SOCIETY OF ACTUARIES®

May 2019

Risks Embedded in Predictive Models

Call for Essays

SOCIETY OF ACTUARIES®

Contents

Introduction

The Predictive Analytics and Futurism and Technology Sections of the Society of Actuaries (SOA), along with the Joint Risk Management Section of the SOA, the Casualty Actuarial Society, and the Canadian Institute of Actuaries, sponsored a call for essays on the theme "Risks Posed by Predictive Models."

Participants were asked to discuss the risks and consequences arising from the use of predictive models. For this contest, an essay was understood to be a short nonfiction form of writing expressing the often subjective opinion of the author.

The call for essays received eight submissions, and the three sections would like to thank all the authors who participated in the contest:

Best Paper

Ratemaking Reformed: The Future of Actuarial Indications in the Wake of Predictive Analytics
Gyasi Dapaa

Best Eligible Paper

Predictiveness vs. Interpretability
Kimberly Steiner and Boyang Meng

Second-best Eligible Paper

Actuarial Fairness in the Era of Machine Learning
Marjorie A. Rosenberg

Unintended Consequences: The Risks Posed by Predictive Analytics
Greg Fann and Kaitlyn Rachelle Fleigle

Risks of Using Predictive Models
John Hegstrom

Risks From Futurism and Ethics in Predictive Modeling
Cameron Rose

A Reality Check for Predictive Modeling
John Wallentine

Reflections on the Day the Models Died
Jim Weiss

Gyasi Dapaa submitted the best paper, titled "Ratemaking Reformed: The Future of Actuarial Indications in the Wake of Predictive Analytics," and is commended for its high quality. However, the paper was not eligible to receive a monetary prize, as it exceeded the length requirement.

Kimberly Steiner and Boyang Meng won the $1,000 cash prize for the paper "Predictiveness vs. Interpretability." Marjorie A. Rosenberg won the $500 cash prize for the paper "Actuarial Fairness in the Era of Machine Learning."

This essay collection contains the three winning papers, which express the opinions and thoughts of a number of authors on the subject. Dapaa's paper has been abridged for inclusion in this collection.

The thoughts and insights shared herein are not necessarily those of the Society of Actuaries, the Casualty Actuarial Society, the Canadian Institute of Actuaries or the corresponding employers of the essayists.

Ratemaking Reformed

The Future of Actuarial Indications in the Wake of Predictive Analytics

Gyasi Dapaa

The journey of achieving the actuarial aspiration of equitable pricing through increased segmentation has not been without twists and turns. It started with one premium for all; moved to meager segmentation with univariate analysis and later with Bailey's minimum bias (BMB) procedure; and finally to generalized linear models (GLMs), which allow us to segment along limitless dimensions. However, despite the capacity of GLMs to determine both the base rate and the policy risk relativity, they are currently used to determine only the latter. This article argues for a necessary cultural change that will enable actuaries to advance pricing excellence and also capitalize on the exploding reserves of data available to them.

Introduction to Ratemaking

Insurance premium for a future year is a product of three factors: current base rate, proposed rate change and risk score relativity. The process of determining insurance premiums is called ratemaking. The base rate ensures that adequate premium is collected on an aggregate basis to cover insurance claims, claim adjustment expenses, underwriting expenses and the targeted profit provision. The proposed rate change ensures that all future dynamics that will affect rate adequacy (such as changes in business mix, changes in tort laws and inflation of insurance commodities, just to mention a few) are accordingly accounted for in future premiums. The exercise of determining the needed rate changes for a future year is called indications. The risk score relativity factor ensures that premiums are actuarially fair—that is, the higher the risk, the higher the premium.

Currently, actuaries obtain proposed rate changes from indications and risk score relativity from GLMs. I argue in this paper how the GLM already contemplates all three factors and is the best machine for them.

Indications: Why It is What It is Today

The current indication process was produced for expediency but not necessarily for excellence. I'll explain. Before the advent of cheap data storages, and hence large databases of granular risk information, actuaries had only aggregate historic claims data to work with. Because the claims data contained claims that were still open, they projected them to their expected ultimate values using a technical procedure called loss development. Also, because insurance losses and expenses change with time due to changes in business mix, technology, insurance commodity prices and tort laws, actuaries trended the historic losses to the future period in which the rates will be implemented. These two adjustments allowed actuaries to derive proposed overall rates (i.e., base rate and proposed rate change) but not the risk score relativity.

The inability to price at the policy level had been the predicament of the actuary for a long time, until perhaps memory became cheap enough, thanks to Moore's law, to allow insurance companies to store dimensional data. However, when actuaries got hold of granular risk data, they unfortunately had no sophisticated methodology to derive risk relativity factors. They therefore started with an approach whereby risk relativities were univariately derived. The bane was that the univariately derived risk relativities also contemplated their correlated effects with each other; hence, their product amounted to double counting of effects and biases in the risk estimate!

Along came the Bailey minimum bias procedure in the 1960s. BMB is a multivariate and more accurate approach but suffers two disadvantages: It is computationally restrictive, and it produces relativities but not base rates. This therefore cemented the ratemaking tradition of manufacturing base rates and relativities in different shops.

In their search for more conducive methods, the actuarial community chanced upon GLM, a statistical methodology that has long been developed and known in the academic world. The GLM has many merits over BMB. It allows actuaries to derive relativities for countless numbers of variables. It affords them the flexibility to model different distributions of insurance losses. Given the varying distributional forms of insurance risk metrics—severity, frequency and purepremiums—this flexibility is not taken lightly. It also allows actuaries to choose the functional form (such as identity, log or power function) of the relationship between the risk measure being modeled and the relativity variables under consideration; and above all, actuaries are able to assess whether their estimated risk relativities are signal or noise using a prolific number of model diagnostic measures such as standard errors, chi-squared statistics, archaic information criterion, F-statistics and many others.

The GLM has one more edge that is far more underutilized than the ones aforementioned: Aside from forging risk relativities, it can also predict the best (minimum mean squared error) loss cost estimate for each insured unit for any exposure period with a greater capacity for segmentation, greater accuracy and lesser effort. This means that the tradition of deriving an overall base rate and the policy risk relativity score separately has to be replaced with a fresher and more powerful culture of directly predicting each policy's loss costs with a GLM. In the remainder of the paper, I argue how GLMs accommodate each of the three main features in traditional ratemaking (base rate, rate changes and risk relativity) and propose a new framework for actuarial indications.[1]

Actuarial Indications are Already Contemplated in GLMs

OVERALL (BASE) RATES

The intercept term in a GLM measures the overall rate level; it can be varied by any dimension the actuary desires: region, state, industry or any broader category. In fact, as with all GLM estimates, it has desirable statistical properties. It is one of, if not *the* most statistically efficient (lowest variance) estimators of the base rate. As a maximum likelihood estimate, it achieves the Cramer Rao lowest bound on variance. It is fair to point out that, in traditional ratemaking, it's not typical to assess the variability of the actuarial base rates, and all are thus banked on the pricing actuary's ability to instinctively determine whether his or her estimate of base rate is noise or signal, a test that even experts steeped in statistics have often failed. (See page 113 of Kahneman's revolutionary book, *Thinking, Fast and Slow*.) However, GLMs force actuaries to know the variability and statistical significance of all of their estimated parameters, including the base rate (i.e., the intercept).

The other statistical benefit of a GLM estimate is that it is consistent (i.e., approaches the true value with enough data) at worst and unbiased at best. Unfortunately, this cannot be said about the actuarial base rate. In fact, because it is derived outside the GLM but combined with risk relativities carved from GLMs, it's likely to pick up effects already contemplated in the GLM, and hence is biased. Suppose a Texas automobile book of business has a disproportionate number of reckless drivers. In the current ratemaking culture, reckless drivers in Texas will be double penalized, one through the actuarially derived Texas base rate and the other through the GLM risk relativity for reckless driving possibly captured by motor vehicle records.

There is, however, a silver lining actuaries may tout in an attempt to save the current system: For base rates of smaller states, an actuary can use credibility analysis to combine the unstable experience of the smaller state with a more stable complement (say, the countrywide base rate) to derive desirably stable base rates. Although this is valid, there are GLM variants, such as generalized linear mixed models (GLMMs), that allow for the sort of credibility

[1] Many other equally viable statistical methods such as classification trees, random forests and neural networks are available for actuaries to use. However, we will continue to use GLM (because of its popularity) to loosely represent all such statistical methods.

weighting done in an actuarial analysis. See Klinker (2011) for an exposition of actuarial application of GLMM and its similarities with Buhlmann credibility. Therefore, there is no good reason, at least known to me as of this writing, for actuaries to derive base rates outside of GLMs.

PROPOSED RATE CHANGES

Current rates cease to be adequate in the future for three main temporal changes: (1) general market factors (technology, tort laws, prices, etc.), (2) business mix and (3) the relationship between losses and risk variables. Points 1 and 2 are easily accommodated in a GLM by including an econometric trend term and risk attributes in a predictive model. The coefficient of the trend term measures how premium is expected to change with time, while those of the risk attributes measure how premiums change with differing risk characteristics. Point 3 is checked by regular updates of the pricing models.

RISK RELATIVITY SCORE

Although actuaries get the risk relativities from a GLM (and so are efficient estimates), how they use them in pricing mitigates their statistical merits. I will describe one such misuse. Most pricing actuaries would multiply the relativities together to get a predicted risk estimate. After this, they would—here comes the first unforced spoiler—partition this product into a number of risk groups. After that, they would map each risk group to a risk score factor, and that becomes the policy factor that gets multiplied by the base rate to get proposed premium. Meanwhile, the predicted policy risk estimate—say, purepremium—obtained directly from the GLM has been proven to be the best estimate of the policy's risk exposure. And therefore, every tweak, apart from consuming time, unnecessarily chips away chunks and chunks of its statistical efficacy.

Other Considerations: Ratemaking for Exceptionally Large Risks

Exceptionally large risks may need special attention even if a GLM is available. This is because they are normally so large and few that the GLM is not able to adequately fit their heterogeneous risk features. Actuaries can use specialized rating techniques such as experience rating, schedule rating, composite rating and retrospective rating to complement the manual estimate obtained from the GLM.

THE FUTURE OF INDICATIONS

The GLM estimate should be the pinnacle of, but not a mere input for, proposed policy premiums. Actuaries should find few, if any, reasons to do any analysis outside of it, such as a derivation of a base rate or rate change. It is, if appropriately parameterized and estimated, actuaries' most accurate (least bias and variance) measure of risk that can be carved from historical data. It can also contemplate most of the technical dynamics that are important in insurance pricing and to actuaries, for that matter: credibility, trends, interactions and experience rating, just to mention a handful. The indications process should occur entirely within a GLM framework, with minor episodes involving merely a refitting of current models with new data, and major ones being a new development of the latest and greatest predictive models.

This proposed framework, of course, preserves actuaries' freedom to exercise their judgment to select away from the GLM factors. However, they must exercise this freedom cautiously so as not to weaken the efficacy of the GLM predictions. Two such legitimate justifications for actuarial deviation from indicated GLM estimates are when there is a business motivation or information not fully reflected in the historical data being used for modeling. For instance, there are cases when it has been strategically decided by business leaders to invest in or disinvest from a market segment; or we may even want to combine the supply (loss cost) factors obtained from GLMs with demand factors to drive a pricing strategy that achieves a targeted financial outcome—pricing optimization. For such economic motivations, actuaries can pick model parameters different from the GLMs. There can also be situations in which actuaries can be privy to future information that is correlated with insurance losses. For instance, actuaries can know of the timing of a new technology that mitigates risk or saves lives, or a new tort that changes the cost of insurance. In this case, too, actuaries can adjust the GLM parameters to reflect their more informed future expectations.

The merits of this new ratemaking system are manifold and consequential:

- **Improves pricing precision.** Pricing precision is one of the most important tenets for any insurance book of business. This is because of the leveraged effect an otherwise small pricing error has on insurance profits: A 1 percent pricing undercharge can eat away as much as 10 percent of profits when underwriting profit provision is 10 percent, and even higher proportions of it under lower profit provisions. A second reason is that pricing precision serves as a guard against the vulnerability to adverse selection, which has the ability to spiral an insurance company out of business. Last, it can impede the growth of a book of business, as systematic overcharges can drive away otherwise profitable business.

- **Saves time and resources.** The current ratemaking methodology involves obtaining relativities from GLMs and base rates outside the GLM framework. Our proposed methodology obtains the base rate and the relativities from the GLMs, and hence saves time and resources without sacrificing accuracy. An actuary with an endless list of responsibilities will appreciate having more time to spend on other bodies of work.

- **No transition costs.** There is no transition cost, as GLMs are already familiar to actuaries and regulators.

- **Big-data friendly.** We're in a defining revolutionary moment in which insurance companies are receiving unthinkable volumes of data from their insured risks, thanks to advances in telematics and the Internet of Things. With this privilege comes the competitive pressure of using every bit of this big data to help paint a coherent picture about risks. All are up for grabs, but the winner will no doubt be the one who leverages the power of machines to process this ceaseless data endowment to understand and write risks profitably. Although GLMs can be designed, fine-tuned and automated to handle such humongous data assets, no human mind, no matter how astute and sharp, can keep up with the processing demands of big data. Hence, transitioning from a framework that relies heavily on human intervention to one that relies minimally on it is proactive!

Bibliography

Anderson, Duncan, Sholom Feldblum, Claudine Modlin, Doris Shirmacher, Ernesto Shirmacher, and Thandi Neeza. 2004. *A Practitioner's Guide to Generalized Linear Models*. CAS Discussion Paper Program.

Bailey, Robert A., and Leroy J. Simone. 1960. Two Studies in Automobile Insurance. *Proceedings* 47, part I.

Klinker, Fred. 2011. Generalized Linear Mixed Models for Ratemaking: A Means of Introducing Credibility into a Generalized Linear Model Setting. *CAS E-Forum*, vol. 2. Arlington, Va.: Casualty Actuarial Society.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus, and Giroux. 113.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

Werner, Geoff, and Claudine Modlin. 2016. *Basic Ratemaking*. Arlington, Va.: Casualty Actuarial Society.

Gyasi Dapaa is a senior manager at Ameriprise Financial (Auto & Home). He can be reached at *gkd5r@virginia.edu*.

Predictiveness vs. Interpretability

Kimberly Steiner and Boyang Meng

A common criterion for the selection of predictive models is predictiveness: one model is considered better than another if it gives more accurate predictions of the outcomes of unknown events. Apart from making intuitive sense, this criterion is attractive because there are measures available (e.g., Gini coefficient, R^2) that allow us to easily rank models by predictiveness. This paper demonstrates that relying on predictiveness alone can result in choosing a model that exhibits behavior that may not be intuitive. It also demonstrates that this unintuitive behavior may not be immediately obvious.

In this paper, we compare two kinds of predictive models, built using the same data, on the criteria of predictiveness *and* interpretability, in the context of life insurance mortality. The two types of models compared are generalized linear models (GLMs) and gradient boosting machines (GBMs). We demonstrate, using a double lift chart on holdout data, that a GBM can give better predictions than a GLM. We also demonstrate that while GLMs are easy to interpret, GBMs can be difficult to interpret, in the sense that profiles that are similar can have very different, and sometimes unintuitive, behaviors.

In conclusion, we emphasize that the desired attributes of a predictive model must be taken into account when determining what type to use, and we discuss some implications for the wider use of machine learning techniques in the insurance industry. We do not dispute the importance of predictiveness. However, we do argue that depending on the context, interpretability is an important consideration, and that, in some contexts, interpretability should not be sacrificed for predictiveness.

This paper is organized into the following sections:

- Predictive Models Considered: General remarks on GLMs and GBMs
- Data Used: Details of the data used for this study
- Details of the Models: Details of the actual models' fit
- Predictiveness: A comparison of the predictiveness of the models
- Interpretability: Discussion of the interpretability of results
- Conclusion: Discussion of these results and some consequences in the context of life insurance, as well as some possible directions for further study

Predictive Models Considered

This section includes a high-level description of GLMs and GBMs. Further details can be found in the predictive analytics literature.

The types of models we chose to compare in this study were generalized linear models and gradient boosting machines. GLMs have been widely used in property and casualty insurance for decades for pricing purposes and have been increasingly used in recent years in life insurance for experience studies. GBMs are a trendy machine learning technique becoming more widely used in many sectors. Models involving the use of GBMs are frequent winners of predictive analytics contests such as Kaggle (*www.kaggle.com*), which determines winners solely based on the Gini coefficient (i.e., a measure of predictiveness is the only consideration).

GENERALIZED LINEAR MODELS

GLMs are a generalization of ordinary least squares regression. They are characterized by the selection of an error structure, which comes from the exponential family of distributions (this includes normal, Poisson, Gamma and binomial distributions), and a link function, the inverse of which relates the linear predictor (the linear combination of features included in the model) to the response or independent variable. Common link functions are the identity, log and logit functions. Features are selected using a combination of statistics, heuristics and judgment. Each feature

has a parameter associated with it, and model-fitted values are calculated by summing parameters of the appropriate features and applying the inverse of the link function.

GRADIENT BOOSTING MACHINES

Gradient boosting involves fitting a model on a randomly selected subset of the data, calculating the ratio between some proportion of the predictions of the previous model and the response on another random subset, fitting another model of that ratio and continuing the process unless some convergence criterion is reached. The model is selected by determining combinations of parameters such as the proportion of data included in each sample, the proportion of predictors available in each model and the proportion of the previous model predictions used at each step (the learning rate), as well as the characteristics of the underlying model. The underlying model is often a classification or regression tree. In this case, the final model is a weighted sum of a (potentially large) number of trees.

Data Used

This study used single life mortality experience data provided by 23 companies for Willis Towers Watson's TOAMS4. The data include $25 trillion face amount of exposure over the four-year study period (calendar years 2011–2015), representing over 123 million policy years of exposure. More than 1.5 million death claims, corresponding to $82 billion, are included in the data. The data were split randomly into training and testing data. Both models were trained on the same training data and compared on the same testing data.

Details of the Models

GENERALIZED LINEAR MODEL

The GLM used a log link function and Poisson error structure. Attained age, issue age and duration were included as polynomials. The model included many interactions, including between categorical variables and polynomials (e.g., smoking status and duration or attained age and gender) and between combinations of polynomials (e.g., between duration and issue age). Categorical variables were grouped as necessary.

GRADIENT BOOSTING MACHINE

The response GBM was assumed to be distributed Poisson. Attained age, issue age and duration were included as continuous variables. Different groupings of categorical variables were experimented with. Hyperparameters were optimized using a grid search and cross-validation on a random split of the training data with four levels.

Predictiveness

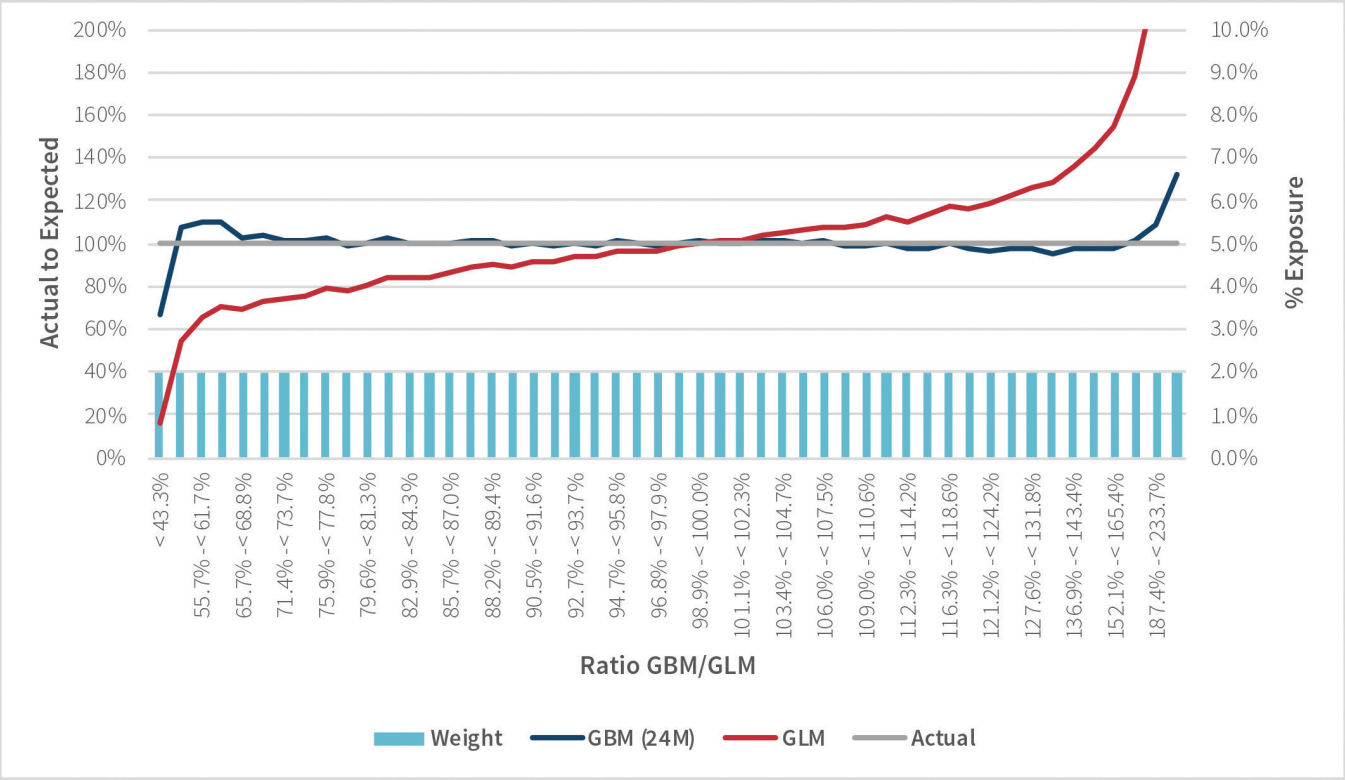
Double lift charts are commonly used to compare predictiveness of two different models. A double lift chart is created as follows:

- For each observation in the testing data, predictions according to each model are calculated.
- The ratio of predictions is calculated for each observation, and the observations are ranked according to this ratio from low to high and segmented into a number of bands (we used 50) of approximately equal exposure.
- In each band, each average model prediction is calculated and divided by the observed (i.e., actual) mortality in that band.

A double lift chart is effectively an actual vs. expected analysis by discrepancies between predictions in a pair of models. Where the model predictions are different, meaning where the ratio is high or low (i.e., in the extreme left and right of the graph), the model that gives better predictions is that for which the actual vs. expected is closer to 1.

To compare the predictiveness of the GLM and GBM, we used a double lift chart on the testing data as shown in Figure 1.

Figure 1 Double Lift Rescaled



According to the double lift chart, the GBM was clearly more predictive than the GLM.

Interpretability

As stated earlier, for a GLM, predicted values are determined by calculating a sum of parameters of the appropriate features and applying the inverse of the link function. In the case of a log link function, this is equivalent to multiplying the exponentials of the model parameters; that is, the model is multiplicative. This allows us to have a complete and interpretable understanding of the variables and combinations of variables driving estimates of mortality and the quantitative impact of each. It also allows us to make statements like, "In segment x, mortality is *y* percent higher than in segment z."

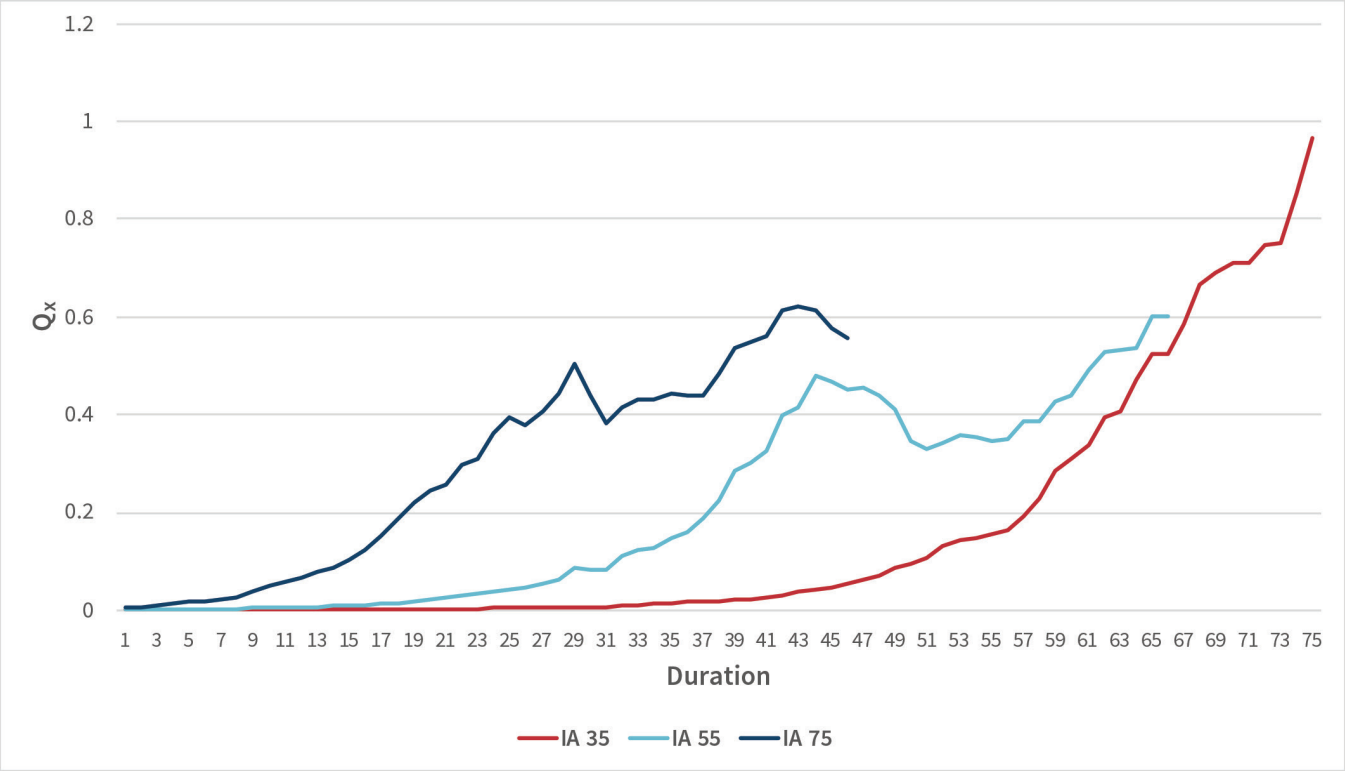
As previously stated, a GBM is a weighted sum of (an often-large number of often tree-based) models. There is no practical way to extract an interpretable characterization of the model predictions. Techniques (e.g., partial dependency plots) do exist that allow a general understanding of drivers of the model, but because of the nature of the model, it is possible for predictions associated with sets of observations to differ in unexpected ways. We illustrate this using several examples. The examples were created by:

- preparing profiles corresponding to different combinations of policy characteristics, including sex, smoking status, underwriting class, face amount, product and issue age;
- for each profile, creating observations corresponding to different durations; and
- calculating the GBM prediction on each observation for each profile.

MORTALITY BY DURATION FOR SELECTED PROFILE

In this example, we used male, nonsmoker, residual standard, face amount band $500,000-$600,000, current assumption universal life with level risk amount (ULNG). We compare the q_x by duration for selected issue ages (Figure 2).

Figure 2 Q_x for Selected Profile

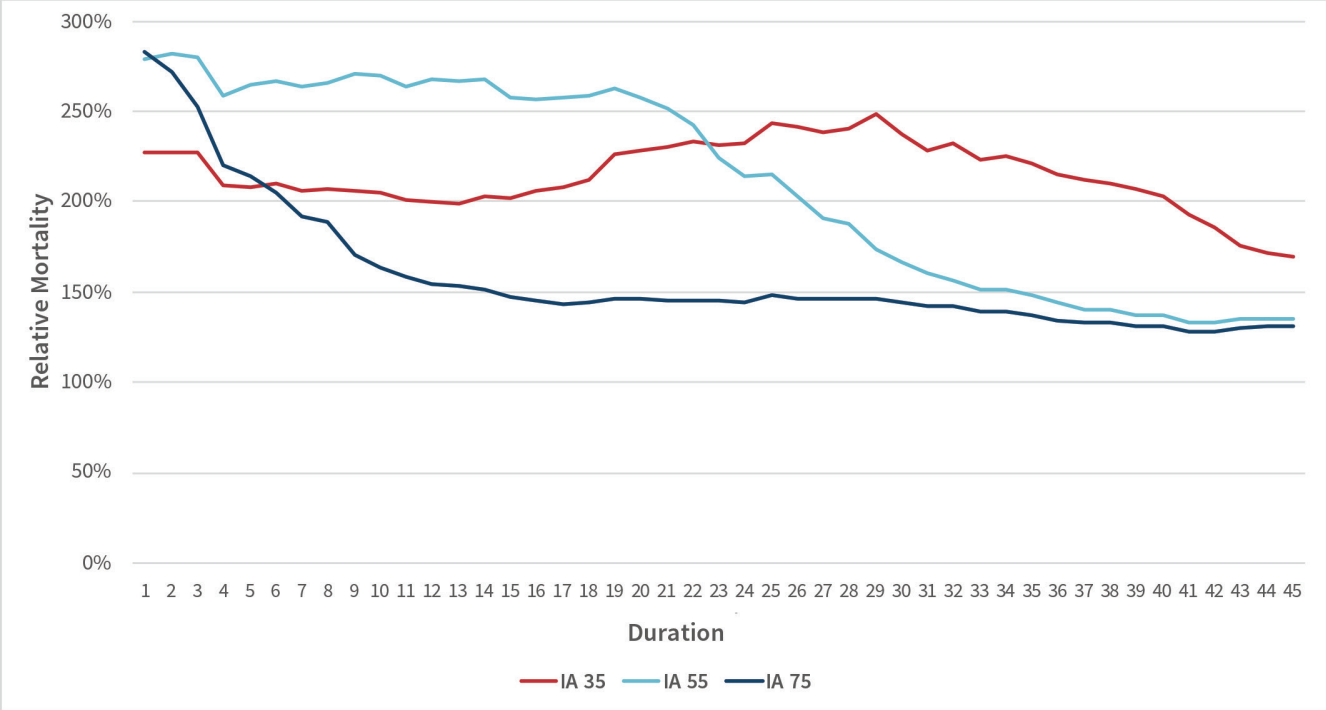


We note that the q_x pattern for issue age 35 is monotonic and might be considered reasonable for all durations, whereas for higher issue ages the pattern breaks down (mortality decreases in certain durations compared to the prior duration) at higher attained ages that lack credibility. While this is not surprising, the duration at which the pattern breaks down will vary by profile, and the only way to determine the point at which it breaks down is to evaluate the curve for all required profiles, of which there may be a very large number. While GLMs also struggle where credibility is lacking, we can identify and understand exactly how they are lacking.

SMOKER RELATIVE TO NONSMOKER MORTALITY BY DURATION FOR SELECTED PROFILE

In this example, we used male, residual standard, face amount band of $500,000–$600,000, male universal life (level net amount at risk), ULNG. We compare the ratio of smoker to nonsmoker mortality by duration for selected issue ages (Figure 3).

Figure 3 Smoker Relative to Nonsmoker for Selected Profile

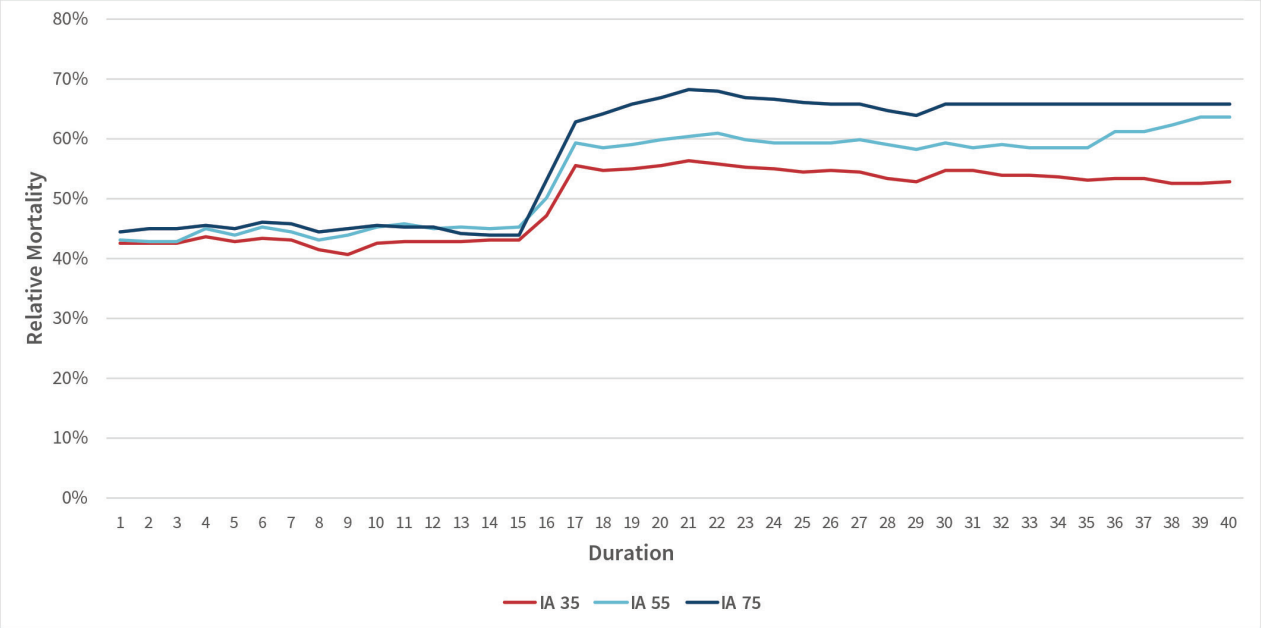


We note that even for combinations of issue age and duration where exposure is high, the ratio between smoker and nonsmoker qx can exhibit patterns, including zigzags, for which there is no obvious explanation. We also note that these patterns can be different for all possible profiles. By way of contrast, GLMs allow a complete understanding of patterns describing relative levels of predictions (i.e., the relationship between smokers and nonsmokers is straightforward to determine with a GLM).

BEST PREFERRED RELATIVE TO RESIDUAL STANDARD BY DURATION FOR SELECTED PROFILE

In this example, we used male, nonsmoker, face amount band of $500,000–$600,000, male universal life (level net amount at risk), ULNG. We compare the ratio of best preferred to residual standard mortality by duration for selected issue ages (Figure 4).

Figure 4 Best Preferred Relative to Residual Standard for Selected Profile



The patterns can contain unexpected "jumps" for which there is no obvious explanation. As explained in previous examples, detecting such behavior inherent in the model requires significant analysis of model results.

Conclusions

We do not suggest that machine learning techniques have no place in experience studies or other applications in life insurance. We do want to emphasize that the characteristics of the model (including interpretability) are considerations that in some contexts are as important as predictiveness. There are serious consequences of not fully understanding the relationships inherent in your assumptions:

- Since virtually no data sets are homogeneous through all durations and ages in life insurance, you may end up with assumptions that are inappropriate for your new business and it will be difficult to evaluate this since relationships are not immediately obvious.
- It will be difficult to set charges such as cost of insurance without knowing all of the patterns inherent in the mortality assumption.
- Modifying the assumption in places where little credibility exists in the data will be difficult given that relationships are not easily identified.

With that said, further areas of research that could help limit these consequences include the following options:

- Exploring ways to detect unintuitive behavior (such as that illustrated in the examples) in GBM predictions
- Exploring ways to limit the GBM (or other machine learning methods) so that results are more likely to be intuitive (e.g., to guarantee that mortality increases with duration)
- Extracting value from the GBM in ways that can result in an improved GLM (e.g., finding more sophisticated features that can be used to improve the predictiveness of a GLM)

Kimberly Steiner, FSA, MAAA, is senior director at Willis Towers Watson. She can be reached at *kim.steiner@willistowerswatson.com*.

Boyang Meng, ASA, is consultant and senior actuarial analyst at Willis Towers Watson. He can be reached at *boyang.meng@willistowerswatson.com*.

Actuarial Fairness in the Era of Machine Learning

Marjorie A. Rosenberg

While the field of machine learning is not new, the level of interest in the tools by actuarial practitioners has been gaining great speed over the past several years. The Society of Actuaries (SOA) professional syllabus has been modified to introduce the concepts of machine learning in two new exams, Statistics for Risk Modeling and Predictive Analytics.

The purpose of this essay is to step back and ask ourselves the meaning of a fundamental tenet of actuarial practice, i.e., the notion of *actuarial fairness*. One can google the term and see thousands of links that all point to the concept of pricing risks related to the benefits. In fact, law has set precedence of establishing unfair or fair discrimination of premiums based on this concept.[1]

The Modelling, Analytics, and Insights from Data working party of the Institute and Faculty of Actuaries recently published a report highlighting areas of actuarial practice that could benefit from machine learning techniques.[2] These include topics ranging from product design and customer behavior to pricing, reserving, claims management, capital modeling, surplus distribution, and asset and liability management/hedging. Some of these functions are based within an organization, while others impact outside constituents, such as customers and regulators.

Those who advocate machine learning techniques focus on the *bias-variance trade-off*.[3] The idea is fundamentally based on the notion of selection of a model whose mean squared error (MSE) is lowest on an independent data set (called *test* or *validation* on the machine learning side and *out-of-sample* on the statistics side). We learn in our first statistics course that the MSE of an estimator is the sum of the variance plus the square of the bias.

The *bias* of an estimator is defined as the expected value (or large sample average) of an estimator minus *truth*. We show in our statistics class that if an estimator is unbiased, then the MSE equals the variance. Our focus in some statistical applications is a search among those estimators that are unbiased to find the one with the smallest variance. In the machine learning framework, we focus on minimizing the MSE without constraining the bias to be zero.

In his article, Breiman discussed two cultures, the data modeling culture (i.e., statisticians) and the algorithmic modeling culture (i.e., machine learning), to analyze data and make decisions.[4] The article was provocative in contrasting statistics and machine learning. Both cultures depend on functions with observed data inputs to produce some sort of prediction. Those advocating for machine learning emphasize the accuracy of the prediction of the outcome. Statisticians are also interested in prediction, but want also to be able to interpret and quantify the impact of an input variable on the outcome (called inference).

In the context of actuarial work in pricing, we define the notion of *actuarially fair* from the perspective of an unbiased estimator of the loss, called the equivalence principle. Here being actuarially fair is the actuary's way of defining premiums from the perspective of a customer. The premiums are calculated in a way that treats one person the same as others with the same risk profile, where the expected value of the premiums is equal to the sum of their expected losses and expenses.

---

1 Landes, Xavier. 2015. How Fair is Actuarial Fairness? *Journal of Business Ethics* 128, no. 3:519–533.

2 Panlilio, Alex, Ben Canagaretna, Steven Perkins, Valerie du Preez, and Zhixin Lim. 2018. *Practical Application of Machine Learning Within Actuarial Work*. Technical report. London: Institute and Faculty of Actuaries.

3 James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. New York: Springer.

4 Breiman, Leo, et al. 2001. Statistical Modeling: The Two Cultures (With Comments and a Rejoinder by the Author). *Statistical Science* 16, no. 3:199–231.

The result is a neutral way of explaining fairness to the public.

Along the same line of thinking, the notion of *risk adjustment* in health care examines the expected value of the claim. We adjust for the severity of an individual by examining its average amount of loss.

The practice of unisex pricing relies on actuarial fair principles.[5]

By removing the constraint of unbiasedness in a machine learning world, how then do we define and defend what is actuarially fair? From an outsider's perspective, if the bias of the premium calculation is negative, then the customer is paying a premium, on average, less than their true costs and expenses. The public could accept this situation, as it is favorable to them. Regulators could possibly be convinced if the MSE is smaller and insurer solvency is not at risk. But if the bias is positive and the customer is paying more than the expected value, then how is this communicated to both consumers and regulators? Then it can appear as if the insurer is charging higher premiums to benefit itself and unfairly penalize the consumer, as the premium charged is higher than its true value.

It seems that the machine learning approach of minimizing MSE without constraining the bias can be advantageous to the insurer to properly manage its total portfolio, so as to minimize the risk of expected outcomes overall. The historical definition of actuarial fairness changes with the use of machine learning tools, along with other actuarial processes, like risk adjustment, and the fundamental meaning of what the premium represents. The premium would no longer be actuarially fair as defined traditionally. The public and regulatory messaging would also need to be altered to reflect solvency and managing insurer risk in a new way.

With the increasing adoption of machine learning techniques by the insurance industry, actuaries need a broader perspective to examine the greater context of what is the definition of actuarial fairness and its impact on the law and ethics, in addition to the prediction accuracy of machine learning methods.

Marjorie A. Rosenberg, FSA, Ph.D., is the Assurant Health Professor of Actuarial Science and Michael E. Lehman Distinguished Chair for Inspired Learning in Business at the University of Wisconsin–Madison. Her research interests are in the application of statistical methods to health care cost and policy issues. She can be reached at *mrosenberg@bus.wisc.edu*.

[5] Schmeiser, Hato, Tina Störmer, and Jöel Wagner. 2016. Unisex Insurance Pricing: Consumers Perception and Market Implications. In *The Geneva Papers*, 102–138. New York: Springer.