Article from
CompAct
October 2019
Issue 63

Predictiveness vs. Interpretability

By 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 article, 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 article 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 based solely 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 more than 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

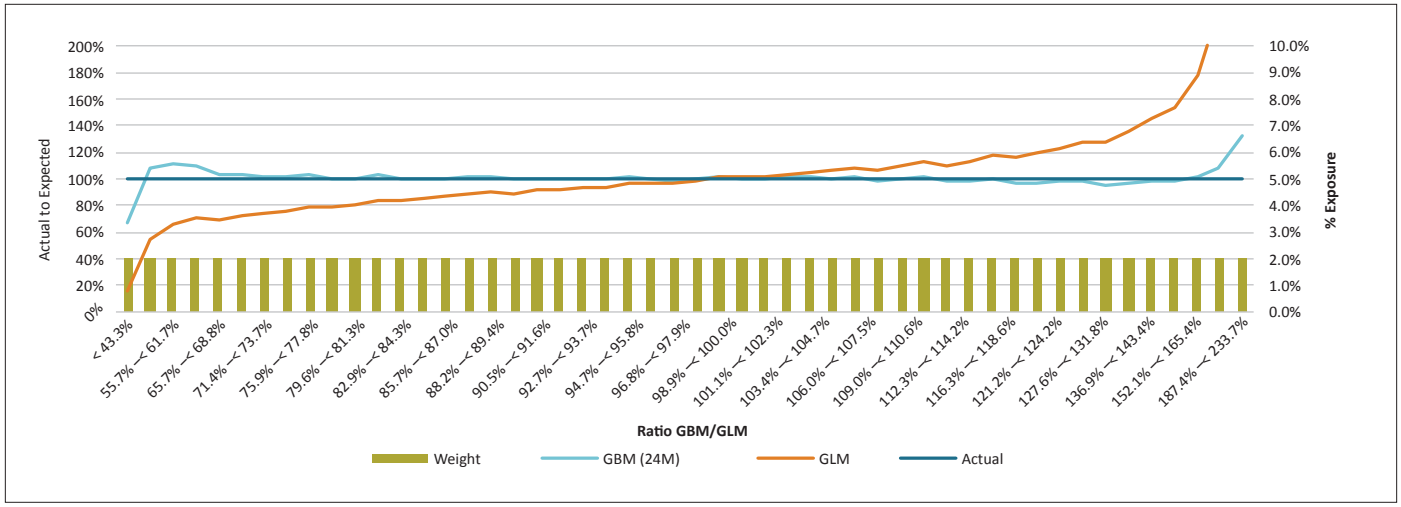
The response GBM was assumed to be distributed Poisson. Attained age, issue age and duration were included as continuous variables. Different groupings of categorical variables were experimented with. Hyperparameters were optimized using a grid search and cross-validation on a random split of the training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

Figure 1
Double Lift Rescaled



- 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.

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 qx by duration for selected issue ages (Figure 2).

We note that the qx 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

Figure 2
Qx for Selected Profile

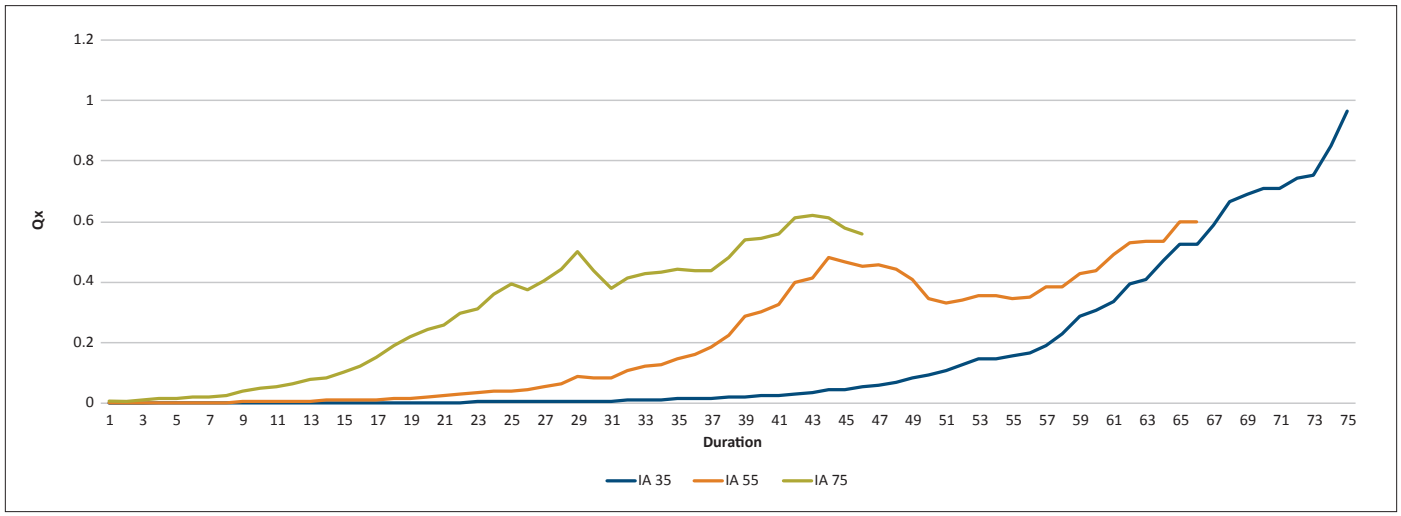
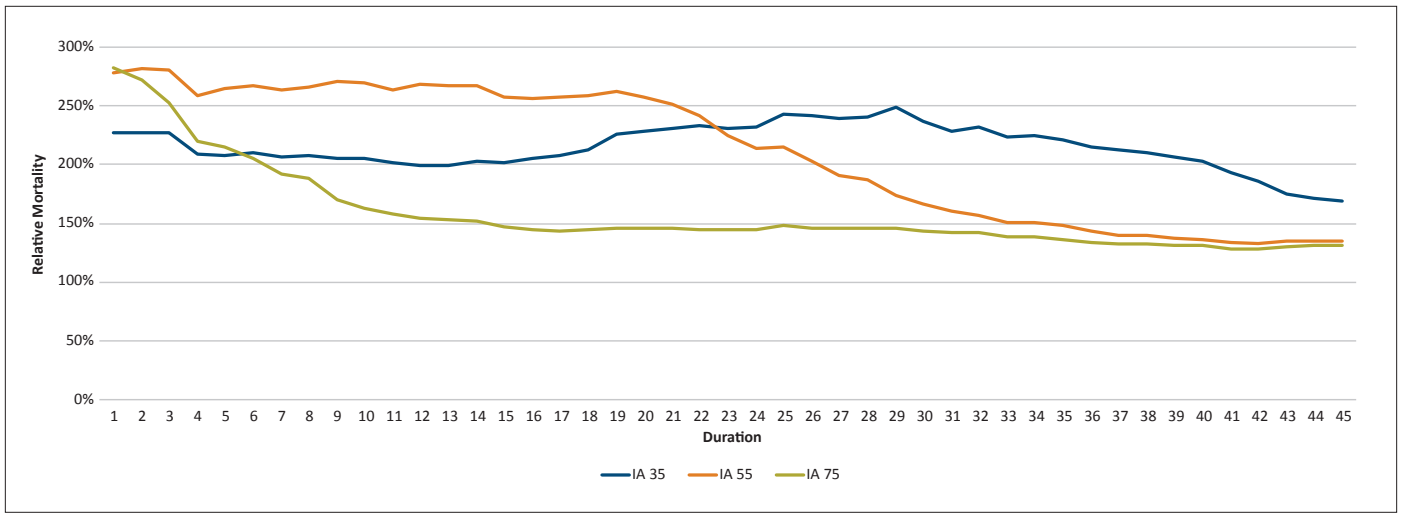


Figure 3
Smoker Relative to Nonsmoker for Selected Profile



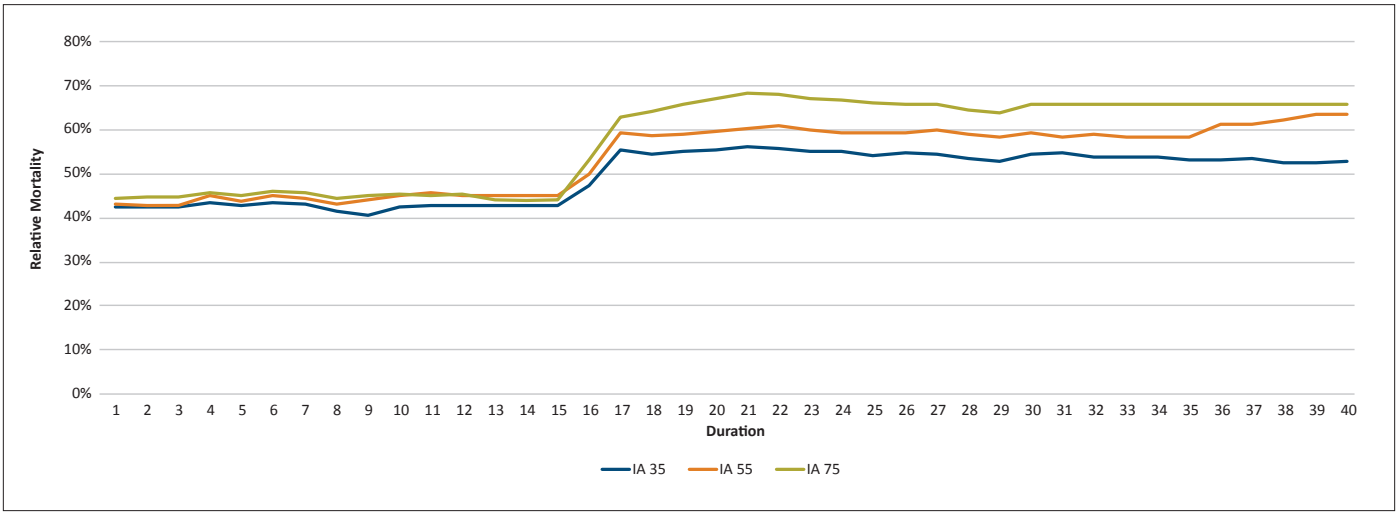
durations compared to the prior duration) at higher attained ages that lack credibility. While this is not surprising, the duration at which the pattern breaks down will vary by profile, and the only way to determine the point at which it breaks down is to evaluate the curve for all required profiles, 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).

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).

Figure 4
Best Preferred Relative to Residual Standard for Selected Profile



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).

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*.