



Analytics 2.0



Marlowe Leibensperger
Senior Vice President,
EagleEye Analytics

Machine learning analytic methods, when combined with traditional actuarial and statistical methods, can help pick out the truly predictive data points and correlations.

The property and casualty insurance industry's underlying analytics, mathematics and statistics have been stable for some time, denying individual companies the opportunity to deploy significant commercial advantage based on analytics. More recently, an important series of technological breakthroughs in the analytics field known as machine learning have gained substantial attention. These technological advances, when applied to insurance data, can enable insurers to establish commercial advantages across their operations — from more sophisticated rating and underwriting (including the replacement of credit scores) to improved customer conversion/retention and the more effective optimization of insurance customer portfolios.

UNIQUE ANALYTICAL CHALLENGES

Insurance is a special case in terms of data and data analysis, very different from typical datasets encountered by mathematicians and statisticians. By its very nature, insurance data is extremely noisy and variable: claims frequency is relatively low and claims severity is volatile, and both metrics are variable by insurance line of busi-

ness. The concept of variable exposures introduces further complexity. As opposed to nearly all other non-insurance data, where data points are each identical in their weights, insurance data yields data points with one day, a few days, one year, etc. as the exposure basis. These are a few main factors differentiating insurance data from the conditions found in traditional datasets used to develop relevant mathematical and statistical methodologies.

Furthermore, insurance has a set of issues that make deployment of advanced analytics a required competence for sustained profitability. Most notably, the ultimate cost of goods sold on any particular policy is not known at the time of sale. Insurance pricing is an exercise in predicting the future. It involves building predictive models that leverage past history. Information known at the time of the sale is used to predict the probability of the insured experiencing a loss, as well as the likely size of that loss. Since no past individual policy truly represents the probable cost of the policy being sold, past policies are grouped together to make credible forecasts of cost and price.

CURRENT ANALYTICAL STANDARD

Traditional insurance analytics based on actuarial science is based primarily on a range of linear models, with the more advanced linear model being the Generalized Linear Model (GLM-1975). From a traditional statistical perspective, these methods provide a solid baseline approach, providing a powerful and flexible

framework for understanding the predictive power of multiple variables and an output form convenient to forming new or adjusting existing rate plans.

These methods are limiting, however, because insurance data exhibits non-linear effects that have been shown to be very significant. Additionally, the nature of insurance data requires the application of assumptions regarding error functions, typically Poisson (claims frequency) and Gamma (claims severity), which have been proven to be imperfect when applied to fitting insurance data. This is particularly true in claims severity modeling. In the absence of something better, these assumptions deliver a useable result, but the real distributions are often different from model assumptions.

These challenges, individually and collectively, present an opportunity to derive new, complementary approaches to augment model development and performance.

Raising the Analytical Bar

In contrast to traditional analytical techniques, machine learning makes no assumptions about linearity or the shape of the underlying error distributions. Machine learning searches the solution space of the data and lets the data speak for itself. The results derived from machine learning methods are a pure, unadulterated representation of the predictive signal.

Like traditional methods, machine learning methods were developed outside the insurance industry; in their raw form, they suffer a number of the same challenges presented by insurance data. Illustrative of this fact, most machine learning methods have been publicly available over the past couple of decades and have had a minimal impact in the insurance industry.

In their raw form, machine learning methods are generally difficult to apply to insurance problems effectively. To apply machine learning methods to insurance data efficiently, substantial research and development is necessary — including the fusion of machine learning and traditional actuar-

ial and statistical methods. Once the problems associated with the matching of base machine learning methods and the complexity of insurance data have been addressed, substantial levels of predictive signal can be extracted from insurance data above and beyond the level achieved using current methods.

Fusion machine learning methods op-

erate on both classification and regression problems, allowing insurance business operations to exploit this additional signal in a number of critical areas. The main benefits from the new fusion methods are found in the form of superior model performance (lift/consistency), improved analytical productivity/speed to results and enhanced ease of implementation.

Superior model performance

Underwriting insurance policies is a very complex task at the best of times. For some time, underwriting has become much more difficult, since the linear methods have established rates in such a way that the signal from individual variables has been captured in rates. (For example, looking at any single variable such as the age of driver, the underwriter is able to accept all risks, since the rates compensate for the appropriate level of risk.) This is also becoming true of a number of two-way interactions such as sex and marital status in auto insurance. If the single one-way and two-way interactions are successfully addressed in rates, how does the personal lines underwriter discover and discriminate good and bad risks?

The answer is to use fusion-based machine learning methods that break portfolios into a controllable number of subdivisions based on variations in rate adequacy (loss ratio) with each subdivision described by a simple intersection of variables. For example, a subdivision might include an intersection of variables such as a vehicle older than 1998, combined with a rated driver male over 35 and a minimum operator age of 16 to 19. This approach has proven to uncover significant rate subsidies in current classification plans, ranging from -50% to 100% or more.

Applying these methods to insurance portfolios will provide the underwriter with new, empirically driven knowledge, promoting the acceptance of previously unidentified good risks and the rejection of previously accepted poor risks. Tools like this enable underwriters, in combination with their own valuable and specific knowledge, to improve their performance.

In many North American insurance markets, the continuing use of credit score as an underwriting or rating variable is under scrutiny. Given that credit score has such a strong effect when applied using current methods, the possibility of its disappearance is distressing to many companies. For many others, examining the removal of credit score from underwriting or rating is on the

forefront of their corporate agendas as they look to take social or public policy stands, or eliminate operational costs or procedures, related to consumer consent. Credit score is very effective due to the stability of the underlying methods used in underwriting or rating, with much of the recent success coming from the introduction of new variables.

The introduction of fusion-based machine learning methods has dramatically altered this scenario: the new methods can deliver more predictive signal using fewer variables. For example, recent research proved the fusion methods delivered a stronger signal



using only six of an insurer's existing rating variables (excluding credit score) compared to traditional methods using 15 variables including credit score. Additionally, the newly derived model was found to be completely uncorrelated with credit score, removing the concerns of regulators.

Fusion methods have consistently demonstrated an ability to find signal that improves the performance of the underlying model — demonstrating amplified dispersion, improved strength of fit relative to the historical data and better generalization of results on unseen data.

Increased analytical productivity

A key factor in developing and deploying any commercial strategy is the time it takes to complete the process. The application of fusion machine learning methods — from data processing, model development and subsequent de-

ployment of a solution — takes a fraction of the time taken to develop and deploy traditional methods. This shortened timeline yields a reduction of cost, both in terms of research costs and the opportunity cost of realizing benefits. This accelerated result is achieved by the methods' ability to handle data with minimum manipulation, as well as requiring no underlying assumptions on the distributions of the data. Flowing from these points, a large amount of manual, iterative work is removed from the modeling process. This far reduces the time and effort taken to execute these fusion machine learning models successfully when compared to the resources consumed using traditional methods.

The use of fusion methods has been proven to accelerate the socialization, adoption and realization of benefits of advanced analytics within insurance carriers, all the while increasing an insurer's analytical productivity and ability to investigate a number of currently unaddressed issues.

Enhanced ease of implementation

Traditionally, operational departments have had issues with analytics in terms of the clarity (or incomprehensibility) of results. Market research points to a sizable chasm between modeling units and business folks, each speaking different languages related to statistically valid and useable business results. The introduction of a wide range of new methods allows these issues to be addressed in many circumstances. The simplest example is the previously discussed underwriting example, in which an unprofitable niche was detected and described to the underwriter with a commercially definable result.

Results from fusion methods have been proven to enhance the dialogue regarding analytical results among insurance company stakeholders, including internal (i.e., non-modeling business units) and external (i.e., agents/brokers) parties. This transparency has resulted in improved collective buy-in and support, ultimately positioning the associated actions for greater acceptance and probability for success. ≡