Article from

Product Matters

June 2018
Issue 110

Predictive Modeling for Life Insurance

Ways Life Insurers Can Participate in the Business Analytics Revolution

By Chris Stehno and Jim Guszcza
Contributors: Mike Batty, Arun Tripathi, Alice Kroll, Cheng-sheng Peter Wu, David Moore, and Mitch Katcher

Editor's Note: This article is part two of two. Part one was published in the February 2018 issue. It has been formatted for the newsletter and is reprinted with permission from Deloitte Development LLC.

BUILDING A PREDICTIVE MODEL

After discussing so much about what can be done with predictive models in life insurance, we have finally come to how to build one. The following section describes the technical process of developing a model.

Data

Predictive modeling is essentially an exercise in empirical data analysis. Modelers search through mountains of data for repeatable, statistically significant relationships with the target (underwriting decision in this case), and generate the algorithm that produces the best fit. Since it is central to the modeling process, the best place to begin the technical discussion is with the data.

Data miners prefer to start with a wide lens and filter out potential data sources as necessary. We start by asking, "What data can be obtained for an individual applicant?" And then move to questions such as, "Which data elements show a relationship with the target?," "Is the penetration of the data enough to generate statistical significance," "Is the correlation strong enough to justify the data's cost?," and finally, "Based upon regulatory and compliance concerns, can the data be included in a predictive model in the underwriting process?" In our experience, working through these questions leads to two different classes of data: a sub-selection of traditional underwriting requirements, and alternative datasets not traditionally used in underwriting assessments.

The traditional underwriting requirements incorporated into the predictive models generally meet several criteria:

- Available within the first one to two days after an application is submitted Transmitted electronically in a machine readable format
- Are typically ordered for all medically underwritten applicants
- Several of the most common data sources are discussed below. The actual sources used by any particular life insurer may vary.

Application Data (including part 2 or tele-interview)—any piece of data submitted to the company by an insurance applicant is a candidate for the predictive model. There are two keys to successfully using the data contained in an insurance application in a model. First, the questions which are easiest to work with are in a format such as multiple choice, Yes/No, or numerical. However, new text mining applications are making free form text possible in some situations. Second, the new business process should capture the application electronically and store the answers in a machine readable format such as a database. Life insurers who do not have application data in a compatible format face considerable manual data entry during model build.

MIB—When member companies receive an application, they will request a report from the Medical Information Bureau (MIB). This report includes MIB codes which provide detail on prior insurance applications submitted to other member companies by the person in question.

MVR—The Motor Vehicle Record (MVR) provides a history of driving criticisms, if any, for a given applicant. This inexpensive and readily available data source provides important information on the applicant's risk profile otherwise unavailable in third-party marketing data. Due to its protective value, it is also a common underwriting requirement for many life insurers.

Electronic Rx Profile—in recent years, several firms have started collecting prescription data records from pharmacy benefit managers nationwide, compiling by individual, and selling this information to insurers. Many users are enthusiastic about its protective value, and as a result it is becoming a standard underwriting requirement for an increasing number of life insurance companies. This is another interesting source for predictive modeling.

Other traditional underwriting requirements, such as blood and urine analysis, EKG's, medical records and exam, etc., would add predictive power to a model, but the time and cost to include them may negate the benefits.

Non-traditional third-party data sets come in a variety of shapes and forms, but most recently we have seen the application of marketing and consumer credit data from companies such as Equifax and Axiom. It is important to distinguish between marketing data and the credit score information for which these consumer reporting agencies are better known. Beyond the credit data, these firms also collect and distribute consumer information for marketing purposes. Whenever you use your credit card to make a purchase, or provide your phone number or zip code to the cashier, this data is being collected, aggregated, and resold.

The third party marketing dataset obtained from the consumer credit company contains thousands of fields of data. In contrast to the credit score data, is not subject to the Fair Credit Reporting Act (FCRA) requirements, and does not require signature authority by the insurance applicant to use it in a model. For the purposes of constructing a model, the data can be returned without personally identifiable information. Our experience indicates that using an individual's name and address, the typical match rate for members of these databases is over 95 percent.

We understand if some people react to this with a feeling of someone looking over your shoulder, and we discuss some of the ethical concerns of using this data in a later section of this article. Here we will simply say that while many of these data fields are quite interesting for life underwriting, it is important to note that model scores are not highly dependent upon any one, or even handful of them. Instead, the picture painted by this data is viewed holistically, trends are identified that are not necessarily noticeable to the naked eye, and the overall messages about lifestyle and mortality risk are communicated. For this reason, it is difficult, if not impossible, to send a powerful message that misrepresents the applicant, or for the applicant to manipulate the data in a misleading fashion.

Modeling Process

The first step in the model building process is to collect and organize all this data. For several reasons, it is collected for applications received by the insurer over the past 12 to 18 months. Depending upon the volume of applications received, this time frame typically produces a sample of underwriting decisions which will be large enough to sufficiently remove the statistical variation in the model, and ensure the third-party data available is still relevant. To clarify, the external data technically reflects the applicant's lifestyle today, but is still an accurate representation of them when they applied for insurance provided that time was in the recent past. Based on our experience, 18 months is about when you may begin to see material changes in the modeling data, and thus question its applicability to the application date.

The actual collection of the third-party marketing data set for model building is typically a painless process facilitated by the provider, but the availability of internal historical underwriting data can vary greatly depending upon individual company practices.

Once the data is collected into one centralized data set and loaded into the statistical package in which the analysis will be performed, data preparation will provide a solid foundation for model development. Data preparation can be summarized into four steps which are described below:

1. Variable Generation
2. Exploratory Data Analysis
3. Variable Transformation
4. Partitioning Model Set for Model Build

Variable Generation

Variable generation is the process of creating variables from the raw data. Every field of data loaded into the system, including the target and predictive variables, is assigned a name and a data format. At times this is a trivial process of mapping one input data field to one variable with a descriptive variable name. However, this step can require more thought to build the most effective predictive models. Individual data fields can be combined in ways that communicate more information than the fields do on their own.

These synthetic variables, as they are called, vary greatly in complexity. Simple examples include combining height and weight to calculate BMI, or home and work address to calculate distance. However, in our experience some of the most informative predictive variables for life insurance underwriting are what we call disease-state models. These are essentially embedded predictive models which quantify the likelihood an individual is afflicted with a particular disease such as diabetes, cardiovascular, or cancer. The results of these models can then be used as independent predictive variables in the overall underwriting model. Synthetic variables are where the science and art of predictive modeling come together. There are well-defined processes which measure the correlations of predictive variables with a target, but knowing which variables to start from relies more on experience and intuition.

Exploratory Data Analysis

Before even considering the relationship between independent and dependent variables, it is first important to become comfortable with the contents of the modeling data by analyzing the distributional properties of each variable. Descriptive statistics such as min, max, mean, median, mode, and frequency provide useful insight. This process tells modelers what they have to work with, and informs them of any data issues they must address before proceeding.

After the initial distributional analysis, the univariate (one variable at a time) review is extended to examine relationship with the target variable. One-by-one, the correlation between predictive and target variable is calculated to preview of each variable's predictive power. The variables that stand out in this process will be highly correlated with the target, well populated, sufficiently distributed, and thus are strong candidates to include in the final model.

In addition to paring down the list of potential predictive variables, the univariate analysis serves as a common sense check on the modeling process. Underwriters, actuaries, and modelers can sit down and discuss the list of variables which show strong relationships. In our experience, most of the variables that appear are those which underwriters will confirm are important in their processes. However, some other variables that are present can be a surprise. In these cases, further investigation into the possible explanations for the correlation is advisable.

Variable Transformation

The exploratory data analysis will most likely reveal some imperfections in the data which must be addressed before constructing the model. Data issues can be mitigated by several variable transformations:

1. Group excessive categorical values
2. Replace missing values
3. Cap extreme values or outliers
4. Capture trends

To increase the credibility of relationships in the data, it is often helpful to group the values of a given predictive variable into buckets. For example, few people in the modeling data are likely to have a salary of exactly $100,000, which means it is difficult to assign statistical significance to the likelihood an individual with that salary to be underwritten into a particular class. However, if people with salaries between $90,000 and $110,000 are viewed as a group, it becomes easier to make credible statements about the pattern of underwriting classes for those people together.

Missing values for different variables among the records in a data set is sometimes problematic. Unfortunately, there is no simple solution to retrieve the true distribution of variables that have missing values, but there are several approaches that help mitigate the problem. Modelers could remove all records in the data set which have missing values for certain variables, but this may be not an ideal solution because it can create a biased sample or remove useful information. A more common and effective solution is to replace the missing values with a neutral estimate or a best estimate. The neutral estimate could be a relatively straightforward metric such as the mean or median value for that variable, or a more in depth analysis of the best estimate could be the average value for that variable among other records that most similar to the one in question.

Almost all data sets a modeler encounters in real life will contain errors. A common manifestation of these errors is extreme values or outliers which distort the distribution of a variable. While not every outlier is a data error, modelers must weigh the risks and benefits of skewing the overall distribution to accommodate a very small number of what may or may not be realistic data points. Smoothing these extreme values may be a poor idea in applications such as risk management where the tail of the distribution is of utmost concern, but for underwriting predictive modeling it is often worthwhile to focus more on the center of the distribution. One approach to reducing the distortion is to transform a variable to a logarithmic scale. While extreme values will be muted, log transformation may minimize the original trend. Capping extreme values at the highest "reasonable" value is another simple alternative.

Finally, transforming variables from text categories to numerical scales can capture trends more readily. For example, BMI ranges have been officially classified into four categories: under-weight, normal, over- weight, and obese. Applicants with normal range of BMI are associated with a lower health risk than the members of the other categories. The trend of the BMI can be captured more effectively by transforming the BMI categories into an ordinal rank with higher numbers representing higher health risks, for example, 1=normal, 2=over-weight, 3=under-weight, and 4=obese.

Partitioning Model Set for Model Build

After collecting the data, preparing each variable, and casting aside those variables which will not be helpful in the model, the data set is divided into three approximately equal parts. Two of these, commonly called the "train" and "validation" sets, are for model building, while the "test" is placed aside until the end of the process where it will be used to assess the results [20].

After the data sets are partitioned, modelers carry out an iterative process that produces the strongest model. Most model builds will test a variety of statistical techniques, but often one effective, and therefore very common approach, is stepwise regression [21]. This is a fairly complicated process, but in essence, a best fit line that maps a set of predictive variables to the target is created. In a linear model, this best fit line will be of the form A * variable1 + B * variable2 + . . . = target variable. Variables are added and removed one-by-one, each time calculating the new best fit line, and comparing the fit of the new line with the fits of those created previously. This process reveals the marginal predictive power of each variable, and produces an equation with the most predictive power that relies upon the smallest number of predictive variables.

Each variable that survives the univariate review should be correlated with the target, but because it may also be correlated with other predictive variables, not every variable that appears strong on its own will add marginal value to the model. Among a group of highly correlated variables, stepwise regression will typically only keep the one or two with the strongest relationships to the target. Another approach for dealing with highly correlated variables is to conduct a principal components analysis. Similar to the disease-state models described above, a principal component is a type of sub-predictive model that identifies the combination of correlated variables which exhibits the strongest relationship with the target. For example, a principal components analysis of a group of financial variables may reveal that A * income + B * net worth + C * mortgage principal, and so forth, is a better predictor of underwriting decision than these variables are on their own. Then result of this equation will then be the input variable used in the stepwise regression.

The model is first built using the training data, but modelers are also concerned about fitting the model too closely to the idiosyncratic features of one sample of data. The initial model is adjusted using the validation data in order to make it more general. Each set is only used once in the modeling process. It cannot be recycled since the information has already become part of the model; and reusing it would result in over-fitting.

To assure the model does not reflect patterns in the modeling data which are not repeated in the hold- out sample, and most importantly, are less likely to be repeated in the applications the company will receive in the future, the test set is used only to assess the results when modeling is completed. This step protects predictive modeling from pitfalls like back-testing investment strategies. It is almost always possible to find a pattern in data looking backwards, but the key question is whether that pattern will continue in the future. Due to the relative efficiency of financial markets, investment strategies which looked so promising in the past usually evaporate in the future. However, in predictive modeling we generally find that the models built on the train and validation data set hold up quite well for the test data. The results shown in Figures 1 and 2 are representative of model fit on past test data sets.

At the end of this process modelers will have identified the equation of predictive variables that has the strongest statistical relationship with the target variable. A high score from this model implies the applicant is a good risk, and low score means the opposite. However, this is not the last step in model development. Layering key guidelines from the existing underwriting process on top of the algorithm is also a powerful tool. For example, certain serious but rare medical impairments may not occur in the data with the sufficient frequency to be included in a statistical model, but should not be overlooked by one either. For these conditions, it can be helpful to program specific rules that a life insurer uses to govern their underwriting. In addition to acting as a fail safe for rare medical conditions, the underwriting guidelines can also serve as the basis for making underwriting decisions. In the applications we have discussed thus far, the model has the authority to determine whether further underwriting is needed, but not to lower an insurance offer from the best underwriting class. Even for applicants where the model would recommend a lower underwriting class, incorporating the underwriting guidelines provides an easily justifiable reason for offering that class.

A final tool to extract useful information out of the modeling data is a decision tree [22]. A decision tree is a structure that divides a large heterogeneous data set into a series of small homogenous subsets by applying rules. Each group father along the branches of the tree will be more homogeneous than the one immediately preceding it. The purpose of the decision tree analysis is to determine a set of if-then logical conditions that improve underwriting classification. As a simple example, the process starts with all applicants, and then splits them based upon whether their BMIs are greater or lower than 30.

Presumably, applicant with BMI's lower than 30 would have been underwritten into a better class than those with higher BMIs. The stronger variables in the regression equation are good candidates for decision tree rules, but any of the data elements generated thus far, including predictive variables, the algorithm score itself, and programmed underwriting rules, can be used to segment the population in this manner. Figure 3 displays this logic graphically.

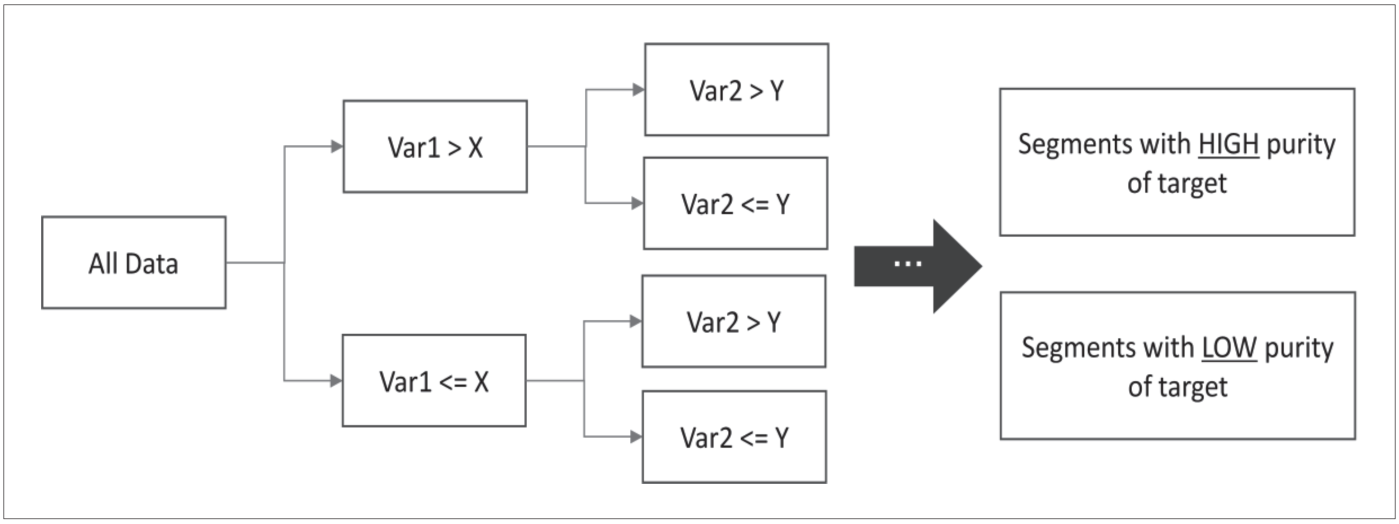
In principal, decision trees could be constructed manually, but in practice, dedicated software packages are much more efficient in identifying the data elements and values upon which to segment the population. These packages essentially take the brute force approach of trial and error, but due to computational efficiency they are able to develop optimal multi-node decision trees in manageable time.

Monitoring Results

In a previous section we discussed how to use the information revealed by predictive models to generate significant operational efficiencies in the underwriting process. From a technical standpoint, implementing a predictive modeling application can occur in many different ways. Given the depth of the topic, this paper leaves these aspects of implementation for a future discussion. However, we would like to address one area which we believe should be strongly considered a focus after implementation.

As with traditional underwriting practices, it is critical to monitor the results of a process change. Since a predictive model is built from a static sample of policyholders who were actually underwritten using the traditional process, it is important to consider how using it to assess the health risk of a dynamic population of new applicants may result in anti-selection. Is there potential for applicants and producers to game the system and

Figure 3
Graphical Representation of Decision Tree



exploit the reduced requirements? There are several avenues through which life insurers can guard against anti-selection.

First, the third party marketing data cannot be easily manipulated by the applicant. It is reported directly by the third-party agency, and is based upon trends captured over time rather than sudden changes in behavior. Moreover, the model does not rely on any one field from this data, but rather uses it to form a general understanding about a person's lifestyle. It would be very difficult for an applicant to systematically alter behavior over time so it presents a false impression. In fact, if the applicant were successful in systematically altering behavior to change his or her profile, more than likely the applicant's mortality risk would have also changed in the same direction.

To supplement the protection offered by the third party data, it is advisable to maintain a certain degree of unpredictability in which applicants will be allowed to forgo additional requirements. The combination of risk factors that qualify an applicant for reduced requirements at each underwriting class is typically sufficiently complex to offer an initial defense against producers seeking to game the system. While the patterns are not simple enough to be picked up upon easily, we also recommend a percentage of applicants who do qualify be selected at random for traditional underwriting. This will both further disguise the profile of applicants who are eligible for streamlined underwriting, and offer a baseline for monitoring results. If evidence of anti-selection is present in these applicants, the insurer will be alerted of the need to alter the process. As in traditional underwriting, producers will seek to exploit differences in criteria to obtain the best offer for their clients, but this application of predictive modeling does offer important safeguards against potentially damaging behavior.

LEGAL AND ETHICAL CONCERNS

Predictive modeling in life insurance may raise ethical and legal questions. Given the regulations and norms that govern the industry, these questions are understandable. The authors of this paper are not legal experts, but we can offer our insight into several issues, and say that in our experience, it is feasible to assuage these concerns.

Collecting any data about individuals is a sensitive subject. Data collection agencies have been around since the late 19th century, but in the 1960s lawmakers became concerned with the availability of this data as they worried that the rapidly developing computing industry would vastly expand its influence, and lead to potential abuses. This concern resulted in the Fair Credit Reporting Act (FCRA) of 1970. The essence of the law is that provided certain consumer protections are maintained around access and transparency, the efficiency gains of making this data available are worthwhile. We tell this story as a kind of aside because it is the first question asked by many with whom we have discussed predictive modeling. However, as described above, the data provided by the aggregators come from their marketing sets which are not subject to the FCRA.

Even though the third-party marketing data does not face explicit FCRA or signature authority legal restrictions, it can still raise ethical question about whether utilizing the consumer data is overly invasive. The first point to realize is that commercial use of this personal data is not new. For many years it has been a valuable tool in selling consumer goods. Marketing firms build personal profiles on individuals which determine what type of catalogs and mailing advertisements they receive. Google scans the text of searches and emails in order to present people with related advertisements. We believe society has accepted this openness, not without hesitation, because on average it provides more of what we want, less of what we do not. In addition to consumer marketing applications, predictive modeling using third-party consumer data has also been accepted for property and casualty insurance underwriting.

Despite its acceptance in other fields, life insurance has a unique culture, set of norms, and regulations, so additional care must be taken to use this data in ways that are acceptable. A critical step in predictive model development is determining which variables to include in the model. We have described the statistical basis on which these decisions are made, but the process also considers regulatory and business concerns. Before beginning the model build, the legal and compliance functions of the life insurer should be the first to review the list of potential variables. No matter what their predictive powers may be, any variable that is deemed to create a legal or public relations risk, or is counter to the company's "values" should be excluded from the model. Even if not explicitly forbidden by regulations, life insurers should err on the side of caution and exclude variables which convey questionable information, and can feel confident that this caution will not cripple the strength of the model.

The legal and ethical concerns raised also depend upon business decisions that the model is allowed to influence. While in principal, predictive models could play the lead role in assigning underwriting classes for many applicants, insurers have been most comfortable from a compliance perspective utilizing models to triage applications. By using the model as described above to inform the insurer when no further requirements are needed, the model does not take adverse actions for any applicant. In fact, the model only has the potential to take a positive action by offering a streamlined underwriting process that would otherwise be unavailable.

We fully expect and understand that questions will be raised when changes occur to a consumer-facing process like underwriting. We also recognize that predictive modeling is a new and growing trend in life insurance, and the industry culture and regulations may evolve to in ways that impact how data and models are used. For both of these reasons, company legal and compliance experts are key members of every predictive modeling project we agree to support. While we do not claim to be the definitive source on this subject, in our experience thus far, it has been possible to utilize predictive modeling for life insurance underwriting in ways that are compatible with regulatory, ethical, and cultural concerns.

THE FUTURE OF LIFE INSURANCE PREDICTIVE MODELING

Due to rapid improvements in computation power, data storage capacity, and statistical modeling techniques, over the last several decades predictive modeling has come into widespread use by corporations looking to gain a competitive advantage. Banking and credit card industries are well known pioneers for modeling credit card fraud, personal financial credit score for mortgage and loan application, credit card mail solicitation, customer cross-sale, and more.

While insurance has lagged behind other industries, more recently it has gained momentum in data mining and predictive modeling. Early developments include the use of personal financial credit history for pricing and underwriting for personal automobile and homeowners insurance. As it proved successful in personal lines, predictive modeling has spread into commercial insurance pricing and underwriting, as well as into a variety of other applications including price optimization models, life-time customer models, claim models, agency recruiting models, and customer retention models. In just the last several years, predictive modeling is beginning to show promise in the life insurance industry.

Until relatively recently, merely using predictive models to support underwriting, pricing and marketing gave property and casualty insurance companies a competitive edge. However, data analytics has sufficiently penetrated the market so first mover advantages no longer exist. Property and casualty companies must now improve their modeling techniques and broaden the applications to stay ahead of their competition [23]. Because application of data mining and predictive modeling is, for the most part, still new and unexplored territory in life insurance, we do believe those who act first will realize similar first mover gains.

Our experience indicates that using predictive modeling for underwriting can empower life companies to segment and underwrite risks through a more consistent and less expensive process. In doing so, they can reduce costs, improve customer and producer experience, and generate substantial business growth. Tomorrow, we anticipate those who ignore this emerging trend will scramble to catch up while the initial users have moved to models of mortality. As a first step in modeling mortality directly, we have experimented with modeling the main cause of death in the short-term, accidents. At younger ages, insured mortality is driven by accidental death rather than by disease.[1] A sample model we have built to segment which members of a population have been involved in severe auto accidents has shown substantial promise, and is being incorporated into the latest projects we have supported. The more we discuss full-scale models of mortality with insurers, the more excited they become about their potential, and committed to unearthing the data to make them a reality. We believe that day is near.

We would like to close by noting that improvements to efficiency and risk selection will not only accrue to insurers, but also to individuals. Over time, competition will drive insurers to not only capture additional profits from their reduced costs, but also charge lower premiums and require fewer medical tests. Because the predictive models we describe do not disadvantage individual applicants, we believe the long run effect of predictive modeling will be to increase access to insurance. And if the final effect of predictive modeling in life underwriting is in some small way to push people toward healthier lifestyles, we would be happy to claim that as the ultimate victory. ■

Chris Stehno is managing director at Deloitte Consulting LLP. He can be reached at *cstehno@deloitte.com*.

Jim Guszczka is the US chief data scientist of Deloitte Consulting, and a member of Deloitte's Advanced Analytics and Modeling practice. He can be reached at *jguszcza@deloitte.com*.

ENDNOTES

1 According to the National Center for Health Statistics National Vital Statistics Reports from March 2005, the top three causes of death among young adults aged 25-29 are each acute injuries. These account for 61.58 percent of all deaths at those ages. The leading cause of death is accidental injury (34.09 percent), followed by homicide (14 percent), and suicide (13.49 percent).

REFERENCES

1 Davenport, T. H., Harris, J. G., *Competing on Analytics: The New Science of Winning*, Harvard Business School Press, (2006).

2 Guszcza, J., "Analyzing Analytics", *Contingencies,* American Academy of Actuaries, July- August, (2008).

3 McCullagh, P., Nelder, J. A., *Generalized Liner Models*, 2nd Edition, Chapman and Hall, (1989).

4 Brockman, M. J., Wright, T. S., "Statistical Motor Rating: Making Effective Use of Your Data," *Journal of the Institute of Actuaries*, Vol. 119, Part III, (1992).

5 Mildenhall, S.J., "A Systematic Relationship between Minimum Bias and Generalized Linear Models," *Proceedings of Casualty Actuarial Society*, Vol. LXXXVI, Casualty Actuarial Society, (1999).

6 Bailey, R. A., Simon, L. J., "An Actuarial Note on the Credibility of a Single Private Passenger Car", *Proceedings of Casualty Actuarial Society*, Vol. LVII, Casualty Actuarial Society, (1959).

7 Bulhmann, H., "Experience Rating and Credibility", *ASTIN Bulletin IV*, Part III, (1967).

8 "Insurance Scoring in Private Passenger Automobile Insurance – Breaking the Silence," *Conning Report,* Conning, (2001).

9 *Report on the Use of Credit History for Personal Line of Insurance*, American Academy of Actuaries, (2002).

10 "The Top 10 Casualty Actuarial Stories of 2009," *Actuarial Review*, Vol. 37, No. 1, Casualty Actuarial Society, (2010).

11 Guszcza, J., Wu, C. P., "Mining the Most of Credit and Non-Credit Data", *Contingencies,* American Academy of Actuaries, March-April, (2003).

12 Meehl, P., *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, University of Minnesota Press, (1954).

13 Gladwell, M., *Blink: The Power of Thinking without Thinking*, Little Brown and Co, (2005).

14 Lewis, M., *Moneyball: the Art of Winning an Unfair Game,* W. W. Norton & Company, (2003).

15 Ayres, I., *Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart,* Bantam Books, (2007).

16 Gladwell, M., "The Formula", *The New Yorker*, (2006).

17 Meehl P., "Causes and Effects of My Disturbing Little Book," *Journal of Personality Assessment*, Vol. 50, (1986).

18 Dawes, R., Faust, D., Meehl, P., "Clinical Versus Actuarial Judgment," *Science,* 243: 1668-1674, (1989).

19 Guszcza, J., Batty, M., Kroll, A., Stehno, C., "Bring Predictive Models to Life," *Actuarial Software Now - A Supplement to Contingencies*, American Academy of Actuaries, Winter, (2009).

20 Hastie, T., Tibshirani, R., Friedman, J. H., *The Elements of Statistical Learning*, Springer, (2003).

21 Berry, M. J. A., Linoff, G. S., *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, John Wiley & Sons, (2003).

22 Breiman, L., Friedman, J., Olshen, R., Stone, C., *Classification and Regression Trees*, New York: Chapman and Hall, (1993).

23 Yan, J., Masud, M., Wu, C. P., "Staying Ahead of the Analytical Competitive Curve: Integrating the Broad Range Applications of Predictive Modeling in a Competitive Market Environment," *CAS E-Forum*, Winter, Casualty Actuarial Society, (2008).

24 Stehno, C., Johns, C., "You Are What You Eat: Using Consumer Data to Predict Health Risk," *Contingencies*, American Academy of Actuaries, Winter, (2006).

25 Yi Zhang , Jonathan Koren, Efficient bayesian hierarchical user modeling for recommendation system, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, July 2007, Amsterdam, The Netherlands.