serve this purpose.

Much research has been completed on using data mining techniques to classify or predict a customer's likelihood to churn (Ngai, *et al.*, 2009). Many data mining techniques are studied, for example, clustering (Bose *et al.*, 2009), decision tree (Lemmens, 2006), neural network, random forest (Xie, *et al.*, 2009), survival analysis (Lu, 2002), generic algorithm (Pendharkar, 2009), and SVM (Archaux, *et al.*, 2004). However, many of them are for the purpose of research, and provided little guidance on "real-world" application. This paper serves the needs for such requests.

This study collected comprehensive customer information from four sources, and built a one-row-per-member-per-month data mart containing 643 attributes. Customer segmentation through supervised clustering was performed in the first stage, and classified the total study population into two clusters. Logistic regression models were then developed for each of the clusters using SAS/STAT and SAS/Enterprise Miner 5.3. Several questions were answered in this paper, such as, how to construct data mart that meets data mining requirement in spite of the data complexity and integrity problems; how to identify key drivers to retention, what's their relative importance, and how to apply the predictive model to your business.

DATA STRUCTURE AND DESCRIPTION

Before any data mining effort can begin, data must be extracted and prepared in a way to maximize the value of the available data and improve the results of complex statistical analysis. The goal is to bring together all of the disparate data in an organization and transform it into meaningful, useful information. Although data construction can consume more than 50 percent of the data mining work effort, little has been published on how best to approach this task.

Many of today's corporations have built vast data warehouses with hundreds of highly normalized tables that can make a simple task of extracting actionable information difficult and, often, cost-prohibitive. The effect of distorted or missing data can have an adverse effect on performance and introduce bias into the investigation. Utilizing the data that is collected to maintain day-to-day business in a way that helps us learn more about our customers is the best tool we have to increase sales, decrease attrition, improve internal processes, reduce call volumes, and to better meet our customer's ever-changing needs.

An enhanced database structure that moves away from the traditional view of *collecting data* to a more focused view of *accumulating facts* can vastly improve data mining results. These two tasks are very different. In the healthcare

industry, a specific life event might trigger a special need for the customer. A life event might include: marriage, divorce, addition of child, a child reaching the age of 18 or about to graduate from college, addition of a parent as a dependent, movement in or out of counties, and retirement. All of these events can and should trigger an action on the part of the business, and many are directly related to the member's propensity to stay (retention).

One of the dilemmas faced by many businesses is the overwhelming volume of data at their disposal. Most businesses have enormous amount of data and not much actionable information. So, should we just proceed with dumping everything we know about a customer into another database, or should we attempt to transform it in a way that would aid in identifying these triggers? "The goal of any model should be parsimony, i.e., to find the simplest explanation of the facts using the fewest variables" (Yu, Wang, & Lai, 2006).

Most databases are very textual in nature. However, computers are very numeric. The closer we can bring the data from its current textual state to a more numeric state, the faster and more predictive the data mining process will be. So, what is the difference? Take for example; a family may consist of 2 adults and 3 children. The existing database structure might typically house 5 rows of data (one for each member) to represent this customer-relationship. However, the new data structure might convert this information into a household fact table that reflects changes in the household over a period of time. See Figure 1.

Cust_Nbr	Begin_Date	End_Date	Total_HH	Married	Total_Deps	Deps_00_05	Deps_06_12	Deps_13_18	Deps_19_25	Deps_26_99
123456	01/01/2004	12/31/2004	1	0	0	0	0	0	0	0
123456	01/01/2005	12/31/2005	4	1	2	1	1	0	0	0
123456	01/01/2006	12/31/2006	4	1	2	0	1	1	0	0
123456	01/01/2007	12/31/2007	1	0	1	1	0	0	0	0

FIGURE 1. HOUSEHOLD FACT TABLE

This structure clearly shows significant changes in the make-up of the household which may prove to be an important predictor of future needs and movement. The data structure supporting this analysis is constructed at a one-row-per-member-per-month level, takes transactional data from four different data sources, and combines them into a wide customer analytic record containing 634 attributes. However, the final retention model is built at the household level in an effort to be consistent with the business process of communicating directly to the subscriber. The four data sources and example data elements are briefly explained in Table 1.

TABLE 1. CUSTOMER ANALYTIC RECORD

DATA CATEGORY	DESCRIPTION	EXAMPLES
ENROLLMENT	Identifies when a customer enrolled and what product(s) they are enrolled in.	- Membership tenure. - Product type, count and migration. - Coverage type and gap. - Method of payment. - Sales channel. - Premium.
DEMOGRAPHIC	Includes internal customer data collected through application processes and purchased customer and household information.	- Member's Age, ethnicity, gender, and marital status. - Number of adults, children in household. - Dwelling type and length of residence. - Hobbies, travel and sport/leisure preferences, and limited credit indicators.

DATA CATEGORY	DESCRIPTION	EXAMPLES
CLAIMS	Includes summarized claims information.	- Allowed, paid, deductible, copay, and coinsurance amount. - Location of service. - In/Out of network indicator.
INQUIRY	Includes customer contact data made through various channels in which the customer has initiated to discuss benefits, claims, and other issues.	- Monthly counts by contact mode and content. - Average inquiry resolution times.

The model was built based on the last active record of a member; the customer status (Active/Cancel) was the target to predict. History data was used to capture member demographic and product changes over time, as well as the cumulative amount of financial information.

SUPERVISED CLUSTERING

The success of predictive modeling depends on statistical similarity, which includes the similarity of behaviors between different customers as well as the similarity between customers' past behaviors and future behaviors. Therefore, clustering analogous customers together and constructing predictive models is a better approach than congregating all customers and building a one-size-fits-all model (Bose *et al.*, 2009).

Clustering is typically applied in an unsupervised learning framework when no prior classification knowledge is available. Based on the value of data attributes, clustering techniques maximize both the similarity of customers within the same cluster and the dissimilarity of customers between the different clusters. When prior classification information is known, the concentration of each class can be incorporated into similarity/dissimilarity measurement to supervise the formation of clusters so that the probability density with respect to one single class is as distinctive as possible among each cluster. For this study, we use SAS/Enterprise Miner as the clustering tool, classification knowledge was unable to be applied during the clustering process, and instead, was used as final supervision to determine the best clusters from a group of clusters formed by applying different clustering methodologies.

SELECTION OF ATTRIBUTES FOR CLUSTERING

When selecting the attributes used for clustering, it is unrealistic and highly inefficient to use all 634 attributes considering the exhaustive computational time and difficult interpretability of the customers' portfolio in each cluster. We picked attributes that are highly correlated with retention, but hard to take action on, for example, member's enrollment tenure. It's known that the longer a customer is enrolled the more likely he is going to continue to stay; such relationship is shown in Figure 2. However, customer's tenure is not an actionable item. There are three different types of tenure in our data mart, i.e., total tenure with the company, tenure with the current product type, and tenure within the policy. In addition, we also use a series of event occurrence flags, such as whether a member had a gap in coverage, whether they switched segments, got married, moved, etc. These event flags are captured for occurrence in the past three months, 4-6 months, and more than 6 months ago.

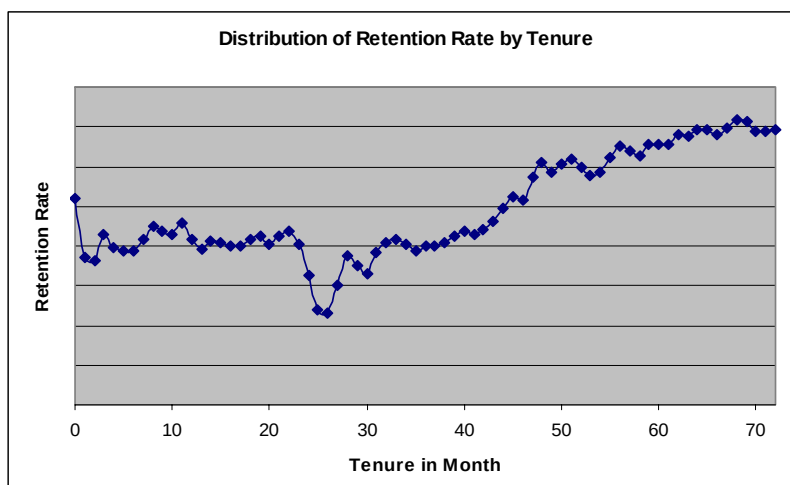


Figure 2. Customer Retention Rate distribution by Enrollment Tenure

RESULTS

Enterprise Miner 5.3 provides the capabilities of Hierarchical clustering, K-means clustering and SOM/Kohonen technique. We tried all three techniques with different settings in the Enterprise Miner property panel, and produced four groupings (see Table 2). Chi-square test between the cluster variable and customer status variable is performed as a significance test for the association of two variables. From Table 2, we see all groupings are highly significant. Due to the large sample sizes and potential for inaccurately rejecting the null hypothesis, further validation of the distinction between clusters was determined by comparing the average retention rate for each cluster, this resulted in the selection of K-means clustering with two clusters.

Table 2. Clustering results of applying different techniques

Methodology	Number of Clusters	Chi-square (DF, Value)
K-Means Clustering	2	(1, 6144.5)
K-Means Clustering	3	(2, 4363.9)
Hierarchical Clustering	3	(2, 5305.6)
SOM/Kohonen	4	(3, 6409.4)

Figure 3 illustrates the target variable distribution within each cluster after clustering. All customers are grouped into either one of the two clusters; cluster 1 is about 70% of total population. The difference of retention rate between two clusters is over 13%, which indicates a successful separation of customers in this first stage.

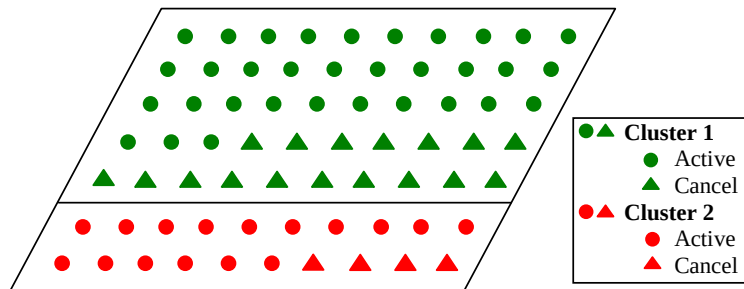


Figure 3. Illustration of clustering results

MODEL BUILDING

The modeling process was completed in the following steps.

EXPLANATORY DATA ANALYSIS

Explanatory data analysis is a necessary first step in the modeling process; data attributes are assessed and audited in this step. Three approaches were taken in this analysis to understand attributes and also discover possible problems. The StatExplore node in Enterprise Miner was used first to take a cursory view at variable distribution and detect variables with missing values and outliers; smoothed scatterplots and logit plots were then generated between each numerical variable and target variable to assess whether linear relationships existed or could exist after transformation (Ratner, 2003); finally, a bar-chart was used to analyze the association between categorical variables and the target variable.

Through this step, variables with a high percentage of missing values and those unlikely to be imputed with reasonable values were thrown away. The data miner also gained a better understanding of all the attributes under study, such as, which nominal variables with multiple levels needed to be re-grouped, and also which numerical variables needed to be capped or transformed.

VARIABLE REDUCTION AND MASSAGING

Numerical variables are selected primarily in Enterprise Miner by the Variable Selection node, and the VarClus node. Categorical variables are chosen by their information value. All of the categorical variables are kept when their information value is greater than 0.01. The reason for using information value instead of univariate analysis through chi-square test is because test results are almost always significant since the size of the study population is large. Besides statistical analysis, business knowledge is also used to combine the information of some variables and create a new representative one. For example, customer out-of-pocket amount is a summation of contract premium,

deductible, copay, coinsurance, and non-covered amounts for both medical and pharmacy claims. This step reduced the total number of attributes to approximately one hundred.

Variable massaging refers to collapsing categorical variables with many levels, and transforming, capping or ranking numerical variables to better fit the model. Variable manipulation increases model stability and accuracy.

MODEL TRAINING

We first partition data into 70% for training and 30% for validation using stratified sampling on the target variable. User defined interactions and second-degree polynomial terms are added in Enterprise Miner Regression node. Stepwise selection with both entry level and stay level of 0.05 is used for final variable selection. Selection criteria were set to minimize the validation misclassification rate.

MODEL VALIDATION

The chart of Cumulative Percent of Captured Target and ROC curve generated from validation dataset were used to assess the model performance. Figure 4 and Figure 5 are the modeling results from cluster 1. From Figure 4 we can see that the top two deciles capture about 53% of the cancellations; the top five deciles capture about 92% of the total cancellations. The area under ROC curve is 0.87.

After building two predictive models for each of the clusters, we combined the results in order to rank the propensity score for the whole population. We adjusted the target prior probability for each cluster and made the probability score comparable.

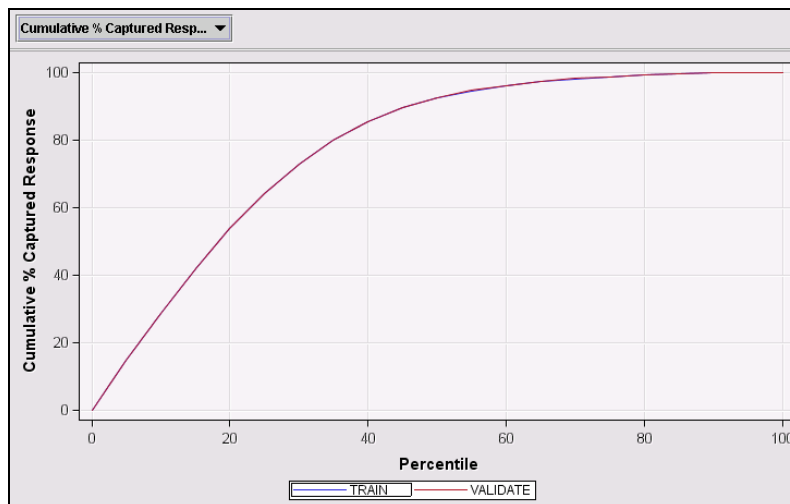


Figure 4. Cumulative Percent of Captured Target

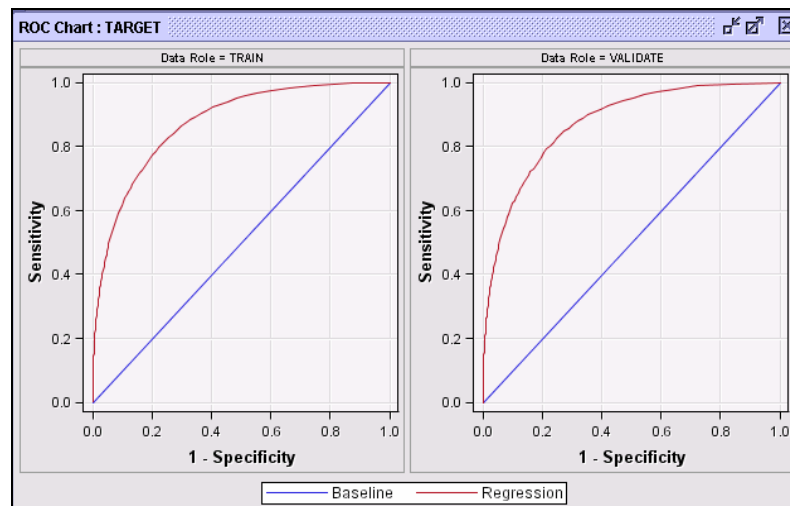


Figure 5. ROC Chart

The variables that are kept in the final model are important factors to retention. Results were validated with experts from various business areas to confirm interpretations and identify any potential caveats. For example, in our initial model, we found that customers who had a written inquiry were more likely to cancel; however, this effect has built-in bias since the company requires customers to submit a written document to cancel their contract, and this business process is recorded as a written service inquiry in our database. Model training, validation and business process checking are an iterative process until a stable and accurate model is built.

FINDINGS AND APPLICATIONS

A number of key drivers of retention were identified during the model building phase of the project. Conveying the model results to business process owners based on their impact to retention was needed before the results could be incorporated into any business processes. In ordinary regression, standardized coefficients are used to compare the importance of attributes in the model. In logistic regression, the standardized coefficients are provided by Enterprise Miner for interval variables in the model results. In order to produce a result that is more understandable to the business we created dummy variables for each value of the categorical variables, calculated their standard deviation, and used the same procedure utilized in Enterprise Miner to get their standardized coefficients (Menard, 2004).

The Top 10 factors for retention and their relative magnitude on a scale of 1-10 are shown in figure 6 and 7 for cluster 1 and 2, respectively. The factors in green on the right side of the graph have positive effect on retention; the factors in red on the left side of the graph are negative. Based on their importance ranking, the company can devote resources to actionable findings in proportion to the effect on retention.

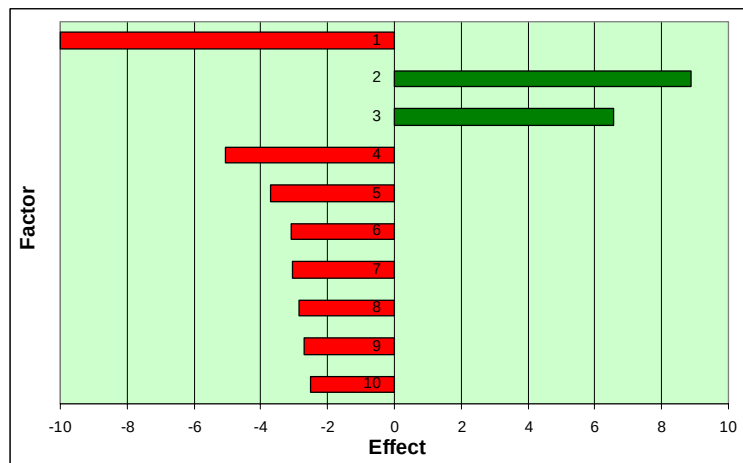


Figure 6. Top 10 Important Factors to Retention (Cluster 1)

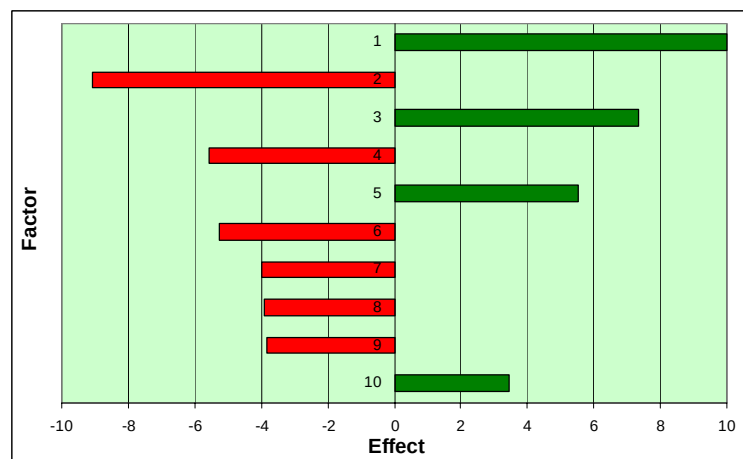


Figure 7. Top 10 Important Factors to Retention (Cluster 2)

In addition to discovering the key factors associated with retention, the model can also be applied to a company's various retention programs. The value that the model brings depends on the member size targeted in each program,

as well as the cost of the program. If the population size is small, such as a few hundred in a win-back program; or the cost of a program is low (i.e., an e-mail communication), then we do not need the model to prioritize members. In cases where a direct communication with thousands of members is needed, and the cost of the intervention is high, the retention scores assigned to each member by the model can greatly reduce the cost of the campaign by focusing the efforts on the members most at risk for attrition, thus increasing the ROI of the intervention.

The company is using both the model scores and the insights gained from the accompanying reports to focus its efforts on the most actionable facts. The added focus on insights bring to the business have been extremely well received and have assisted the company in prioritizing its efforts on the activities most likely to retain its members. The alternative is a shot-gun approach where many programs and activities are undertaken, many of which will have marginal effects on member retention or tenure.

CONCLUSION

This study presented what we believe to be a practical process of how to build a reliable data structure used for statistical analysis and predictive modeling. Although we use health care industry to illustrate the usefulness of this approach, it can be used for any industry with large amounts of data about customers. Comprehensive data about a member was collected from four subject areas, but the structure that has been developed is scalable and extensible as other useful business facts are discovered. A retention model was developed based on the new data structure that can derive attrition scores for each customer in two stages 1) segmentation on customers according to key attributes and 2) logistic regression models are developed by using SAS/STAT and SAS/Enterprise Miner 5.3. The top 10 factors impacting retention have been identified and their relative importance was quantified. The modeling results and accompanying customer profile reports show that the data mining techniques can help companies allocate limited resources based on facts, rather than intuition. Allocating resources to programs that have a demonstrable affect on retention/attrition achieves a higher return on investment for retention programs.

REFERENCES

- Archaux, C.; Martin, A.; Khenchaf, A, *An SVM based churn detector in prepaid mobile telephony*, International Conference on Information and Communication Technologies (ICTTA), Damas, Syrie, 19-23, 2004
- Bose, Indranil; Chen, Xi, *Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn*, Journal of Organizational Computing and Electronic Commerce, 19: 133-151, 2009.
- Dick, Alan S.; Basu, Kunal, *Customer Loyalty: Toward an Integrated Conceptual Framework*, Journal of the Academy of Marketing Science, 22(2): 99-113, 1994
- Lemmens, Aurélie; Croux, Christophe, *Bagging and Boosting Classification Trees to Predict Churn*, Journal of Marketing Research, 43(2): 276-286, 2006.
- Lu, Junxiang, *Predicting Customer Churn in the Telecommunications Industry – An Application of Survival Analysis Techniques*, Proceedings of the 27th Annual SAS Users Group International Conference, Cary, NC: SAS Institute Inc., 2002.
- Menard, Scott, *Six Approaches to Calculating Standardized Logistic Regression Coefficients*, The American Statistician, 58(3): 218-223, 2004
- Ngai, E.W.T.; Xiu, Li; Chau, D.C.K., *Application of data mining techniques in customer relationship management: A literature review and classification*, Expert Systems with Applications, 36: 2592 - 2602, 2009
- Pendharkar, Parag C., *Genetic Algorithm Based Neural Network Approaches for Predicting Churn in Cellular Wireless Network Services*, Expert Systems with Applications, 36: 6714-6720, 2009
- Pogol, Gina, *Tips for Cost-Effective Customer Retention Management*, www.crm2day.com/library/docs/50577-0.pdf, 2007
- Ratner, Bruce, *Statistical Modeling and Analysis for Database Marketing: effective techniques for mining big data*, Boca Raton: Chapman & Hall/CRC, 2003
- Svolba, Gerhard, *Efficient "One-Row-per-Subject" Data Mart Construction for Data Mining*, Proceedings of the 31st Annual SAS Users Group International Conference, San Francisco, CA: SAS Institute Inc., 2006.

Xie, Yaya; Li, Xiu, Ngai, E.W.T.; Ying, Weiyun, *Customer Churn Prediction Using Improved Balanced Random Forests*, Expert Systems with Applications 36,: 5445–5449, 2009

Yu, L.; Wang, S.; & Lai, K., *An Integrated Data Preparation Scheme for Neural Network Data Analysis*. IEEE Transactions On Knowledge And Data Engineering, 18(2): 217-230, 2006

CONTACT INFORMATION

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

Jin Su
Marketing Analytics
BlueCross BlueShield of Florida
4800 Deerwood Campus Parkway
Jacksonville, FL, 32246
(904) 905-7679
jin.su@bcbsfl.com

Tina Robinson
Marketing Analytics
BlueCross BlueShield of Florida
4800 Deerwood Campus Parkway
Jacksonville, FL, 32246
(904) 905-5851
tina.robinson@bcbsfl.com

Kimberly Cooper
Marketing Analytics
BlueCross BlueShield of Florida
4800 Deerwood Campus Parkway
Jacksonville, FL, 32246
(904) 905-0067
kimberly.cooper@bcbsfl.com

Brad Jordan
Marketing Analytics
BlueCross BlueShield of Florida
4800 Deerwood Campus Parkway
Jacksonville, FL, 32246
(904) 905-6003
brad.jordan@bcbsfl.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 and product names are trademarks of their respective companies.