Predictive Modeling for Life Insurance

Ways Life Insurers Can Participate in the Business Analytics Revolution

Prepared by

Mike Batty, FSA, CERA
Arun Tripathi, Ph.D.
Alice Kroll, FSA, MAAA
Cheng-sheng Peter Wu, ASA, MAAA, FCAS
David Moore, FSA, MAAA
Chris Stehno
Lucas Lau
Jim Guszcza, MAAA, FCAS, Ph.D.
Mitch Katcher, FSA, MAAA

Deloitte Consulting LLP
April 2010

Predictive Modeling for Life Insurance

Ways Life Insurers Can Participate in the Business Analytics Revolution

Abstract

The use of advanced data mining techniques to improve decision making has already taken root in property and casualty insurance as well as in many other industries [1, 2]. However, the application of such techniques for more objective, consistent and optimal decision making in the life insurance industry is still in a nascent stage. This article will describe ways data mining and multivariate analytics techniques can be used to improve decision making processes in such functions as life insurance underwriting and marketing, resulting in more profitable and efficient operations. Case studies will illustrate the general processes that can be used to implement predictive modeling in life insurance underwriting and marketing. These case studies will also demonstrate the segmentation power of predictive modeling and resulting business benefits.

Keywords: Predictive Modeling, Data Mining, Analytics, Business Intelligence, Life Insurance Predictive Modeling

Predictive Modeling for Life Insurance

Ways Life Insurers Can Participate in the Business Analytics Revolution

Contents

Predictive Modeling for Life Insurance

Ways Life Insurers Can Participate in the Business Analytics Revolution

The Rise of "Analytic" Decision Making

Predictive modeling can be defined as the analysis of large data sets to make inferences or identify meaningful relationships, and the use of these relationships to better predict future events [1,2]. It uses statistical tools to separate systematic patterns from random noise, and turns this information into business rules, which should lead to better decision making. In a sense, this is a discipline that actuaries have practiced for quite a long time. Indeed, one of the oldest examples of statistical analysis guiding business decisions is the use of mortality tables to price annuities and life insurance policies (which originated in the work of John Graunt and Edmund Halley in the 17th century). Likewise, throughout much of the 20th century, general insurance actuaries have either implicitly or explicitly used Generalized Linear Models [3,4,5] and Empirical Bayes (a.k.a. credibility) techniques [6,7] for the pricing of short-term insurance policies. Therefore, predictive models are in a sense, "old news." Yet in recent years, the power of statistical analysis for solving business problems and improving business processes has entered popular consciousness and become a fixture in the business press. "Analytics," as the field has come to be known, now takes on a striking variety of forms in an impressive array of business and other domains.

Credit scoring is the classic example of predictive modeling in the modern sense of "business analytics." Credit scores were initially developed to more accurately and economically underwrite and determine interest rates for home loans. Personal auto and home insurers subsequently began using credit scores to improve their selection and pricing of personal auto and home risks [8,9]. It is worth noting that one of the more significant analytical innovations in personal property-casualty insurance in recent decades originated outside the actuarial disciplines. Still more recently, U.S. insurers have widely adopted scoring models – often containing commercial credit information – for pricing and underwriting complex and heterogeneous commercial insurance risks [10].

The use of credit and other scoring models represents a subtle shift in actuarial practice. This shift has two related aspects. First, credit data is behavioral in nature and, unlike most traditional rating variables, bears no direct causal relationship to insurance losses. Rather, it most likely serves as a proxy measure for non-observable, latent variables such as "risk-seeking temperament" or "careful personality" that are not captured by more traditional insurance rating dimensions. From here it is a natural leap to consider other sources of external information, such as lifestyle, purchasing, household, social network, and environmental data, likely to be useful for making actuarial predictions [11, 24].

Second, the use of credit and other scoring models has served as an early example of a widening domain for predictive models in insurance. It is certainly natural for actuaries to employ modern analytical and predictive modeling techniques to arrive at better solutions to traditional actuarial problems such as estimating mortality, setting loss reserves, and establishing classification ratemaking schemes. But actuaries and other insurance analytics are increasingly using predictive modeling techniques to improve business processes that traditionally have been largely in the purview of human experts.

For example, the classification ratemaking paradigm for pricing insurance is of limited applicability for the pricing of commercial insurance policies. Commercial insurance pricing has traditionally been driven more by underwriting judgment than by actuarial data analysis. This is because commercial policies are few in number relative to personal insurance policies, are more heterogeneous, and are described by fewer straightforward rating dimensions. Here, the scoring model paradigm is especially useful. In recent years it has become common for scoring models containing a large number of commercial credit and non-credit variables to ground the underwriting and pricing process more in actuarial analysis of data, and less in the vagaries of expert judgment. To be sure, expert underwriters remain integral to the process, but scoring models replace the blunt instrument of table- and judgment-driven credits and debits with the precision tool of modeled conditional expectations.

Similarly, insurers have begun to turn to predictive models for scientific guidance of expert decisions in areas such as claims management, fraud detection, premium audit, target marketing, cross-selling, and agency recruiting and placement. In short, the modern paradigm of predictive modeling has made possible a broadening, as well as a deepening, of actuarial work.

As in actuarial science, so in the larger worlds of business, education, medicine, sports, and entertainment. Predictive modeling techniques have been effective in a strikingly diverse array of applications such as:

- Predicting criminal recidivism [12]
- Making psychological diagnoses [12]
- Helping emergency room physicians more effectively triage patients [13]
- Selecting players for professional sports teams [14]
- Forecasting the auction price of Bordeaux wine vintages [15]
- Estimating the walk-away "pain points" of gamblers at Las Vegas casinos to guide casino personnel who intervene with free meal coupons [15]
- Forecasting the box office returns of Hollywood movies [16]

A common theme runs through both these and the above insurance applications of predictive modeling. Namely, in each case predictive models have been effective in domains traditionally thought to be in the sole purview of human experts. Such findings are often met with surprise and even disbelief. Psychologists, emergency room physicians, wine critics, baseball scouts, and indeed insurance underwriters are often and understandably surprised at the seemingly uncanny power of predictive models to outperform unaided expert judgment. Nevertheless, substantial academic research, predating the recent enthusiasm for business analytics by many decades, underpins these findings. Paul Meehl, the seminal figure in the study of statistical versus clinical prediction, summed up his life's work thus [17]:

> There is no controversy in social science which shows such a large body of quantitatively diverse studies coming out so uniformly in the same direction as this one. When you are pushing over 100 investigations, predicting everything from the outcome of football games to the diagnosis of liver disease, and when you can hardly come up with half a dozen studies showing even a weak tendency in favor of the clinician, it is time to draw a practical conclusion.

Certainly not all applications of predictive modeling have a "clinical versus actuarial judgment" character [18]. For example, amazon.com and netflix.com make book and movie recommendations without any human intervention [25]. Similarly, the elasticity-optimized pricing of personal auto insurance policies can be completely automated (barring regulatory restrictions) through the use of statistical algorithms. Applications such as these are clearly in the domain of machine, rather than human, learning. However, when seeking out ways to improve business processes, it is important to be cognizant of the often surprising ability of predictive models to improve judgment-driven decision-making.

Current State of Life Insurance Predictive Modeling

While life insurers are noted among the early users of statistics and data analysis, they are absent from the above list of businesses where statistical algorithms have been used to improve expert-driven decisions processes. Still, early applications of predictive modeling in life insurance are beginning to bear fruit, and we foresee a robust future in the industry [19].

Life insurance buffers society from the full effects of our uncertain mortality. Firms compete with each other in part based on their ability to replace that uncertainty with (in aggregate) remarkably accurate estimates of life expectancy. Years of fine-tuning these estimates have resulted in actuarial tables that mirror aggregate insured population mortality, while underwriting techniques assess the relative risk of an individual. These methods produce relatively reliable risk selection, and as a result have been accepted in broadly similar fashion across the industry. Nonetheless, standard life insurance underwriting techniques are still quite costly and time consuming. A life insurer will typically spend approximately one month and several hundred dollars underwriting each applicant[1].

Many marginal improvements to the underwriting process have taken hold: simplified applications for smaller face amounts, refinement of underwriting requirements based upon protective value studies, and streamlined data processing via automated software packages are all examples. However, the examples in the previous section suggest that property-casualty insurers have gone farther in developing analytics-based approaches to underwriting that make better use of available information to yield more accurate, consistent, and efficient decision-making. Based on our experience, life insurance underwriting is also ripe for this revolution in business intelligence and predictive analytics. Perhaps

[1] According to the Deloitte 2008 LIONS benchmarking study of 15 life insurers, the median service time to issue a new policy ranges between 30 and 35 days for policies with face amounts between $100k to $5 million, and the average cost of requirements (excluding underwriter time) is $130 per applicant.

motivated by the success of analytics in other industries, life insurers are now beginning to explore the possibilities[2].

Despite our enthusiasm, we recognize that life underwriting presents its own unique set of modeling challenges which have made it a less obvious candidate for predictive analytics. To illustrate these challenges it is useful to compare auto underwriting, where predictive modeling has achieved remarkable success, with life underwriting, where modeling is a recent entry. Imagine everything an insurer could learn about a prospective customer: age, type of car, accident history, credit history, geographic location, personal and family medical history, behavioral risk factors, and so on. A predictive model provides a mapping of all these factors combine onto the expected cost of insuring the customer. Producing this map has several prerequisites:

- A clearly defined target variable, i.e. what the model is trying to predict
- The availability of a suitably rich data set, in which at least some predictive variables correlated with the target can be identified
- A large number of observations upon which to build the model, allowing the abiding relationships to surface and be separated from random noise
- An application by which model results are translated into business actions

While these requirements are satisfied with relative ease in our auto insurance example, life insurers may struggle with several of them.

	Auto Insurer	**Life Insurer**
Target Variable	Claims over six-month contract	Mortality experience over life of product (10, 20+ years)
Modeling Data	Underwriting requirements supplemented by third-party data	Underwriting requirements supplemented by third-party data
Frequency of Loss	Approximately 10 percent of drivers make claims annually	Typically, fewer than 1 first year death per 1,000 new policies issued
Business Action	Underwriting Decision	Underwriting Decision

Statisticians in either domain can use underwriting requirements, which are selected based upon their association with insurance risk, supplement them with additional external data sources, and develop predictive models that will inform their underwriting decisions. However, the target variable and volume of data required for life insurance models raise practical concerns.

For the auto insurer, the amount of insurance loss over the six-month contract is an obvious candidate for a model's target variable. But because most life insurance is sold through long duration contracts, the analogous target variable is mortality experience over a period of 10, 20, or often many more years. Because the contribution of a given risk factor to mortality may change over time, it is insufficient to analyze mortality experience over a short time horizon. Further, auto insurers can correct underwriting

[2] As reported in an SOA sponsored 2009 study, "Automated Life Underwriting," only 1 percent of North American life insurers surveyed are currently utilizing predictive modeling in their underwriting process.

mistakes through rate increases in subsequent policy renewals, whereas life insurers must price appropriately from the outset.

The low frequency of life insurance claims (which is good news in all other respects) also presents a challenge to modelers seeking to break ground in the industry. Modeling statistically significant variation in either auto claims or mortality requires a large sample of loss events. But whereas approximately 10 percent of drivers will make a claim in a given year, providing an ample data set, life insurers can typically expect less than one death in the first year of every 1,000 policies issued[3]. Auto insurers can therefore build robust models using loss data from the most recent years of experience, while life insurers will most likely find the data afforded by a similar time frame insufficient for modeling mortality.

The low frequency of death and importance of monitoring mortality experience over time leaves statisticians looking for life insurance modeling data that spans many (possibly 20) years. Ideally this would be a minor impediment, but in practice, accessing usable historical data in the life insurance industry is often a significant challenge. Even today, not all life insurers capture underwriting data in an electronic, machine-readable format. Many of those that do have such data only implemented the process in recent years. Even when underwriting data capture has been in place for years, the contents of the older data (i.e. which requirements were ordered) may be very different from the data gathered for current applicants.

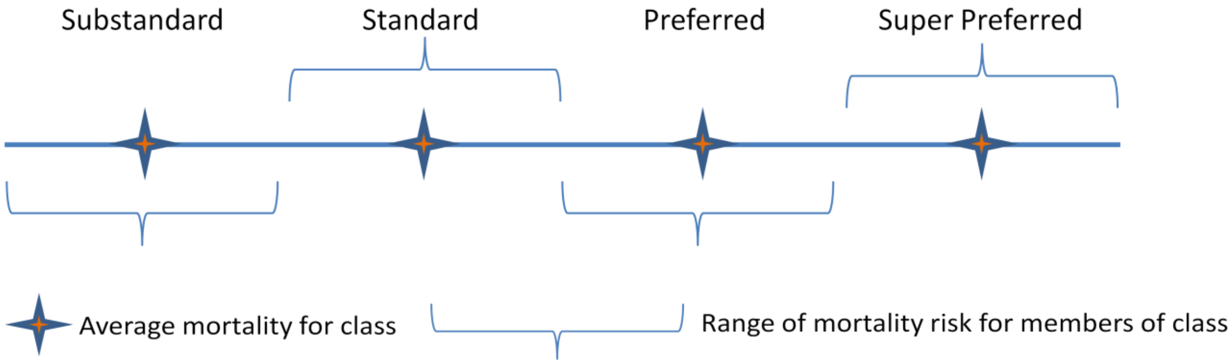
These challenges do not preclude the possibility of using predictive modeling to produce refined estimates of mortality. However, in the short term they have motivated a small, but growing number of insurers to begin working with a closely related yet more immediately feasible modeling target: the underwriting decision on a newly issued policy. Modeling underwriting decisions rather than mortality offers the crucial advantage that underwriting decisions provide informative short term feedback in high volumes. Virtually every application received by a life insurer will have an underwriting decision rendered within several months. Further, based upon both historical insurer experience and medical expertise, the underwriting process is designed to gather all cost-effective information available about an applicant's risk and translate it into a best estimate of future expected mortality. Therefore, using the underwriting decision as the target variable addresses both key concerns that hinder mortality-predicting models.

Of course, underwriting decisions are imperfect proxies for future mortality. First, life underwriting is subject to the idiosyncrasies, inconsistencies, and psychological biases of human decision-making. Indeed this is a major motivation for bringing predictive models to bear in this domain. But do these idiosyncrasies and inconsistencies invalidate underwriting decisions as a candidate target variable? No. To the extent that individual underwriters' decisions are independent of one another and are not affected by common biases, their individual shortcomings tend to "diversify away." A famous example

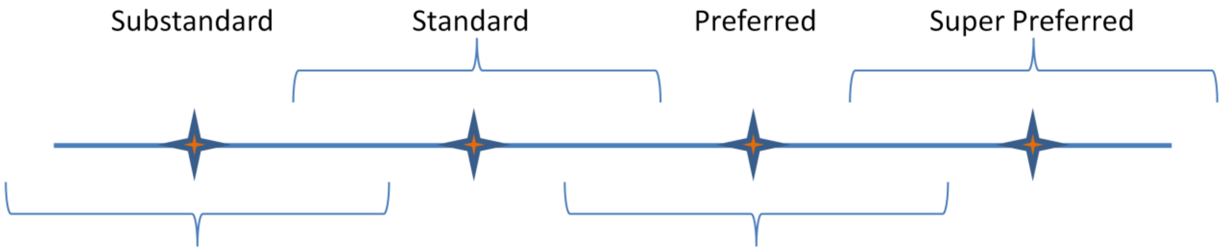
[3] This is an estimate based upon industry mortality tables. Mortality experience varies across companies with insured population demographics. In the 2001 CSO table, the first-year select, weighted average mortality rate (across gender and smoker status) first exceeds 1 death per thousand at age 45.

illustrates this concept. When Francis Galton analyzed 787 guesses of the weight of an ox from a contest at a county fair, he found that the errors of the individual guesses essentially offset one another, and their average came within one pound of the true weight of the ox. This illustrates how regression and other types of predictive models provide a powerful tool for separating “signal” from “noise”.

In fact, the Galton example is quite similar to how life insurers manage mortality. Although individual mortality risk in fact falls along a continuum, insurers group policyholders into discrete underwriting classes and treat each member as if they are of average risk for that class. When the risks are segmented sufficiently, insurers are able to adequately price for the aggregate mortality risk of each class.


Substandard
Standard
Preferred
Super Preferred
Average mortality for class
Range of mortality risk for members of class

However, to avoid anti-selection and maintain the integrity of the risk pools insurers must segment risks into classes that are homogenous. While the “noise” in underwriting offers may diversify, these offers are accepted or rejected by applicants strategically. On average, applicants who have knowledge of their own health statuses will be more likely to accept offers that are in their favor, and reject those that are disadvantageous. For example, in the figure below an applicant at the upper range of the standard class may qualify for preferred with another insurer, thus leaving the risk profile of the original standard class worse than expected.



Therefore, anything that widens the range of mortality risks in each class, and thus blurs the lines between them, poses a threat to a life insurer. In addition to the inconsistency of human decision making, global bias resulting from company-wide underwriting guidelines that may not perfectly represent expected mortality can also contribute to this potential problem.

While modeling underwriting decisions may ultimately become a step along the path towards modeling mortality directly, we do believe today it is a pragmatic approach that provides the maximal return on modeling investment today. Specifically, utilizing underwriting decisions as the target variable is advantageous because they are in generous supply, contain a great deal of information and expert judgment, and do not require long "development" periods as do insurance claims. At the same time they contain diversifiable "noise" that can be dampened through the use of predictive modeling. Although building models for mortality and improving risk segmentation remain future objectives, utilizing predictive models based upon historical underwriting decisions represents a significant improvement on current practice, and is a practical alternative in the common scenario where mortality data is not available in sufficient quantities for modeling.

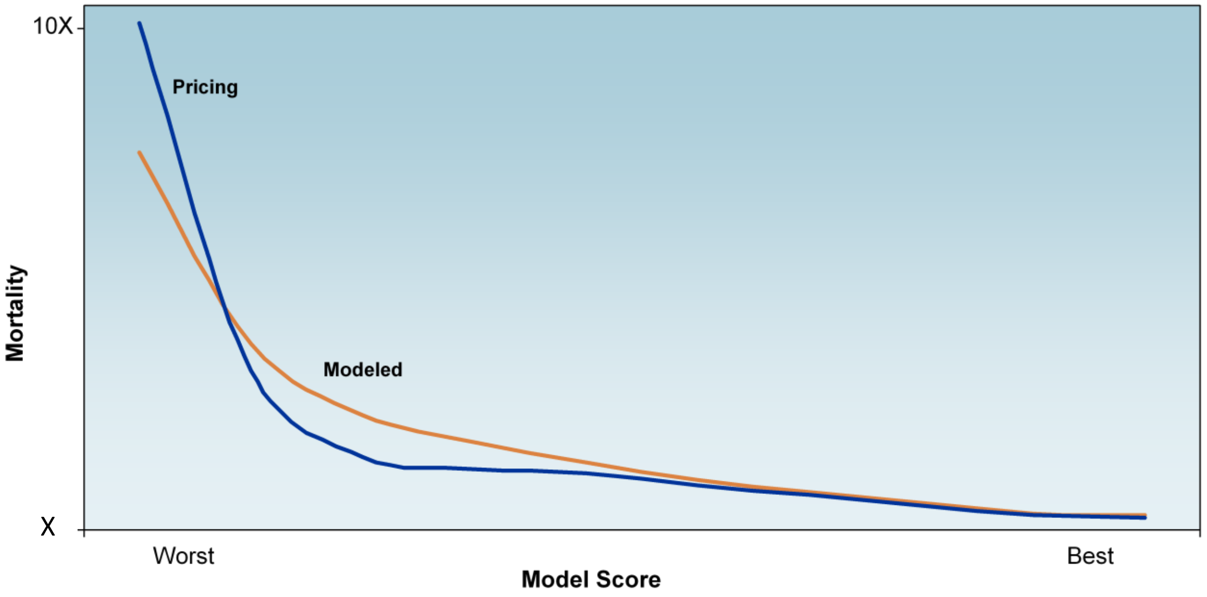
Business Application That Can Help Deliver a Competitive Advantage

We will describe the technical aspects of underwriting predictive models in some detail in a subsequent section. While that discussion may beguile certain members of the audience (the authors included), others will be more interested in understanding how predictive modeling can deliver a competitive advantage to life insurers.

Life Underwriting

Unsurprisingly, one compelling application has been to leverage models that predict underwriting decisions directly within the underwriting process. As mentioned above, underwriting is a very costly and time consuming, but necessary, exercise for direct life insurance writers. Simply put, the underwriting process can be made faster, more economical, more efficient, and more consistent when a predictive model is used to analyze a limited set of underwriting requirements and inexpensive third-party marketing data sources (both described below) to provide an early glimpse of the likely underwriting result. As illustrated in Figure 1, the underwriting predictive models that Deloitte has helped insurers develop have been able to essentially match the expected mortality for many applicants. These insurers are beginning to leverage model results to issue many of their policies in just several days, thus foregoing the more costly, time consuming, and invasive underwriting requirements.

Figure 1: Mortality of Predictive Model vs. Full Underwriting



Risks which had been underwritten by the insurer and kept in a holdout sample were rank-ordered by model score and divided into equal-sized deciles. Modeled mortality is computed by taking a weighted average of the insurer's mortality estimates for each actual underwriting class in proportion to their representation within each decile. Pricing mortality represents the fully underwriting pricing mortality assumptions.

Issuing more policies with fewer requirements may initially seem like a radical change in underwriting practices, but we think of it as an expansion of a protective value study. Just as insurers currently must judge when to begin ordering lab tests, medical exams records, and EKGs, the models are designed to identify which applicant profiles do and do not justify the cost of these additional requirements. Based on the results of the models we've helped insurers build thus far, the additional insight they provide has allowed these insurers to significantly change the bar on when additional tests are likely to reveal latent risk factors. As indicated by the quality of fit between the model mortality and pricing assumptions, these models have been able to identify approximately 30 percent to 50 percent of the applicants that can be issued policies through a streamlined process, and thus avoid the traditional requirements.

With impressive frequency, the underwriting decision recommended by these models matched the decision produced through full underwriting. For cases when they disagree, however, we offer two possibilities: 1) the models do not have access to information contained in the more expensive requirements which may provide reason to change the decision, or 2) models are not subject to biases or bounded cognition in the same way that underwriters, who do not always act with perfect consistency or optimally weigh disparate pieces of evidence, are. The latter possibility comports with Paul Meehl's and his colleagues' studies of the superiority of statistical over clinical decision making, and is further motivation for augmenting human decision-making processes with algorithmic support.

In our analyses of discrepancies between models and underwriting decisions we did encounter cases where additional underwriting inputs provided valuable results, but they were rivaled by instances of underwriting inconsistency. When implementing a model, business rules are used to capitalize upon the model's ability to smooth inconsistency, and channel cases where requirements are likely to be of value to the traditional underwriting process. Thus, our experience therefore suggests that insurance underwriting can be added to the Meehl school's long list of domains where decision-making can be materially improved through the use of models.

These results point to potentially significant cost savings for life insurers. Based on a typical company's volume, the annual savings from reduced requirements and processing time are in the millions, easily justifying the cost of model development. Table 1 shows a rough example of the potential annual savings for a representative life insurer. It lists standard underwriting requirements and roughly typical costs and frequencies with which they would be ordered in both a traditional and a model-enhanced underwriting process. It then draws a comparison between the costs of underwriting using both methods.

Table 1: Illustrative Underwriting Savings from Predictive Model

	Requirement Cost	Requirement Utilization	
		Traditional Underwriting	Predictive Model
Paramedical Exam	$55	50%	0%
Oral Fluids Analysis	$25	20%	0%
Blood and Urine Analysis	$55	70%	0%
MVR Report	$6	70%	75%
Attending Physician Statement	$100	20%	0%
Medical Exam	$120	20%	0%
EKG	$75	10%	0%
Stress Test	$450	1%	0%
Third-Party Data	$0.50	0%	100%
Total Cost Per Applicant		**$130**	**$5**
Savings Per Applicant		**$125**	
Annual Applications Received		**50,000**	
Annual Savings (over 30% to 50% of applications)		**$2 to $3 million**	

In addition to hard dollars saved, using a predictive model in underwriting can generate opportunities for meaningful improvements in efficiency and productivity. For example, predictive modeling can shorten and reduce the invasiveness of the underwriting. The time and expense required to underwrite an application for life insurance and make an offer is an investment in ensuring that risks are engaged at an appropriate price. However, the effort associated with the underwriting process can be considered a deterrent to purchasing life insurance. Resources spent while developing a lead, submitting an application, and underwriting a customer who does not ultimately purchase a policy are wasted from the perspective of both the producer and home office. The longer that process lasts, and the more tests the applicant must endure, the more opportunity the applicant has to become frustrated and abort the purchase entirely, or receive an offer from a competitor. Further, complications with the underwriting process also provide a disincentive for an independent producer to bring an applicant to a given insurer. Enhancing underwriting efficiency with a model can potentially help life insurers generate more applications, and place a higher fraction of those they do receive. In addition, the underwriting staff, which is becoming an increasingly scarce resource[4], will be better able to handle larger volumes as more routine work is being completed by the model.

We should emphasize that we do not propose predictive models as replacements for underwriters. Underwriters make indispensible contributions, most notably for applicants where medical tests are likely to reveal risk factors requiring careful consideration. Ideally, models could be used to identify the higher risk applicants early in the underwriting process, streamline the experience for more straightforward risks, and thus free up the underwriter's time for analysis of the complex risks. In addition, underwriters can and should provide insight during the construction, evaluation, and future refinements of predictive models. This is an oft overlooked but significant point. Particularly in complex domains such as insurance, superior models result when the analyst works in collaboration with the experts for whom the models are intended.

How exactly does the process work? The rough sequence is that the insurer receives an application, then a predictive model score is calculated, then a policy is either offered or sent through traditional underwriting. In more detail, the predictive model is typically used not to make the underwriting decisions, but rather to triage applications and suggest whether additional requirements are needed before making an offer. To that end, the model takes in information from any source that is available in near-real time for a given applicant. This can include third-party marketing data and more traditional underwriting data such as the application/tele-interview, MIB, MVR, and electronic prescription database records. For most insurers, this data can be obtained within two days of receiving the application[5].

We should point out one key change some insurers must endure. It is essential that producers do not order traditional requirements at the time an application is taken. If all requirements are ordered immediately at the application, eliminating them based upon model results is impossible. For some insurers, this is a major process change for the producer group.

After loading the necessary data for model inputs, the model algorithm runs and produces a score for the application. From here, several approaches can lead to an underwriting decision. One central issue insurers may wrestle with is how to use the model output when justifying an adverse action (i.e. not offering an individual applicant the lowest premium rate). Due to regulatory requirements and industry conventions, it is customary to explain to applicants and producers the specific reasons in cases where

[4] According to the Bureau of Labor Statistics 2010-2011 Occupational Outlook Handbook, despite reduced employment due to increased automation, the job outlook of insurance underwriters is classified as "good" because "the need to replace workers who retire or transfer to another occupation will create many job openings."

[5] Receiving the application is defined as when all application questions have been answered and/or the tele-interview has been conducted. If applicable, this includes the medical supplement portion of the application.

the best rates are not offered. It is possible to fashion a reason message algorithm that “decomposes” the model score into a set of intuitively meaningful messages that convey the various high-level factors pushing an individual score in a positive or negative direction. There is considerable latitude in the details of the reason message algorithm, as well as the wording of the messages themselves.

While allowing the model algorithm to place applicants in lower underwriting classes while delivering reason codes is a viable, given the novelty of using predictive modeling in underwriting, the approach life insurers have been most comfortable with thus far is underwriting triage. That is, allowing the model to judge which cases require further underwriting tests and analysis, and which can be issued immediately. From a business application perspective, the central model implementation question then becomes: what model score qualifies an applicant for the best underwriting class that would otherwise be available based upon existing underwriting guidelines? The information contained in the application and initial requirements will set a ceiling upon the best class available for that policy. For example, let us assume an insurer has set an underwriting criterion that says children of parents with heart disease cannot qualify for super preferred rates. Then for applicants that disclose parents with this condition on the application, a model can recommend an offer at preferred rates without taking the decisive step in the disqualification from super preferred.

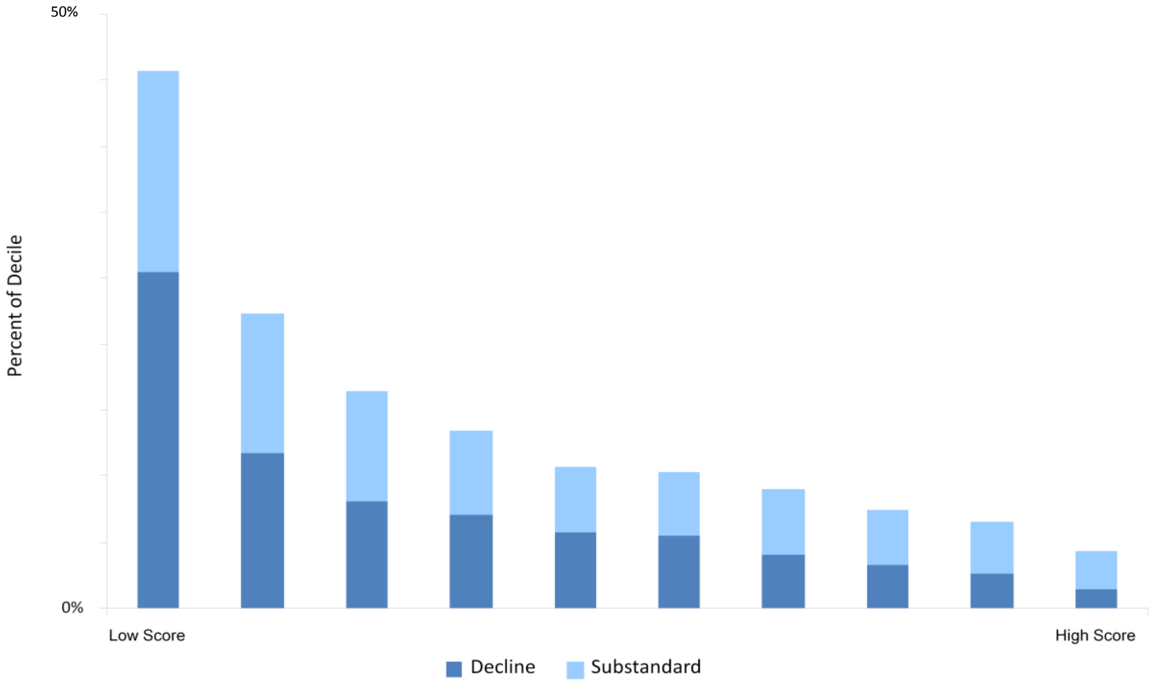
That is, the role of the model is to determine whether an applicant’s risk score is similar enough to other applicants who were offered preferred after full underwriting. If so, the insurer can offer preferred to this applicant knowing the chance that additional requirements will reveal grounds for a further downgrade (the protective value) will be too small to justify their cost. If the applicant’s risk score is not comparable to other preferred applicants, the insurer can continue with the traditional underwriting.

Marketing

In addition to making the underwriting process more efficient, modeling underwriting decisions can be of assistance in selling life insurance by identifying potential customers who are more likely to qualify for life insurance products. Marketing expenses are significant portions of life insurance company budgets, and utilizing them efficiently is a key operational strategy. For example, a company may have a pool of potential customers, but know little about their health risks at the individual level. Spreading the marketing efforts evenly over the pool will yield applicants with average health. However, this company could likely increase sales by focusing marketing resources on the most qualified customers.

The models supporting underwriting decisions that we have discussed thus far leverage both third-party marketing data and a limited set of traditional underwriting requirements. Alternatively, we can build predictive models using only the marketing data. While these marketing models do not deliver the same predictive power as those that utilize traditional underwriting data, they still segment risks well enough to inform direct marketing campaigns. Scoring the entire marketing pool and employing a targeted approach should help reduce the dollars spent marketing to those who will later be declined or less likely to accept an expensive offer, and result in an applicant pool that contains more healthy lives.

Figure 2: Marketing Model Segmentation



Like Figure 1, risks which had been underwritten by the insurer and kept in a holdout sample were rank-ordered by model score (using third-party data only) and divided into equal-sized deciles. However, this graph shows fractions of those deciles which contain declined or substandard applicants.

In addition to general target marketing efforts, models of underwriting decisions can also serve more specific sales campaigns. For example, multiline insurers, or even broader financial institutions often attempt to increase sales by cross-marketing life products to existing customers. However, they run the risk of alienating a current customer if the post-underwriting offer is worse than what the marketing suggests. Instead of selling an additional product, the company may be at risk of losing the customer. In dealing with this challenge, predictive modeling can be used to conduct an initial review of the customer pool and assist in determining which existing customers should receive offers for life insurance.

Predictive modeling can also aid in targeting specific products to the markets for which they were designed. For example, a given company may sell a product with preferred rates that are competitive, but standard rates that are less attractive. Other products may contain incentives for the insured to maintain healthy lifestyle. To whom might these products appeal? A person who the model indicates is currently living a healthy lifestyle is a prime target for such marketing programs.

In-Force Management

It is well known that the effects of underwriting wear off over time. Lives that were initially healthy may have degraded, and people who appeared to be poor risks initially may have improved. Products are priced to anticipate a reversion to mean health risk, but considerable variation in the health status of in-force policyholders will both remain and be unobservable without new underwriting. While full underwriting is cost prohibitive in these situations, a predictive model could be an inexpensive and transparent alternative. Scoring the in-force block could provide more insight to emerging mortality experience, inform changes to nonguaranteed policy features, help insurers know where to focus efforts to retain policyholders, and guide both direct writers and reinsurers in block transactions.

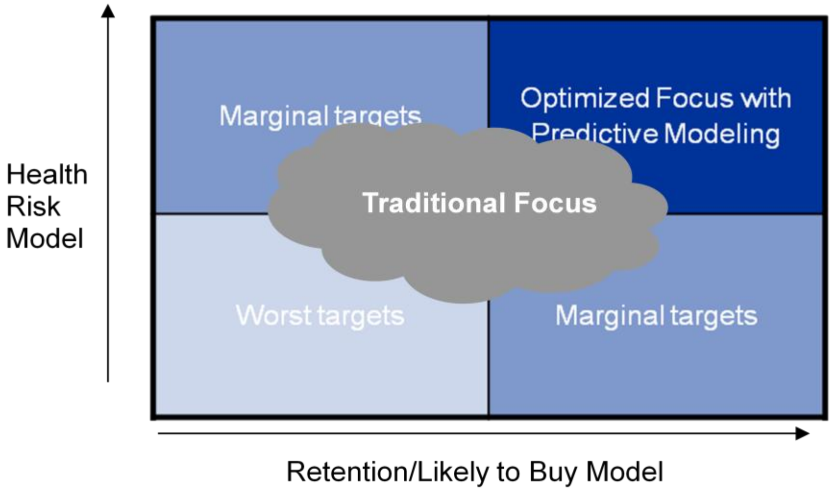
Additional Predictive Model Applications

We have focused our discussion on modeling health risk for life insurers because it is arguably the latest advancement, but there are many other areas of uncertainty for life insurers where a predictive model could reveal valuable information. We will present several potential applications in brief.

Analogous to models used to market consumer products, predictive algorithms can also estimate how interested a potential customer would be in purchasing a product from a life insurance company. Insurance customers are often relatively affluent, or have recently undergone life-changing events such as getting married, having children, or purchasing a house. All of these traits and events (among other characteristics) can be identified in the marketing data. More specifically, a predictive model can be built to identify which characteristics are most highly correlated with the purchase of life insurance. Again, scoring a direct marketing database can help a life insurer determine where to focus limited resources for marketing and sales.

We have discussed retention in terms of which customers an insurer would most like to keep, but equally important is which customers are most likely to leave. As many of the same life event and lifestyle indicators in the marketing data communicate when a person is likely to purchase a product, they also hint when a person is likely to surrender a product. In addition to third-party data, insurers also can see indicators of impending surrenders in transactional data such as how policyholders are paying premiums (automatic bank debits vs. having to physically write each year, or month), whether a policyholder is taking policy loans, whether they are calling the home office asking for cash values, account balances, and in-force illustrations, etc. Since neither producers nor the home office can give complete attention to each policyholder, a predictive model can sort these different indicators and help prioritize where to focus policy-saving effort.

Predictive modeling becomes even more powerful when models are used in combination. Not only can they answer who is most likely to purchase or surrender, but they can simultaneously identify the customers the company would most like to insure. Layering the underwriting health risk model on top of either the purchase or surrender models will tell the insurer which quadrant of the population will likely yield the highest return.



A final application we will mention is workforce analytics. Becoming a successful life insurance agent is notoriously difficult. The home office spends significant resources recruiting and training these agents, and the high turnover rate is a considerable drain. Predictive models can be used to help improve the efficiency of agent recruiting by scoring applicants on the basis of how similar their profile is to that of a company's existing successful field force. Such a tool can help prioritize which applicants to pursue.

When considering all the potential applications for predictive modeling in life insurance, it becomes apparent that analytics is truly an enterprise capability rather than a niche solution. After an insurer begins with one application that proves successful, the next application follows more easily than the first. Expertise, data, and infrastructure can be leveraged throughout the organization, but more importantly, decision makers come to realize and respect the power of predictive modeling.

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.

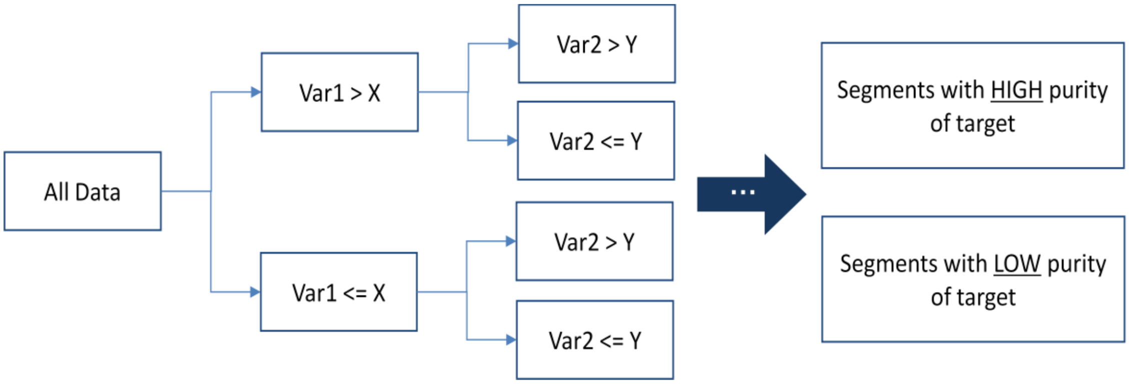
The model is first built using the training data, but modelers are also concerned about fitting the model too closely to the idiosyncratic features of one sample of data. The initial model is adjusted using the validation data in order to make it more general. Each set is only used once in the modeling 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 indentified 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.

Figure 3: Graphical Representation of Decision Tree


All Data
Var1 > X
Var1 <= X
Var2 > Y
Var2 <= Y
Var2 > Y
Var2 <= Y
...
Segments with HIGH purity of target
Segments with LOW purity of target

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 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[6]. 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

[6] 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).

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.

This publication contains general information only and is based on the experiences and research of Deloitte practitioners. Deloitte is not, by means of this publication, rendering business, financial, investment, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte, its affiliates, and related entities shall not be responsible for any loss sustained by any person who relies on this publication.

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