PREDICTIVE MODELING, A LIFE UNDERWRITER'S PRIMER

Mark S. Dion, FALU, FLMI
Assistant Chief Underwriter
RGA Reinsurance Company
Chesterfield, MO
mdion@rgare.com

Executive Summary *Use of predictive models is becoming more common throughout the business landscape. Underwriters need to understand the basic concepts as these models impact pricing, marketing and underwriting of life insurance products. This primer introduces and describes predictive modeling, the development of predictive models, types of models, advantages and disadvantages of such models, and closes with a glossary of terms commonly encountered when discussing models with other professionals in your organization. The article also addresses the distinction between predictive modeling and lifestyle-based analytics.*

This primer does not explore the statistical modeling mathematics in detail. Nor does it represent an endorsement for, or an argument against, any implementation or use of predictive modeling.

Predictive Modeling: Introduction and Description

Life insurance companies collect enormous amounts of data on clients and prospects. Since the inception of computer-based data gathering techniques, businesses have pursued methods to put the collected data to work to bring in more profitable business. Data mining, the process of culling and collating the huge volumes of collected transactions, laid the groundwork for bringing statistical analysis to bear on that vast repository. Predictive modeling, drawn from the realm of inferential statistical analysis, has been used for many years as a way to better understand and use the collected data to become more efficient, and hopefully, more profitable (Table 1).

Simply stated, predictive models are analytical tools to predict the probability of an outcome or future behavior. These models build from a number of variables (factors) that are likely to influence future results or behaviors. Use of the term in this article is an extension of predictive analytics, in the context of data mining concerned with forecasting trends and probabilities. Additionally, predictive modeling sometimes uses non-parametric statistical analysis as a form of predictive inference, i.e., the models do not involve the estimation of parameters before the analysis is completed. Rather, these predictive models build from previously observed phenomena or observed outcomes. Analysts often use Bayesian methodology (q.v.) to build predictive models.

In predictive modeling, data is collected for the relevant predictors, the data is cleaned, a statistical model is formulated, predictions are made and the model is validated (or revised) as additional data becomes available. Models may employ simple linear equations, decision trees or complex neural networks. Most are built using sophisticated mathematical modeling software. During the development of a model there are usually many models and factors reviewed.

Use of Predictive Models

- Capacity planning for scarce resources
- Change management
- City planning
- Customer relationship management (CRM)
- Disaster recovery
- Engineering
- Fraud protection - Automotive insurance claim fraud
- Geology and oil exploration
- Health insurance utilization and renewals
- Marketing – Credit card campaigns, Amazon, NetFlix and iTunes recommendations, uplift marketing
- Medical diagnosis and testing
- Meteorology
- Security – Spam filtering
- Security management

Table 1

The goal is to find the optimal uplift, known as "lift," to the hoped-for outcome. Lift could be described as the model's effectiveness, its ability to predict outcomes. Those outcomes might include reduced costs of acquisition, higher direct marketing response rates, more cross-selling opportunities or greater numbers of returning customers.

Models generally add weight to more-recent behaviors and actions. Predictive models should not be allowed to stagnate because conditions and factors upon which a model is built can change over a period of time. Well-constructed models can increase efficiency and decrease costs; poorly constructed models may increase risks to a company's bottom line.

A simple example:
A customer's gender, age and purchase history might predict the likelihood of a future sale. If the purchaser's history included details about genre, characters or author of book, or movie purchases, the model may allow specific types of offerings to their clients – e.g., iTunes, Netflix and Amazon recommendations–the model's lift providing for increased sales. Offer the client items similar to those purchased previously, although the similarities might not be obvious without a model. When purchasing movies, was it the genre, star, co-star, director or plot that captured the purchaser's interest? Would the purchaser be interested in items similar to those purchased by people who purchased the same item? A well-constructed model looks at all of those factors and more, to provide recommendations that optimize profitability and increase sales.

Life insurance example:
After data analysis by a life insurance company, a model is applied to incoming applicants that looks at age, gender, admitted tobacco use, face amount, admitted family history, negative admitted personal medical history, current lipid findings, current hemoglobin A1c, current GGTP, BMI, cotinine results and a pharmacy record check. The model score provides insight as to which applicants require an APS ordered as an automatic requirement and which do not. A well-constructed model would provide a targeted population significantly smaller than a simple age-amount requirement grid. The resulting cost savings may be applied to the wider use of underwriting requirement sentinels that provide a better cost benefit ratio.

Do not confuse predictive modeling with lifestyle analytics (LSA) also known as lifestyle based analytics (LBA). Predictive models may be built using a multitude of factors and data sets, including aggregate results of the many underwriting requirements that underwriters have been using for decades. The example described above uses familiar underwriting tools. In addition, it is not out of the question when a model uses various LSA factors: geographic regions, gym memberships, customer purchase histories, magazine subscriptions, daily travel distance, etc. LSA combined with predictive modeling might be used in straight-through processing environments. The implications of such models are beyond the scope of this introduction.

Life Insurance Predictive Models
- Agency evaluation
- Claims review
- Marketing – Customer segmentation, cross-selling, price optimization
- Pricing
- Reserving
- Risk selection scoring

Risk Selection Applications
- Determining the advisability for ordering Attending Physician Statements or other requirements
- Fraud investigation triggers and over-insurance identifiers
- Preferred screening tools
- Risk scoring for final decisions

Table 2

Developing a Predictive Model

Your company may well be using predictive models in various ways. Predictive modeling is used by various insurers to identify auto or health claim fraud, utilization reviews, marketing approaches, direct response mailings, etc. Incurred but not reported claims (IBNR) calculations, done for financial reporting purposes, might involve a form of predictive modeling. Application of predictive modeling techniques for life insurance underwriting is not widely in use currently, but interest is growing. The Society of Actuaries has sponsored symposia and presentations, and the number of published articles on the topic is growing. In the future, predictive models may become common practice in our industry. The following describes only the barest outline of the work involved with development of a useful model. See Figure 1 (next page).

Models Begin with Data Mining
The goal of data mining is to find patterns in data. Data, once separated from the statistical noise, may unveil patterns of behavior leading to efficient and profitable business.

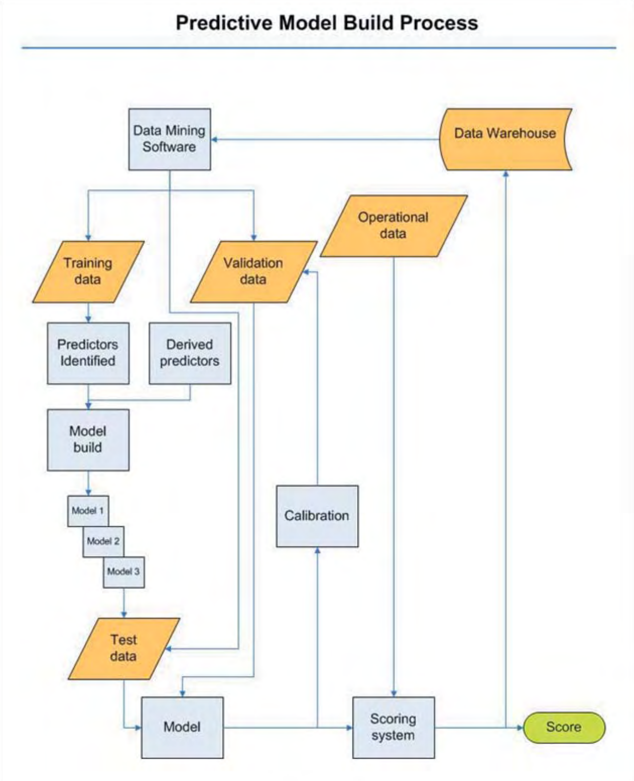

Figure 1

1) Data and Algorithm Preparation
   a) Collect data for predictive variables
   b) Clean the data
   c) Assign data distribution to three portions: training, test, validation
      - Data should be drawn from the same source, at the same time, properly randomized and divided, to be used through the stages of model development

Establish the logic and algorithm

2) Develop decision trees
3) Use training data to identify factors and predictors

Model builds require some trial and error to get the most lift

4) Build the model
   a) Build multiple models attempting to optimize lift
   b) Test the models using a portion of the identified data
   c) Watch for discordancy
   d) Calibrate
5) Validation

Implementation and maintaining a good model

6) Implementation and use
7) Scoring system established
8) Establish audit procedures and feedback to data warehouse
9) Calibration
10) Once a model moves to a production environment, the results require regular monitoring and recalibration.

Models in Production

When the model has proven itself and provides sufficient lift, a scoring system will provide guidance. For example, a scoring system might provide a score designed by the model development team to suggest whether to order an APS. Based on the factors used in the model, a score of 0-60 might yield a pass result with no further action. A score of 61-74 might generate a "refer to underwriter" with a recommendation for an APS. The next class of 75-95 might generate an automatic APS. Finally, a score of 96-100 might raise a red flag to the underwriter that, based on this sample model, there may be significant anti-selection or perhaps a suspicion of attempted fraud.

Aside from the score, the system should feedback to the data warehouse providing additional data elements and outcome results that can assist in further improvement of the model.

Types of Predictive Models

The following list provides a glimpse into the number of model possibilities. Unfortunately, fuller descriptions exceed this article's intent. The reader should consult the references for additional description and examples.

1) Classification and Regression Trees (CART) – Sort groups and population into smaller discrete branches and nodes.
2) Cox Proportional Hazard – A form of survival modeling and a hazard function. Hazard functions are an estimate of the relative risk of a terminal event, such as a death. This survival model is a multivariate technique for analyzing the effect of two or more metric and/or non-metric variables on survival.
3) Decision Tree Analysis – Sorts decisions into smaller branches and nodes, weighting the decisions based on factors identified.
4) Generalized Linear Model (GLM) – Linear or logistic regression models. The most commonly described in life insurance literature, relatively speaking, this is a simple way to model using multiple variables that interact in ways that are not obvious using univariate analysis. See expanded discussion below.
5) k-Nearest Neighbor (kNN) – Uses data that aggregates based on classification. Predictions are based on population density of the factors in the training sample. A simple system often applied to "machine learning."

6) Logistic Regression – Logistic regression is a category of statistical models referred to as generalized linear models (q.v.). The goal of logistic regression is to predict the category of outcome for individual cases using the most efficient model.
7) Naïve Bayes Classifier – Assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.
8) Neural Networks (NN) – As the name suggests, models that have the structural appearance of animal neurons. The concept suggests that many inputs can result in a single (or at least smaller number) output. NNs can produce multiple outcomes however. (Figure 2)
9) Regression Splines – Implies regression analysis that requires data points that must be interpolated, or somehow smoothed.

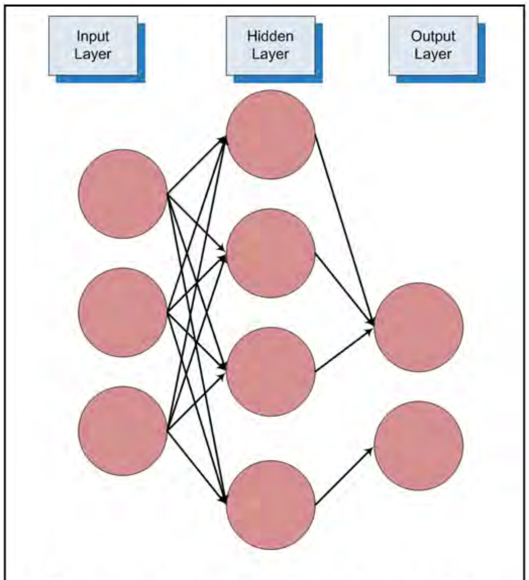

Figure 2: a simple Neural Net (NN)

While this article has not to this point dwelt on the statistical details of predictive modeling, an example chosen from above will expand the description of one of the model types encountered in the actuarial literature.

Generalized Linear Modeling (GLM) – Extends linear regression models to both non-normal distributions and linear transformation (transformation of linearity). First we can describe a conventional linear model that specifies the relationship between a dependent variable Y, and a set of predictor variables, the X's, so that

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k$$

In this equation b_0 is the regression coefficient for the intercept and the b_i values are the regression coefficients (for variables 1 through k) computed from the data. Linear models as described make a set of somewhat restrictive assumptions:
- Dependent variable y is in normal distribution and conditioned on the value of predictors
- A constant variance, regardless of the predicted response

The advantages of linear models with the above restrictions are:
- An easy-to-interpret model form
- Relatively simple computations
- Readily analyzed to determine the quality of the fit

Generalized linear models relax these restrictions. They adjust responses that violate the linear model assumptions in two ways:
1) Link functions model responses when a dependent variable is assumed to be related to the predictors in a non-linear fashion. These functions, and there are many, transform a target range so the simple form of linear models can be maintained.
2) Variance functions used, which express the variance as a function of the predicted response. This allows responses with variances that are not constant.

Advantages of Predictive Models

1) Ability to detect complex non-linear relationships between dependent and independent variables.
2) Ability to detect all possible interactions between predictor variables.
3) Availability of multiple training algorithms.
4) Some forms of predictive modeling, neural nets for example, require less formal statistical training.
5) Bayesian methods can arguably be used to discriminate between conflicting hypotheses: hypotheses with very high support should be accepted as true, and those with very low support should be rejected as false.

Disadvantages of Predictive Models

1) A perceived "black box" nature, making it difficult to describe and explain results; the proprietary nature and structure of each model reinforces the perception.
2) Subject to bias, some may argue that inference methodology might be biased by perceptions held before any evidence is collected.
3) Computational intensity requiring adequate technology infrastructure and statistical analysis acumen.
4) Prone to overfitting and therefore inaccuracies caused by fluctuations in irrelevant or erroneous predictors.

5) Sensitive to changes in conditions and therefore require close monitoring of the environment to which the model is being applied.
6) Models as mathematical constructs can at times yield results that fly in the face of common sense
7) Regulatory concerns if models are not adequately understood or explained.

Modeling Concepts and Terms

The following terms and concepts provide readers new to the area of predictive modeling a basic lexicon, useful for discussions within their respective organizations.

Algorithms – Expression of a problem as a sequence of logical steps.

Bayesian Methods – Based on Bayes' theorem. Bayesian inference presumes collection of evidence that is either consistent with or inconsistent with a given hypothesis (H1). As we gather evidence, our confidence in the hypothesis ought to change. Given sufficient evidence, the degree of confidence should become either much higher or much lower. Underwriters should already be familiar with other Bayesian models, specifically the specificity and sensitivity of a test, and the test's corresponding positive or negative predictive values.

Calibration – During development and thereafter, as the environment changes, or to maintain lift, predictive models must be calibrated, reworking predictors and factors. Some circumstances will change frequently, for example, a new life insurance marketing plan, a new product with different pricing and expense assumptions, or outside influences such as competition with similar products.

Data Mining – Data mining is the process of finding patterns and correlations among dozens of data elements in relational databases, especially large databases. This process allows a user the ability to analyze data from different perspectives and summarize the data into useful information. The software to mine the data is often a different tool than one used to build a predictive model.

Decision Trees – A decision tree as discussed here depicts rules for dividing data into groups. The first rule splits the entire data set into a number of pieces, and then another rule may be applied to a piece, different rules to different pieces, forming a second generation of pieces. In general, a piece may be either split or left alone to form a final group. (Figure 3)

Derived Predictors – Mathematically transformed predictor variables, sometimes known as synthetic predictors. These predictors are derived from mined data and through application of algorithms developed during the model-building process. They were not identified as useful until discovered during the development phase.

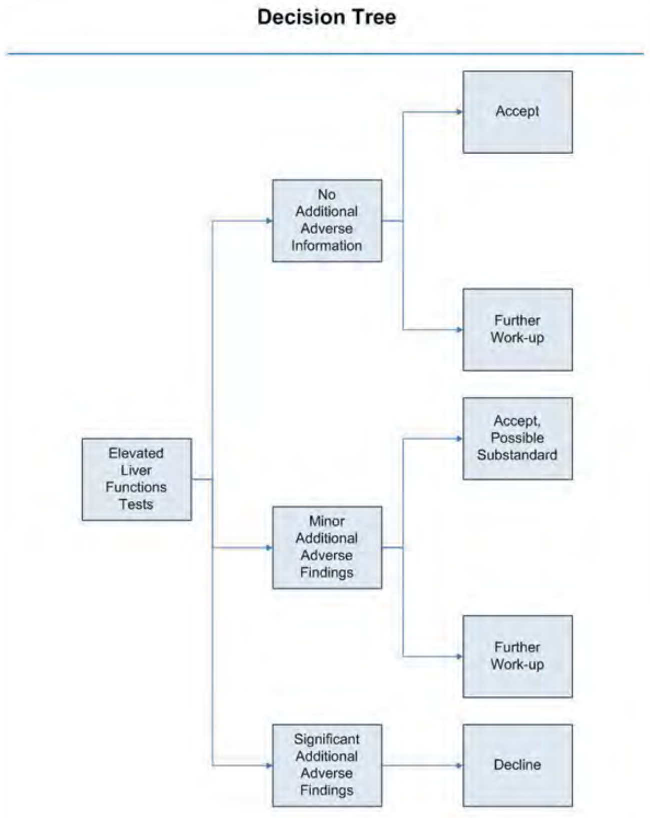
Decision Tree
Elevated Liver Functions Tests
No Additional Adverse Information
Minor Additional Adverse Findings
Significant Additional Adverse Findings
Accept
Further Work-up
Accept, Possible Substandard
Further Work-up
Decline

Figure 3

Lifestyle Based Analytics – Most readers are familiar with the links between lifestyle characteristics and various medical conditions. Various data aggregators and vendors now offer over 2,000 pieces of information (data fields) on various applicants. While a powerful tool for marketing segmentation, which can arguably raise flags concerning higher incidence of disease based on the population predictors, LBA cannot provide the same predictive value to which we are currently accustomed from routine fully underwritten business on an individual case-by-case basis. LBA may be considered in building predictive models for life insurance risk selection, but it should not be confused with the term *predictive modeling*. Robust life insurance predictive models can be prepared without LBA.

Linear Regression – Regression uses the value of one variable in a pair of variables, to predict the value of the second variable. Linear regression attempts to pass a line through the observed variable pairs in a given sample data set. Models built using linear regression are fine when the value of one variable changes, is conditional, upon the value of the other variable. If we wish to study the probability distribution of both variables (or more variables) as they both change, we move to the realm of multivariate analysis (q.v.).

Logistic Regression – Logistic regression predicts the probability of occurrence of an event by fitting data to a logistic curve. Consider it a generalized linear model used for binomial regression. One example, the probability that a person has or will develop diabetes within a specified period might be predicted from knowledge of the person's age, sex and body mass index (BMI). Logistic regression is used extensively in medical literature.

Multivariate Analysis – A statistical technique that allows analysts to review several dependent variables simultaneously. Multivariate analysis helps summarize data and reduce the number of variables necessary to describe it. The multiple dimensions involved in data mining require statistical tools capable of measuring the effects of the interactions of many variables in action at the same time.

Neural Net – Named for the resemblance to an anatomical neural system, a non-linear data modeling tool, used to model complex relationships between inputs and outputs or to find patterns in data. The multiple inputs, once processed by the NN model, produce a single output, if things turn out as hoped. Input is acted upon by factors not observed, called the hidden layer, to provide an output. The hidden layer allows the action based on recognized patterns. While the hidden layer can theoretically become quite large, the result can be overfitting (q.v.).

N.B. Predictive models may produce more than one output.

Overfitting – Overfitting exists when the model in fact describes noise, another expression for statistical random error, rather than the studied relationship. Overfitting will usually cause a model to perform poorly as small fluctuations in data can cause inaccurate results. Too many predictors included in a model make this problem more likely.

Parametric vs. Non-Parametric Statistical Analysis – When certain assumptions about the underlying population are questionable, non-parametric tests can be used in place of their parametric counterparts.

Closing

The literature relating to life insurance applications of predictive modeling techniques is relatively sparse compared with other disciplines. While the techniques do lend themselves to life insurance business, they are not as widely used as in auto or health insurance. Recently interest by actuaries and business heads suggests these tools will become much more common and be with us as a risk selection tool going forward.

The reader should refer to the Society of Actuaries website for additional information on this topic at www.soa.org. For additional information regarding data mining, refer to the excellent introduction by Berry & Linoff, 2004.

References

Batty, Mike, Tripathi, Arun, Kroll, Alice, Wu, Cheng-sheng Peter, Moore, David, Stehno, Chris, Lau, Lucas, Guszcza, Jim, Katcher, Mitch, *Predictive Modeling for Life Insurance, Ways Life Insurers Can Participate in the Business Analytics Revolution*; Deloitte Consulting LLP. April 2010

Berry, Michael J. A. and Linoff, Gordon; *Data Mining Techniques for Marketing, Sales and Customer Management*; John Wiley and Sons, Inc; 2004

Cox, D. R. Regression models with life tables; *Journal of the Royal Statistical Society*, 34, 187-220. 1972.

Galen, Robert S., Gambino, S. Raymond; *Beyond Normality: the Predictive Value and Efficacy of Medical Diagnoses*; John Wiley and Sons, Inc; 2001.

Geisser, Seymour. *Modes of Parametric Statistical Inference*; Wiley, 2006.

Geisser, Seymour. *Predictive Inference: An Introduction. Monographs on Statistics and Applied Probability*; 55 New York: Chapman & Hall, 1993.

Grossman, Robert, Bailey, Stewart, Hallstrom, Philip, et al. *The Management and Mining of Multiple Predictive Models Using the Predictive Modeling Markup Language (PPML)*; 1999

Kahneman, Daniel, Slovic, Paul, and Tversky (editors); *Judgment Under Uncertainty: Heuristics and Biases*; Cambridge University Press; 1982

Leonard, Thomas, and Hsu, John S.J.; *Bayesian Methods, An Analysis for Statisticians and Interdisciplinary Researchers*; Cambridge University Press 1999.

London, Dick; *Survival Models and Their Estimation 2nd Ed*; ACTEX Publications, Winstead Connecticut; 1988.

McCullagh, Peter, and Nelder, J.A.; *Generalized Linear Models – Second Edition*; *Monographs on Statistics and Applied Probability* 37, New York: Chapman & Hall 1999

Montgomery, Douglas C., Peck, Elizabeth A., and Vining, G Geoffrey; *Introduction to Linear Regression Analysis*; John Wiley and Sons, Inc; 2001

Mosley, Roosevelt C.; *The Use of Predictive Modeling in the Insurance Industry*; Pinnacle Actuarial Resources, Inc., January 2005.

Rozar, Tim, Rushing, Susan, Willeat Susan, *Report on the Lapse and Mortality Experience of the Post-Level Premium Period Term Plans*; Society of Actuaries July 2010.

Shapiro, A.F. and Jain, L.C. (editors); *Intelligent and Other Computational Techniques in Insurance*; World Scientific Publishing Company; 2003.

Smith, J. Maynard; *Mathematical Ideas in Biology*; Cambridge University Press; 1968.

Speicher MD, Carl E., Smith Jr. MD MS, Jack W; *Choosing Effective Laboratory Tests*; W.B. Saunders; 1983.

Staudt, Andy; Why Predictive Modeling for Life Insurance and Annuities? *Product Matters!* February 2010.

About the Author

Mark Dion is Vice President, Underwriting Rules and Education, for RGA Reinsurance Company, and has responsibilities for facultative risk management in RGA's U.S. division. In this role, he provides support for various department projects including underwriter training, underwriting manual support, simplified issue, Bancassurance and automated rules for underwriting. He also coordinates the training program for new RGA underwriters. Mark received a Bachelor of Science degree in biology with a philosophy co-major from Creighton University in Omaha, NE. He is a Fellow of the Academy of Life Underwriting and a Fellow of the Life Management Institute. He currently serves on the Society of Actuaries Life Insurance Mortality and Underwriting Survey Committee.