Paper 170-2013

What Do Your Consumer Habits Say About Your Health? Using Third-Party Data to Predict Individual Health Risk and Costs

Satish Garla, SAS Institute; Albert Hopping, SAS Institute; Rick Monaco, SAS Institute; Sarah Rittman, SAS Institute

ABSTRACT

The Affordable Care Act is bringing dramatic changes to the health care industry. Previously uninsured individuals are buying health insurance and consuming health care services differently. These changes are forcing insurers to reevaluate marketing, engagement, and product design strategies.

The first step in addressing these challenges is understanding the financial risk of new entrants into the marketplace. How do you predict the risk of a person without any historical cost information? What if all you know is the name and address?

Historically, the finance industry has long used third-party consumer data to predict future finance habits and credit risk. This paper takes a look at applying advanced analytics from SAS to third-party data in order to predict health care utilization risk and costs.

INTRODUCTION

The Patient Protection and Affordable Care Act (PPACA), also sometimes referred to as Health Care Reform, is sweeping legislation that is dramatically and permanently altering the US health insurance landscape. Presently, health insurers across the US are scrambling to prepare for some of the most critical elements of the reform package, which are scheduled to take effect January 1, 2014.

Some questions asked include: What are these changes, and how will they impact the way health insurers run their business? What are the major risks and threats brought about by the law? Where are the opportunities for successful market expansion?

This paper discusses the components of the law that effectively create a new "consumer" marketplace for health insurers. It also reveals why this new marketplace is creating the necessity for health insurers to develop capabilities to predict and manage individual level health risk in non-traditional ways. In addition, this paper delves into a case study that explores how SAS partnered with a US health insurer to uncover ways to leverage consumer data and powerful predictive models in order to understand the health care utilization risk of unknown individuals. Finally, the future analytical and business applications of the lessons from this work are explored.

HEALTH CARE REFORM BRINGS MILLIONS OF NEW, UNKNOWN PEOPLE TO MARKET

Beginning in 2014, in an effort to bring as many individuals as possible into the health insurance market, all American citizens will be required to purchase "qualified" health insurance plans – or pay an IRS penalty. To encourage the purchase of coverage, the law provides a government-funded subsidy to all citizens with household incomes at or below 400% of the federal poverty limit (FPL). Individuals and families will use this subsidy to purchase insurance via a new distribution channel known as public health insurance exchanges. The exchanges will provide "apples to apples" comparisons of available health plans to all consumers. These exchanges will enable consumers to shop, select, purchase, and receive their subsidy for coverage – all online. The exchanges will be established and maintained either by the state or federal government based on the individual state's preference. As of February, 2013 only 17 states have elected to operate their own public exchange and seven states will partner with the federal government. This means that the federal government will establish and operate the other 26 exchanges.

As a result of these components of the law, health insurers anticipate millions of currently uninsured individuals will enter the market sometime in the next few years. Estimates range anywhere from 20 to 40 million individuals with some models exceeding 60 million, depending on what happens in the employer group segment.

UNCERTAINTY ABOUNDS

Here is where the challenges and opportunities begin. Health insurers have historically placed little to no focus on individual consumers of their product. They have sold to primarily large and mid-sized employer groups. In many cases, health insurers have also sold to benefits consultants who act as third party brokers or agents for multiple carriers – thereby removing themselves even further from the end user of their product. Consequently**, insurers know very little about individual consumers of health care.** They are finding themselves asking questions such as: "What will these individuals want to buy?", "How will they purchase?", "What do they value?", "What is their price

sensitivity?", and perhaps most importantly, **"How do I understand and predict future medical expense of these individuals?"**

This lack of understanding is further exacerbated by the fact that many of the individuals entering the new marketplace are currently uninsured, and may actually be habitually uninsured; These individuals do not resemble the traditional end user for health insurance. Health insurers are struggling to understand and predict what this new population will look like as well as what the ultimate business and financial impact will be once the new markets become activated.

Another layer of uncertainty that plagues the health insurance strategist today is the question of whether the subsidies and penalties will actually be effective at driving the purchase of insurance with younger, healthier populations. The mandate is relatively weak, especially in the first few years of post-reform implementation. Many industry observers believe that young, healthy individuals will choose to forego the purchase of insurance and pay the penalty – reserving their decision to purchase insurance until they really need it; This could enable these individuals to save thousands of dollars annually in the process.

NEW UNDERWRITING REQUIREMENTS AND PROFIT & LOSS LIMITATIONS CHANGE THE GAME

In most business models, instantly gaining access to a new marketplace with millions of new consumers (even unknown consumers) would be a dream come true. In the post-reform health insurance market however, this situation is full of uncertainty and potential risk. This is true primarily because insurers will be required to price, underwrite, and issue policies to individuals, having very little ability to incorporate medical and financial risk into this process.

Under the law, health insurers are required to issue policies to any US citizen who applies for coverage, regardless of pre-existing condition or medical history. This is called "guaranteed issue." Many insurers already operate this way today. Therefore, this provision would not be a significant change for them. In fact, most insurers enthusiastically support this clause, as it represents a true movement towards reform and solid access to coverage and care for the population. Where guaranteed issue becomes potentially problematic for insurers is when it is combined with the new underwriting requirements.

In the post-reform marketplace of individual and small employer groups, health insurers may use only three pieces of information to underwrite policies:

1. age
2. geographic location
3. tobacco use

Anyone familiar with traditional actuarial principles understands that statistical models used to predict risk and health care utilization are based on many individual characteristics. These models can be quite accurate in predicting future expense when applied consistently. As health insurers contemplate losing control of and access to these traditional analytical approaches, they are concerned about how their lines of business will perform moving forward. As a result, there is tremendous interest and a growing sense of urgency for defining new and different ways to understand and forecast risk of populations, as well as utilization trends.

Post-reform, health insurers will be able to use the three data elements listed above to underwrite individual and group policies. However, they will be fairly limited when it comes to pricing (rating) individuals based on these criteria. The law states that the premium charged by a health insurer to an individual in the highest "risk" category cannot be more than three times the premium charged to an individual in the lowest "risk" category for similar products. Effectively, this clause in the legislation restricts a health insurer's ability to price according to actual projected risk.

Industry observers believe this requirement may force insurers to skew their pricing models higher than they would otherwise. This could have the unintended, and very negative, impact of pricing lower risk (typically younger, healthier) individuals out of the market. This could further exacerbating the anticipated situation in which a higher proportion of high risk, high cost individuals purchase insurance, while lower risk, lower cost individuals postpone their decision to enter the market.

In an effort to avoid any one health insurer suffering catastrophic consequences as a result of taking on too much of the high risk, high cost population, the legislation provides a mechanism for the transfer of funds to insurers who bear a larger portion of the high risk population. This mechanism is called risk adjustment. Risk adjustment creates a very interesting opportunity for health plans to maximize revenue by acquiring high risk individuals and helping them to effectively manage their health. The transfer of funds is based exclusively on diagnoses ("risk") and **not on actual medical expense**. Theoretically, a health plan could acquire a high volume of individuals with a diagnosed condition, keep them healthy and out of the hospital (thereby reducing their own expense), and receive additional revenue for the aggregated "risk" of taking on these individuals.

WHAT DOES IT ALL MEAN?

Insurers are faced with a new market where they must accept all comers, may not be able to price adequately for differences in future medical expense, and know very little about their new customer base. As a result, insurers are extremely concerned that they will attract a disproportionate volume of high risk, high cost members who cannot be adequately managed from a health and wellness perspective. This scenario would be devastating for the average insurer, potentially leading to catastrophic failure of the business. Health insurers also envision the very real possibility that a nontraditional player or series of players might enter the market and completely flip the model on its head. For example; What if Wal-Mart began offering health insurance to the 40 million Americans who find themselves suddenly in the market for a health plan?

HEALTH INSURERS HAVE A LOT OF WORK TO DO, TO QUICKLY PREPARE FOR THE NEW CONSUMER MARKETS

The most logical first step is to develop the ability to understand and predict individual consumer behavior. Many insurers are looking to other, more mature, consumer-centric industries for clues on what to do first. Leaders in financial services, retail, and telecom have been using publicly available consumer data for years to predict behavior, preference, propensity, and services utilization of their individual customers. It's no surprise that insurers are beginning to follow suit, exploring the possibilities and asking the following questions:

- **Can we use the same approach to combining consumer data with our own data to predict financial risk, health and other important pieces of information we want to know?**
- **If it's possible, how powerful would the insights be – and how would it work?**

And maybe most importantly:

- **In a restricted and complex market, how can I use these insights to improve the health of my members and grow my business?**

In a recent analytical project with one such large health insurer, SAS explored all three of the previous questions, with very compelling results.

THE CASE: TEST THE HYPOTHESIS THAT CONSUMER DATA PROVIDES ENHANCED VALUE TO PREDICTIVE RISK MODELS

In 2012 SAS worked closely with a major US health insurer to answer the question of whether publicly available consumer data could enhance the quality and effectiveness of predictive risk models and to what degree.

In this project, there were two tasks:

1. demonstrate the value of applying advanced analytical methodologies to our health plan partner's consumer and traditional data to better predict health care costs of existing members
2. determine whether this model can be used effectively to predict cost or classify new individuals without having access to historical claims and utilization data

This paper focuses on the second task. The project team consisted of SAS consultants with strong data management and analytical capabilities, supported by industry experts from healthcare and financial services. Actuarial experts were also involved to validate the models and compare them with traditional risk models. The team worked for a period of five months to develop these models. The duration for a project of this kind depends on the number of models being developed and also largely on the size of the data. As explained in the previous section, consumer data itself could constitute a massive database of information. In addition to cleansing this data, significant data management effort is needed to make it ready for analytics. Some of the key challenges with consumer data are discussed in later sections of this paper.

At a high level, some of the key tasks undertaken in this project include the following:

- data extraction and exploration
- data management
- model development and testing
- scoring and analysis

In an effort to ensure that the business problem and solution were clearly defined and agreed upon prior to execution of the project, a very specific approach from the business context perspective was followed.

STEP 1: DEFINE RISK IN THE HEALTH INSURANCE CONTEXT

The first step in the process was to ensure a consistent and accurate definition of risk. A point to consider is: How is risk used in health care and how does it apply to this discussion?

At its core, risk is simply uncertainty. More exactingly, risk comes from uncertainty. Uncertainty comes first as the concept of uncertainty is greater than that of risk. While the details of risk are complex, the concept of risk itself is straight forward. Risk is the likelihood that something we quantify will be an amount worse than we expect. With a general definition, the term risk can be used correctly and still mean many things. Adding to the confusion, the term risk is very often misused and its quantification complicated.

In the health insurance context, risk can be defined in a number of ways. The most common among these focus on either financial loss or the deterioration of health status by a member or group of members. Health insurance companies are very focused on protecting against both forms of risk.

The first form of risk, financial loss, is obvious. Insurers have built their entire business model around predicting medical expense and charging premiums across a large pool of individuals, sufficient to cover the aggregate claims incurred while returning a reasonable profit after other expenses. Avoiding excessive risk brought about by acquiring and retaining a disproportionate number of potentially high cost members is critical to financial viability of health insurance organizations.

The second form of risk, deterioration of health status, is a form of risk closely linked to financial loss for an insurer. This second risk has additional measurement criteria that include severity of condition, progression of disease, quality of care provided by doctors and hospitals, and a number of additional elements. Deterioration of health status is often difficult to measure and requires long-term surveillance and comparative analysis.

For the purposes of this project, the definition of risk was limited to potential financial loss. More specifically, financial loss is the potential expense associated with the expected medical utilization of an individual for a future time period.

STEP 2: UNDERSTAND BASELINE/TRADITIONAL DATA AND AN ANALYTICAL APPROACH WITHIN THE HEALTH INSURANCE LANDSCAPE

In order to develop a meaningful and effective approach to leveraging non-traditional techniques, an understanding of the traditional data sources and analytical approaches utilized today by health plans for risk prediction was needed. Healthcare data is massive and complex, and insurers have access to a tremendous volume of longitudinal data about their members.

Every organization has made use of data mining to derive valuable insights from this data and make effective decisions. Data mining, often (erroneously) synonymously used with predictive modeling, is the process of identifying and extracting patterns from data. Predictive modeling involves the application of analytical methods to predict the future, based on past information. A key objective is to derive and estimate the effect of relationship between a specific healthcare outcome and its related risk factors. Analytical models are used in the healthcare domain for underwriting, fraud and abuse detection, disease prediction, care management, and so forth. Changing healthcare regulations results in a great need for innovative methods and advanced analytics for understanding the individual member.

Examples of the type of traditional data that is useful in predicting individual level behavior and risk include, but is not limited to the following:

- claims data
    - place of service
    - dollar amount of service
    - primary and secondary diagnoses codes
    - provider (doctor, hospital) performing service
- enrollment data
    - date of enrollment with health insurer (and history of enrollment)
    - employer information
- demographic data
    - age, sex, geography
- risk assessment data

    - health information collected from member individually through questionnaires and employer programs
- participation in programs/case management
    - information about the member collected as a result of their interaction directly with the plan on health – related issues
- service interaction
    - information collected about a member as a result of an inbound or outbound communication with the insurer

In the last few decades, the decreasing cost of digital storage and the development of management information systems has enabled healthcare organizations to collect and store significant amounts of data about their customers. Unfortunately, in many cases, this data is not organized in a way that it can be consolidated and leveraged for analytical purposes. Additionally, most of the data collected does not yield insights that can be utilized alone for predicting risk of **unknown populations**.

The vast majority of data mining and predictive modeling that occurs within the health insurance environment today leverages only what the health insurer knows of its member base. Very little augmentation of data has occurred. Furthermore, key analytical leaders within these organizations have primarily leveraged analytics for the purpose of traditional underwriting and actuarial services or for predicting utilization and trend patterns for other clinical purposes.

Risk prediction models typically rely on past claims data. However, for newly insured individuals or groups and prospective consumers, absence of past claims data makes it difficult to predict future costs. Therefore, innovative methods are starting to be used to predict health risk. For example, non-traditional or external data such as life-style factors are being used for modeling.

STEP 3: UNDERSTAND AND STRATIFY RISK OF EXISTING MEMBERS USING CLAIMS DATA

Once risk is defined and baseline data and models for the purposes of this project are understood, the next step in the process is to understand and categorize the level of risk of individual members based on historical, actual claims data. The purpose of this step in the process is to develop a workable framework and understanding of levels of medical expense and correlate those levels of expense to risk.

For this analysis, a health insurance partner provided detailed claims data on several million members over a period of three years. The data included:

- dental claims
- inpatient claims
- outpatient claims
- membership information
- disease registry

STEP 4: APPLY CONSUMER DATA TO RISK STRATIFIED MEMBERS AND BUILD/TEST MODELS

After stratifying the membership file, work began to append the membership data with rich consumer data and attempt to draw correlations between consumer data variables and high or low medical expense or risk.

The consumer data leveraged for this project was a commonly available data set from a well-known commodity data provider. It includes a set of approximately 1,500 variables, organized at the household level and includes data about all known individuals residing within the household.

The majority of the analytical work took place during this phase of the project, and the remainder of this paper focuses on the technical aspects of this work as well as the results and future business applications of our findings.

METHODOLOGY

Both supervised and unsupervised data mining methods are considered for modeling. At a very high level, three different approaches that are considered for model development are described below.

ANALYTICS, TECHNIQUES, AND TECHNOLOGY

Developing the analytical models involved a series of steps from data extraction, data integration, data transformation, and modeling. A great deal of time was spent in preparing an Analytical Base Table (ABT). An ABT is a structure of data suitable for predictive modeling. Creating an ABT involves extracting data from various source systems (or tables) and consolidating the tables based on common identifiers.

SAS has recently introduced a new solution, SAS® Membership Portfolio Optimization. This is the technology utilized for data extraction, transformation, and loading. It is used to profile, cleanse, augment, and monitor data in order to create consistent and reliable information. SAS® Membership Portfolio Optimization also provides a number of transformations and functions that improved the quality of the data.

Once the data was available as an ABT, advanced predictive models were built using a powerful graphical interface that simplified many of the common tasks associated with our data mining, including text mining. This allowed for more time and resources to be devoted to model design and testing.

Supervised Modeling

This approach involved applying supervised techniques such as decision trees, regression, and neural networks to the complete data set. A risk prediction model is a typical example of a supervised model where there is a target variable to predict using a set of input variables. The target variable in this setting can either be total annual cost or a classification variable such as high, medium, or low risk that is derived from total, annual cost.

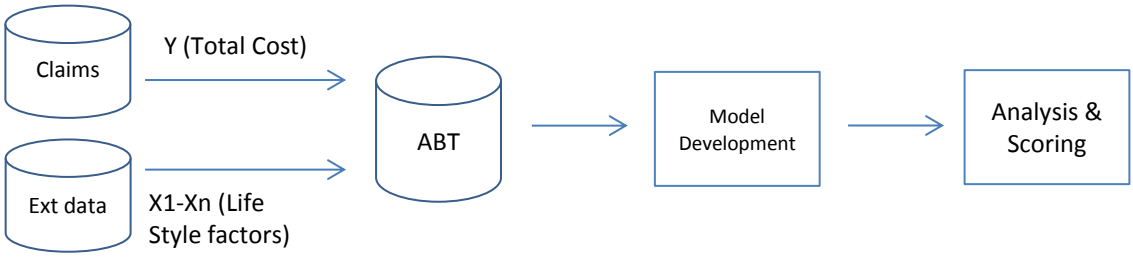


Figure 1. Supervised Modeling Diagram

Hybrid Modeling

In this approach, diverse groups of individuals were first identified by applying unsupervised data mining techniques such as cluster analysis on external data. Different types of predictive models were then built separately for each group. Groups were validated and selected for model development based on total annual costs and those that made good business sense.

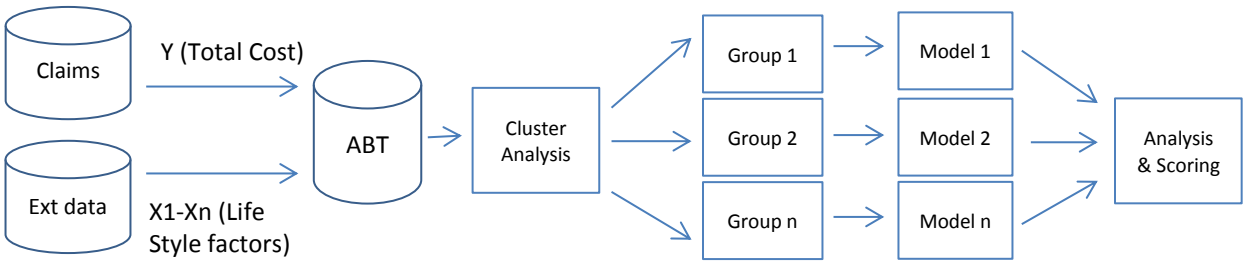


Figure 2. Hybrid Modeling Diagram

Two-Stage Modeling

When the objective is to predict costs, a two-stage modeling approach can be used to do the following tasks:

1. predict if somebody is going to experience any health illness (1/0 prediction)
2. predict costs only for those members who are predicted as positive in the first stage

This type of approach is best suited to situations where you have significant number of members in your input data with zero health care costs.

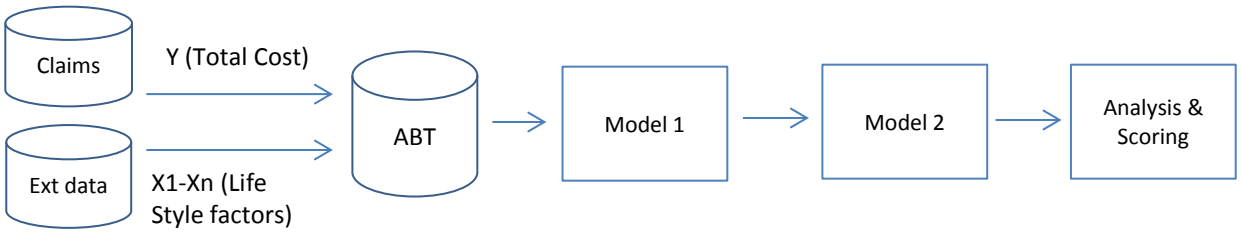


Figure 3. Two-Stage Modeling Diagram

DATA

Individuals enrolled for the entire 12 month period of 2011 are included for model development. The total health care cost of these individuals in 2011 is considered for creating the target variable. External data is assumed to represent these same members' interests and behavior in 2010. This information is used to predict health care cost in 2011. Around 1,500 external data elements were used for this analysis. These input elements can be categorized as following:

- demographics (age, gender, race, etc.)
- household information (household size, children, length of residence, etc.)
- financial stability (net worth, rent/own a home, invest in stocks, etc.)
- behavioral (book reader, travel interests, hobbies, community involvement, etc.)
- propensity (media channel, vehicle brands, telecommunication trends, etc.)
- census

Using the health plan's detailed claims data, we calculated the following information:

- average number of claims
- average number of claims for different claims types (inpatient, outpatient, professional, and office)
- average costs for different claims types (2 months, quarterly, six months) for different age and gender groups at a postal code level

Below are the additional variables that were considered as inputs for modeling:

- dental claims
- inpatient claims
- outpatient claims
- total amount from ancillary claims
- total amount from inpatient claims
- total amount from office professional claims
- total amount from outpatient professional claims

MODELS AND RESULTS

Model development involves a series of tasks that include sampling, exploration, transformation, variable selection, modeling, scoring, and reporting. Extensive data preparation tasks such as creating new variables, imputation, and replacement were performed before trying modeling techniques. Various models were built following the above discussed methodologies.

Models

A large pool of models is essential to attain the best model. The most exciting aspect of predictive modeling is trying a variety of options during model development. Approaches can include; using a subset of observations, using a subset of input variables, or using different forms of target variables. All of these models should be evaluated against a baseline model or against each other, using a standard set of metrics.

Numerical vs. Categorical Target

The initial approach taken was to deploy a numerical target variable to predict exact health care costs in dollars, using consumer data. This approach is extremely challenging, primarily due to an extremely skewed distribution of the target. It was noted that performance could be slightly improved when using a classification model where individuals are classified into risk categories such as high, medium, and low. It was very useful to try both of these models and validate the inputs that are selected. These two types of models can be compared by classifying the predictions from a numerical target model into categories and validating against the results from a categorical target model.

Target Cut-Off

In the case of classification models, identifying the right cut-off points for “high and low” or “high, medium, and low” from cost information is a sensitive issue. Many marketing and product pricing strategies rely heavily on this classification. Therefore, ensuring that the cut-off points are representative of the industry business case is critical. From a data point of view, the distribution of the cost variable plays a vital role. For this project, it is important that the variable be split in such a way that we had sufficient data points for each category for modeling. Using the cost distribution, the following cut-off points were tested:

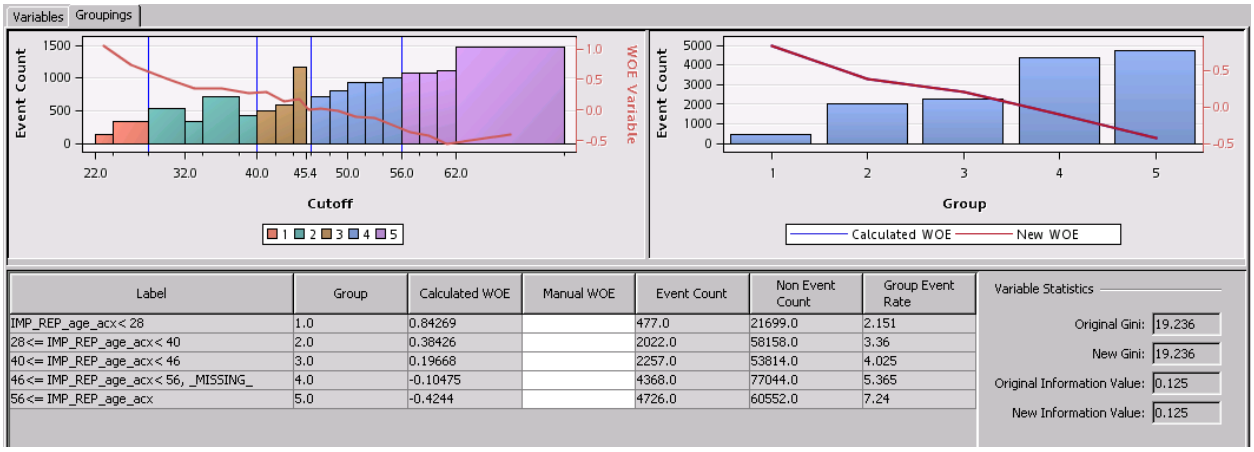
- top 5% as high risk vs. rest 95% as low risk
- top 10% as high risk vs. rest 90% as low risk
- top 10% as high, 50-90% as medium and below 50% as low risk
- top 10% as high and below 50% as low risk. The medium risk population is excluded.

Target Transformation

Usually, cost variables are highly skewed. For a better and more stable model, the target variable needs to be transformed using many available transformation techniques. However, in the case of non-linear transformations (like log transformations, square root transformations, power transformations, and so on) it is very difficult to interpret the results. Log transformation yielded better performance than other techniques. Other methods that were used to tackle skewness are; winsorizing or ordinally-stacking values, exceeding a certain threshold.

Age Groups

Age plays a very significant role in most of the models. Since consumer characteristics significantly differ based on age, it makes good business sense to develop models for various age groups. Figure 4 shows a sample age group classification using weight of evidence (WOE) approach. Models were built separately for each segment and the predictions from all the individual models were combined to measure the overall performance.



Label	Group	Calculated WOE	Manual WOE	Event Count	Non Event Count	Group Event Rate
IMP_REP_age_acx< 28	1.0	0.84269		477.0	21699.0	2.151
28<= IMP_REP_age_acx< 40	2.0	0.38426		2022.0	58158.0	3.36
40<= IMP_REP_age_acx< 46	3.0	0.19668		2257.0	53814.0	4.025
46<= IMP_REP_age_acx< 56, _MISSING_	4.0	-0.10475		4368.0	77044.0	5.365
56<= IMP_REP_age_acx	5.0	-0.4244		4726.0	60552.0	7.24

Variable Statistics

Original Gini:	19.236
New Gini:	19.236
Original Information Value:	0.125
New Information Value:	0.125

Figure 4. Age Groups Identified using WOE

Missing Data

It is quite common to find data missing in any data set. In fact, a data set is never said to be perfect without missing data. In the case of consumer data, it is even more common. Many modeling techniques do not consider records with missing data. When the proportion of observations with missing data increases, it poses a serious problem for the model building process. This situation was encountered in this project. It was decided to make the best use of the

available data by following a segmented or stratified approach to modeling, based on the availability of data. As an example, the individual consumer data was segmented into several categories:

1. individuals with complete (100% of data columns)
2. individuals with most of the data (i.e., an above average amount)
3. individuals with “a little” data (i.e., a below average amount)

Interestingly, it was observed that some of the models that were developed using “a little” data performed well compared to the models with complete data. This was due to some of the strong predictor variables that had a decent amount of missing data. Building models using these criteria helped the scoring of individuals for our project and will be imperative for health insurers wishing to use this same approach to score new individuals appropriately in the presence of missing data.

Results

For the purposes of this paper, only results from one of the best performing classification models are included. In this case, the target variable is binary with a value of 1 representing “high-cost” individuals and value 0 representing “low-cost” individuals. High-cost individuals are the top 5% (most expensive) of individuals with annual costs exceeding approximately $20,000. Input data was partitioned into two groups: 60% for training and 40% for validation. Some of the predictor variables that show value in predicting individual level risk are:

- age of the Individual
- gender
- frequency of purchase of general apparel
- total amount from inpatient claims
- consumer prominence indicator
- primetime television usage
- smoking
- propensity to buy general merchandise
- ethnicity
- geography – district and region
- mail order buyer - female apparel
- mail order buyer - sports goods

This list of variables is illustrative and varies with both the studied population and the consumer data available for modeling.

The neural network model outperformed other models with an ROC index value of 0.65 for the model. Although this is not a significant achievement, a great deal of individual-level insight can be gained by leveraging this model. Table 1 displays classification statistics for the neural network model on both validation and training data sets. The validation data contains approximately 100,000 individuals in the age group 22-65. Model stability is confirmed because the metrics in both the data sets are very close. Sensitivity is the true positive rate and 1-Specificity is the false positive rate. By targeting top decile (approximately 10,000 individuals) in the validation data, 19.6% of the high cost members (sensitivity) are correctly captured.

	Validation Data			Training Data		
Statistic / Top Percentile	**5%**	**10%**	**15%**	**5%**	**10%**	**15%**
Sensitivity	11.9%	19.6%	27.4%	11.1%	20.2%	28.1%
1- Specificity	4.6%	9.5%	14.4%	4.7%	9.5%	14.3%
+ Predictive	10.9%	9.8%	9.1%	11.0%	10.0%	9.4%

Table 1. Statistics for Top Percentiles for Consumer Data Only Model

For the same validation data, a baseline model with age and gender as the input variables performed as shown in Table 2.

Statistic / Top Percentile	5%	10%	15%
Sensitivity	7.7%	15.8%	23.6%
1- Specificity	4.8%	9.7%	14.5%
+ Predictive	7.7%	7.9%	7.9%

Table 2. Statistics for Top Percentiles for Age/Gender Model

Positive improvement in the model sensitivity measure for all cutoff points is noted. In addition, about 7% improvement in ROC metric in validation data is also seen.

Performing cluster analysis promised to reveal a wealth of information and is accomplished by splitting individuals into meaningful groups such that each group of customers is collectively different from the customers in the other groups. Many methods of cluster analysis can be performed using SAS® Enterprise Miner. In this project, as discussed earlier in the hybrid modeleing approach, cluster analysis was performed before the modeling task and models were built at cluster level.

The popular K-Means cluster analysis technique with average method for selecting the initial seeds was performed. The input variables were divided into bases (used for clustering) and descriptors (used for profiling clusters). The bases include propensity and health related behavior variables. The descriptors include demographic and census variables. This method gives a six cluster solution.

Health care cost and medical condition information were used to understand and profile the risk severity for each group. In addition, models were built at the cluster level for better predictions. Figure 5 shows the six-cluster soution obtained in this analysis and cluster proximities in two-dimensinal space. The cluster sizes are reasonable for modeling. However, several clusters were combined, based on risk severity and cluster proximity for better model performance.

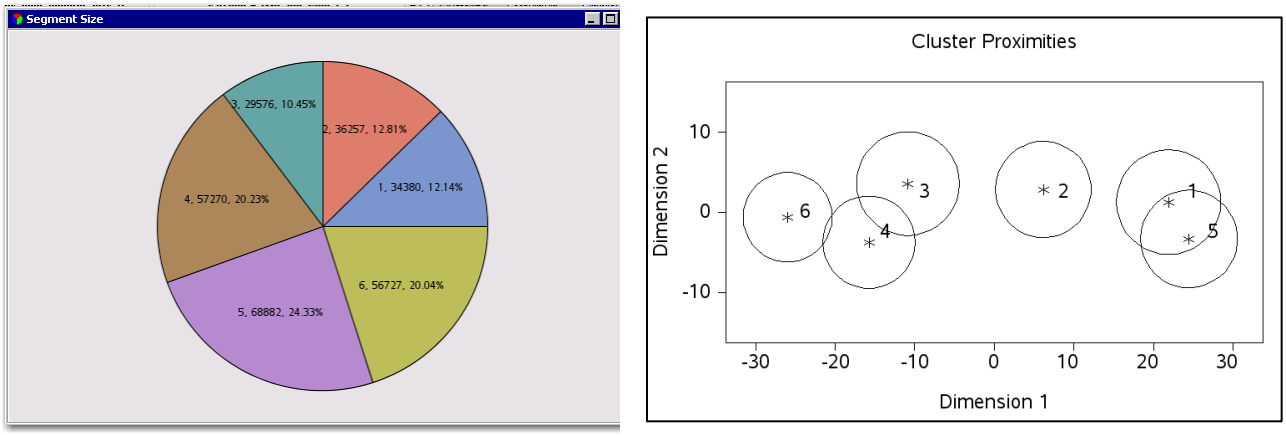


Figure 5. Cluster Results

Table 3 shows profiles of example clusters identied in this exercise,

Cluster Name	Characteristics
Single Adults	low economic stability, low cellphone use, under banked, low investors, have kids, no phone or mail buying, higher renters, single
Health Conscious Adults	high economic stability, own a premium card, religious, involved in kid activities, high investment activity, mail-order prescriptions, high cellular user
High Net Worth Adults	high net worth, heavy readers of newspaper, followers of outdoor advertising, most likely to buy prescriptions by mail, savvy investors, heavily involve in charity activities, mostly adult male between 45 -64

Table 3. Sample Cluster Profiles

CHALLENGES

The results of the project are significant, but there were also significant challenges throughout the process. Working with consumer data is a grueling task; a task made further daunting because consumer data has not been historically organized or designed to support solving health insurance business problems. Additionally, consumer data is collected from various external sources. As a result, there are a wide variety of issues related to data integrity, data quality, missing data, time frame, and so forth. Below are some of the key challenges that were encountered during this project.

Large Set of Inputs Results in Redundancy

Much of a predictive model performance depends on input variables in the model. In the context of consumer data, thousands of various data elements are collected about individuals and from a wide variety of sources and reporting mechanisms. When data is collected from numerous sources, there is high chance for redundancy. Redundant inputs complicate model construction and lead to model instability. Another issue with redundancy is the challenge of selecting which redundant variable you chose for model development. This leads to the question: “Which version of the variable is most relevant to our specific business problem?”

A variety of techniques exist to reduce the number of variables. Some of the supervised modeling techniques such as linear regression using R-Squares, stepwise regression, and other variants of regression are commonly used for variable reduction. One of the most commonly used, unsupervised statistical approaches is principal component analysis (PCA). PCA tries to identify a series of orthogonal vectors that best describe the direction of variation in the data. The goal is to identify a small set of components that characterize most of the variation. While variable redundancy is addressed using PCA, it is highly difficult to interpret the resulting components, thereby making it difficult to interpret the models. Variable clustering is another approach and is similar to PCA in that it can address both input redundancy and variable relevance.

Variable clustering algorithm (PROC VARCLUS in SAS/STAT®) provides an iterative variable clustering approach that uses oblique rotation of principal components. The procedure groups correlated inputs that have correlated as possible among themselves and uncorrelated with inputs from other clusters. A representative input for each cluster is identified. The advantage of this technique over PCA is that reduced sets of inputs are actual variables rather than transformations. In addition, cluster components, which are similar to PCA components, can be used for each cluster. Cluster components can be used; rather than using input variables as representative variables.

In this project, variable clustering is heavily used to exclude correlated variables. This is mainly used to reduce the number of census variables where you encounter many similar variables. Variable clustering does not consider the target variable when identifying the representative variable. Therefore, we accompany variable clustering with correlation analysis of input variables with the target variable. We hand-picked some of the variables rejected by the variable clustering exercise that had significant correlation with the target and dropped the least correlated variables. In-depth analysis on each cluster was performed to see the business value of the variable identified along with the statistical significance in the model. Considerable focus is placed on identifying and including variables that would be correlated to health and financial behavior. Table 4 lists one of the clusters of variables identified in the consumer data. The variable clustering result selected “Economic Stability” as the representative variable. However, when calculating the correlation of the inputs with a target variable and then looking at the business value, the variable “Propensity to buy prescriptions via mail” seems to be relevant for this analysis.

Cluster of similar variables	Cluster R-Square	Variable Clustering	Manual Selection
Affordability	0.83	No	No
Economic Stability	0.87	Yes	No
Invests in stocks and bonds	0.72	No	No
Propensity to buy via mail	0.62	No	No
Propensity to buy via phone	0.73	No	No
Propensity to buy prescriptions via mail	0.55	No	Yes
Propensity to go on cruise vacation	0.63	No	No
Read newspapers	0.48	No	No

Table 4. Example of Variable Clustering Results

A significant amount of time and effort has been invested in analyzing each cluster composition and identifying the right variables for model development. Among various other input reduction techniques, this technique is very useful in working with consumer data. For more technical details and working information on this technique, refer to the 2008 SAS Global Forum paper, "Two-stage variable clustering for large datasets", by Taiyeong Lee, David Duling, Song Liu, and Dominique Latour.

High Dimensional Variables

Data mining has always suffered from the curse of dimensionality. Having more input variables does not always help in modeling. It becomes more complicated with categorical variables that contain many levels. A categorical variable with K levels means having k-1 continuous variables in the model. Many variables in consumer data are categorical variables with numerous levels. For example, data on geography, ethnicity, occupation, and zip code can be captured at a very detailed level that can generate hundreds of levels for each variable. These variables are difficult to manage and significantly impact model performance. It is critical to explore ways to reduce the number of levels either by binning the levels, replacing a level with a higher level in the hierarchy, or excluding rarely occurring levels. For example, the number of levels in occupation or income ranges can be binned and zip codes can be combined, by county or state.

In addition to these approaches, advanced grouping techniques such as weight-of-evidence coding were utilized for dimension reduction. Effectively, a categorical variable is transformed into a numerical variable so that it can be used as a continuous variable in the model.

The weight-of-evidence technique has been modified via Bayesian statistical methods which assume a priori distributions for the target average within each level. These prior distributions reflect an analyst's state of knowledge about the expected value of the target with each level of the input variable. Observations from the training data are then combined with prior observations to form updated estimates of the target average distribution. The expected value of this posteriori distribution serves as the estimated target value within each level of the input variable. These posterior estimates generally show substantial reduction in the prediction bias compared to the basic weight of evidence approach. This is especially true for non-numeric inputs with tens of hundreds of levels. This technique is known as smoothed-weight-of-evidence.

If the target is binary, smoothed-weight-of-evidence (SWOE) for level *i* is calculated as,

$$SWOE_1 = \log\left(\frac{n_{1i} + rho_1 * smooth}{n_{0i} + rho_0 * smooth}\right)$$

Equation 1.

Where, n_{1i} is the number of target=1 for level *i*, n_{0i} is the number of target=0 in level *i*, rho_1 is the proportion of target=1 in the training data set, $rho_0 = 1 - rho_1$, and $smooth$ is the value of the smoothing parameter.

If the target is interval, a Bayesian-inspired estimate for the target mean is calculated in place of the smoothed-weight-of-evidence. This estimate can be thought of as a weighted average of the overall target mean and the observed target mean in level i and is given as,

$$SMOOTHMEAN = \frac{smooth * \hat{Y} + n_i * \hat{Y}_i}{smooth + n_i}$$

Equation 2.

Where, $\hat{Y}$ is the overall target mean, $\hat{Y}_i$ is the level *i* target mean, n_i is the number of cases in level *i*, and smooth is the value of the smoothing parameter.

In this project, this technique was relied heavily on to transform categorical variables into numerical inputs. Also observed are recoded variables for zip code, ethnic code, and congressional district selected in the model which, were otherwise never picked by the model.

Missing Data

Surveys, campaigns and publicly available financial and purchasing data are the primary tools used to collect information on consumer characteristics. It is highly impossible to collect every piece of information for the whole population. The more the types of characteristics collected, the more the sparse data. Out of the approximately 1,500 variables used in analysis, 50% of them were present with more than 90% missing values. A variety of techniques such as mean and mode imputation, tree imputation and distribution were tested. Innovative data imputation techniques such as imputing values at each geographical level were tried for better accuracy. Determining the threshold for excluding certain variables for inclusion in the model building process is an important and critical decision-making process.

Imputed Data

Data imputed by the vendor can be a majority of the records in some columns. Using this high volume of imputed data in a model means that you are building models on models. A change in the vendor's imputation model may impact future results.

Zip Code Level Data

What is the value of zip code level data? Why not make the zip code itself a nominal variable? In terms of the target variable, zip code level data has only the value of the zip code itself. However, it consumes more memory and computing power. A record is no more unique when adding zip code level data to the zip code itself. Therefore, why is it used?

Consider zip code level census data. Replacing zip codes with the census data has two major benefits:

1. no issue with zip codes containing too few records
2. regional knowledge can be applied to new areas

The key value here is extrapolation. For example, if a model was built on zip codes in California, you would know nothing about how to score the model with Virginia zip codes. Census variables are uniform across the many states. The California model built on census variables instead of zip codes could be used to score data from Virginia.

If you have very few members in rural zip codes you may be over-fitting when using zip codes. In this case, replacing zip codes with zip code level census data may increase model accuracy. As with the previous example, this can be thought of as extrapolating, except applied to a location that has little data versus no data.

Claims Cost Distribution and Zero Claims Data

Modeling the numerical claims cost of individuals is made more problematic by individuals with no claims. The distribution of medical costs for individuals with claims is roughly lognormal. However, it is very important to predict the risk of all individuals in a population and not just for those with claims. Figure 6 shows a histogram of the log of claims costs. One dollar was added to the claim records so that the log could be taken.

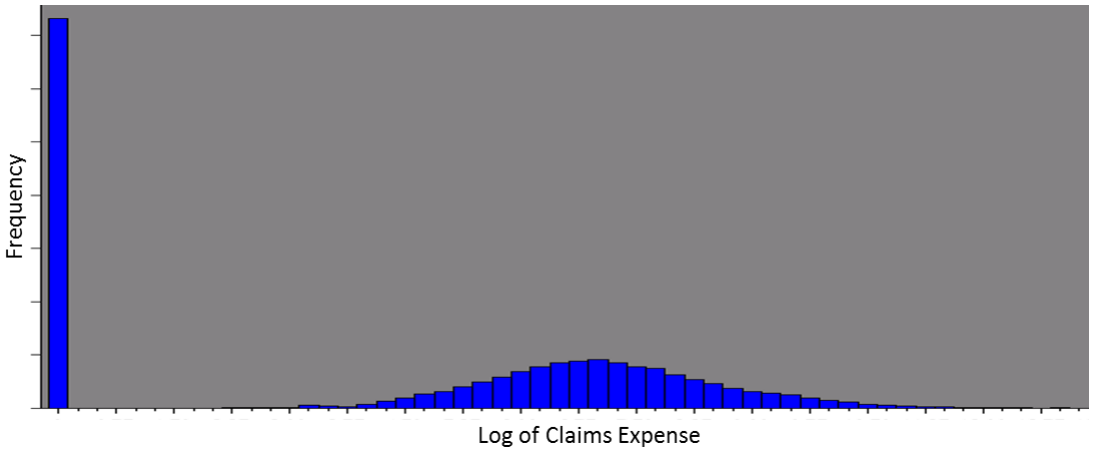


Figure 6. Histogram of the Log of Claims Expense

Many models work better when the data that is being modeled is normal in distribution. As seen in figure 6, the tall bar representing individuals with zero claims records destroys the normality of the data. Winsorizing claims data can produce a similar effect on the right side of the histogram.

In the Analytics, Techniques, and Technology section of this paper it is noted that a two-stage model can be helpful in this situation. One model can be used to predict whether an individual would have any costs. A second model is then used to predict the level of the claims costs, if any. This approach is less affective for the problem of winsorized claims data.

Using a categorical target instead of a numeric interval target is an additional approach to this problem. However, converting to a categorical target, does by its nature, lose detail of the data. This balance should be considered when choosing a modeling approach.

Time Frame

Consumer data is collected over a period of time. The life-style of an individual can be expected to remain unchanged over a period time. However this depends entirely on numerous socio-economic factors. It is very essential to identify constant and time-varying factors in the consumer data before model development.

Individual data is the most granular. However, for some variables it does not add more value. Consider a variable such as the number of children in a household. Clearly, if the data is good, this value would be the same for all living at an address. Having this variable at the individual level would add no additional value.

CONCLUSION

There were two goals for this project:

1. demonstrate the value of applying advanced analytical methodologies to our health plan partner's consumer and traditional data to better predict health care costs of existing members
2. determine whether this model can be used effectively to predict cost or classify new individuals without having access to historical claims and utilization data

Based on the results of this project, it is clear that publicly available consumer data can indeed be leveraged to improve the prediction of high, medium, and low health care utilization and cost. The degree to which this data is predictive and useful depends on the quality and prevalence of the data as well as the type of statistical models applied to the problem. Consequently, there are significant challenges to overcome along the way.

As would be expected, even the best performing models generated from third party data do not yield a tremendous lift in predictability of risk. However, the lift is sufficient to provide a competitive advantage for those health plans that choose to deploy this data and advanced analytical models to the problem. Additionally, as this data and these types of models are adapted within the industry and leveraged for consumer marketing and engagement, the power and predictability of the models will likely improve in a dramatic and rapid way.

Lifestyle factors such as smoking, drinking, and regular exercise are used by insurance companies and care providers to understand the health profile of their customers. However, a much wider set of behavioral attributes such as online and offline purchasing frequency, media preferences, and affinity measures, help to identify proxy risk factors. This external information can also be used for existing members along with the medical condition data to design better care management and wellness programs.

As previously discussed, health insurers face tremendous challenges in the new marketplace. Understanding, predicting, and developing strategies to address consumer risk and behavior will become critical, competitive differentiators for those who choose to pursue these capabilities.

While in the very early stages of leveraging consumer and external data to generate risk and behavioral insights, business applications are vast and transformational for this industry. It is estimated that by operationalizing the models built with this consumer data (for marketing and consumer engagement purposes), the insurer referenced throughout this paper could experience medical savings of anywhere between $4 and $8 per member per month. To put this into perspective, for a health plan with one million members, this could result in anywhere from $48 million to $96 million in medical expense savings annually.

Obviously, these types of potential, financial results warrant continued exploration and capability development. The real opportunity for long-term business transformation and financial return will come over time. With a mature consumer-level risk and behavior prediction capability, health insurers will have the ability to:

- optimize portfolios of individual and group members – for profitability or health manageability – across a wide variety of geographies and product suites
- maximize revenue from government reimbursement programs by predicting health of members better than the competition
- develop and deploy disease management and prevention programs to the right member at the right time – to improve health of populations
- develop the right service strategies and product offers to keep members happy and satisfied with their coverage; thereby retaining more members than the competition
- provide point of service predictive information to providers to help them deliver the best care option to at-risk members
- ultimately, remove millions – if not billions – of dollars in unnecessary medical expense from the system by better engaging and managing the care of individual consumers of health care

Clearly, there is a long and arduous road ahead for US health insurers as they attempt to transform their business model to focus on and cater to individual consumers. The first step in this process is being taken now, with insurers attempting to understand and predict behavior using publicly available consumer data. There is tremendous opportunity to leverage new and emerging data sources and elements to enhance this understanding and prediction. Social media data, electronic medical records, remote health monitoring devices, retail health insurance marketplaces, health insurance exchanges, and so many more sources of consumer health data can be used.

REFERENCES

Duncan, Ian. 2011. *Healthcare Risk Adjustment and Predictive Modeling*. Winsted, CT. ACTEX Publications, Inc.

Weilenga, Doug. (2007). Identifying and Overcoming Common Data Mining Mistakes. SAS Global Forum 2007.

Cathie, A., (2011). The Anti-Curse: Creating an Extension to SAS® Enterprise Miner™ Using PROC ARBORETUM. SAS Global Forum 2010.

Lee, Taiyeong, Duling, David and Latour, Dominique.(2008). Two-stage clustering for large data sets. SAS Global Forum 2008.

American Medical Association. 2009. An Introduction to risk assessment and risk adjustment models. Available at: http://www.ama-assn.org/resources/doc/psa/risk-assessment.pdf

Gay, Barry. *Better care, better system: The role of advanced analytics in improving efficiency and patient outcomes*. SAS Institute Inc. 2011. Available at: www.sas.com/healthinsights

Leigh JP, Hubert HB, Romano PS. 2005. Lifestyle risk factors predict healthcare costs in an aging cohort. Am J Prev Med. 2005; 29:379–387

Koh, Hian Chye and Tan, Gerald. 2005. Data Mining Applications in Healthcare. Journal of healthcare information management, Vol. 19, Issue 2, Pages 64-72.

Srinivas, K., B.K. Rani, and A. Govrdhan, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering (IJCSE), 2010. Vol. 02, No. 02: p. 250-255.

Advanced Predictive Modeling Using SAS® Enterprise Miner™ 6.1 Course Notes developed by Jim Georges, and revised by Dan Kelly and Bob Lucas. SAS Institute Inc.

RECOMMENDED READING

- *Predictive Modeling With SAS® Enterprise Miner: Practical Solutions for Business Applications*
- *Customer Segmentation and Clustering Using SAS® Enterprise Miner, Second Edition*
- *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management, Second Edition*