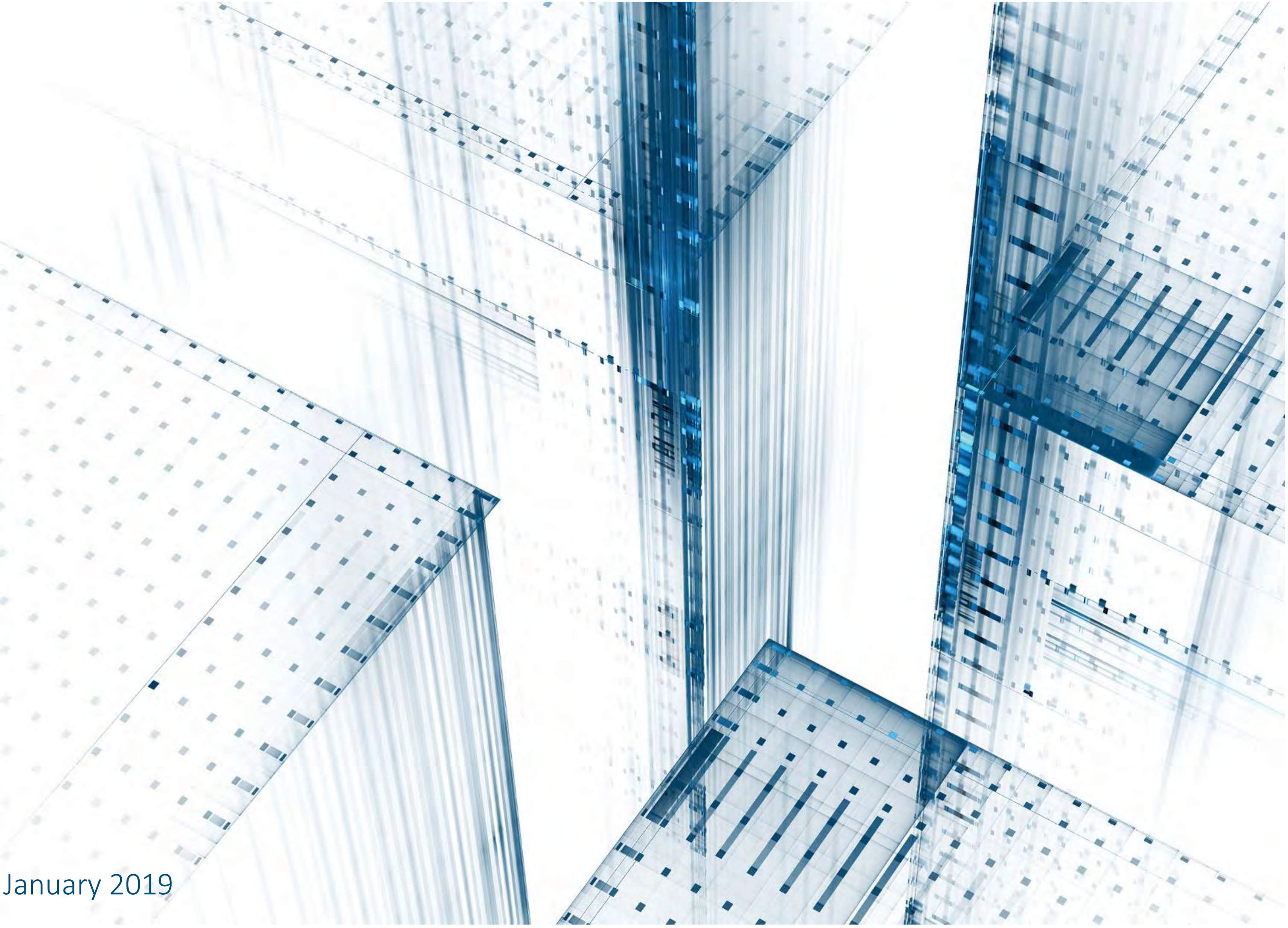
SOCIETY OF ACTUARIES®

Machine-Learning Methods for Insurance Applications

A Survey



January 2019

Machine-Learning Methods for Insurance Applications

A Survey

ALEX DIANA	PhD Candidate in Statistics University of Kent	SPONSOR	Society of Actuaries
JIM E. GRIFFIN	PhD, Professor of Statistics University of Kent	PROJECT OVERSIGHT GROUP	Syed Danish Ali Mary Pat Campbell Thomas Farmer Valerie Gingras Lie Jen Houng Mervyn Kopinsky Min Mercer Michael Niemerg Shisheng Qian Steven Siegel Ronora Stryker Jim Wright
JAIDEEP OBEROI	PhD, Lecturer in Finance University of Kent		
JI YAO	PhD, FIA, CERA Ernst & Young, China, and Shanghai Jiao Tong University		

Caveat and Disclaimer

The opinions expressed and conclusions reached by the authors are their own and do not represent any official position or opinion of the Society of Actuaries or its members. The Society of Actuaries makes no representation or warranty to the accuracy of the information

Copyright © 2019 by the Society of Actuaries. All rights reserved.

CONTENTS

Executive Summary **4**

Introduction **5**

Machine Learning 5

Data 6

Generalized Linear Model **7**

Issues With GLM in Insurance Modeling 7

Regularization 7

LASSO, Elastic Net and Ridge Regression 7

Feature Selection: Subset Selection, Stepwise Regression and Bayesian Variable Selection 8

Nonlinear Classification and Regression Models **9**

Classification and Regression Trees (CART) 9

Ensemble Methods **11**

Bagging 11

Random Forests 12

Boosting and Gradient Boosting 14

Bayesian Additive Regression Trees (BART) 17

Other Machine Learning Methods and Topics **18**

Multivariate Adaptive Regression Splines (MARS) 18

Neural Nets and Deep Learning 18

Unsupervised Learning 19

Cluster Analysis 19

Additional Methods and Combinations 19

Existing Comparative Studies **20**

Applications to Wider Problems for Actuaries **20**

Estimation Packages **20**

Challenges **21**

Interpretation 21

Discovering Nonlinear Relationships, Dealing With Large Numbers of Variables and Sparse Data 21

Conclusion **21**

References **23**

About The Society of Actuaries **27**

Copyright © 2019 Society of Actuaries

Machine-Learning Methods for Insurance Applications

Executive Summary

This project has been awarded funding by the Society of Actuaries from the Research Expanding Boundaries (REX) pool to contribute to actuarial practice expansion. The objective of the current report is to provide a literature survey of methodological improvements from the field of machine learning to insurance claim modeling.

In this report, we describe and illustrate a range of machine-learning approaches that have been used in the insurance literature or have the potential to be used. We emphasize variable selection methods, including those applicable to the generalized linear model (GLM), the current workhorse for the industry. The methods covered include LASSO, elastic net, ridge regression and Bayesian variable selection. The former three methods involve penalizing the objective function as a means of shrinking certain parameter estimates toward zero. The latter approach assigns weights to alternative models in order to combine them to select features that have predictive value.

We then discuss classification and regression trees (CART), a method that helps capture the nonlinearities that are challenging for the linear approaches. However, it also suffers from high variance and lack of smoothness. CART can in many cases be augmented by a range of ensemble methods, that combine more trees (estimated in parallel or sequentially) to improve the trade-off between bias and variance. The methods we cover are bagging, random forests, boosting and Bayesian additive regression trees (BART). In bagging and random forests, the idea is to average across a number of trees fitted to subsets of the data, whereas boosting involves sequential estimation using the residuals from the previous model fit. These methods are suited to trees but are not restricted to them. BART involves combining simple trees, which uses a prior to discourage complicated tree structures.

We also include multivariate adaptive regression splines (MARS), a more traditional nonparametric regression method that generates a nonlinear regression function by combining shorter line segments. In our illustrations, this method is outperformed by the other methods considered.

This report is accompanied by two working files, prepared in the Jupyter software, that illustrate the implementation of the models covered in the report using R. Each file uses a different data set. The first is the Group Long Term Disability data set (see Kopinsky 2017), and the second is the Long Term Care Incidence data set (see Bodnar et al. 2015).

In addition to the regularization methods listed here, an important consideration for evaluating model predictions is the loss function itself. We illustrate two alternatives, the mean-squared error (MSE, applied to the disability data) and the receiver operating curve (ROC, applied to the long-term care data).

The evidence in the public domain about the performance of the various machine-learning methods in claims prediction is still limited. Overall, the literature has found that different methods offer improvements for specific applications, suggesting the need for more robust examinations of model performance. Although we compare the models using simple out-of-sample predictive ability, we also find that there is potential for improving predictions relative to linear methods. We recommend further detailed comparisons, which could be made possible by the increased availability of standard public data sets for researchers to use.

Introduction

The actuarial profession strives to develop or adopt the best approaches to understanding claims data. Recent trends in data collection and data science call for the attention of actuaries to examine whether existing approaches to modeling can be improved and lead to better decisions. Data sets are getting larger as researchers collect and combine information from a broad range of sources. This has necessitated the development of methods to select the most relevant variables in predictive models and to expunge spurious relationships. Claims data tend to be sparse, volatile, skewed and time variant. This makes them excellent candidates for analysis using machine-learning approaches, particularly variable selection methods.

The industry has widely adopted generalized linear models (GLMs) as a standard approach to modeling claims. Typically, claims are modeled either as a combination of two components (frequency and severity) or directly using a distribution such as the Tweedie distribution. While this approach is more accurate than a multiway table approach, its popularity is also based on the speed with which it can be implemented and its easy interpretability. However, the standard models do have certain limitations, and there is a potential that some of these limitations may be addressed by the methods in this survey.

This survey attempts to link the challenges of using GLMs with the motivation for using the methods covered here. One such challenge is the time taken to build models when incorporating new information. This becomes more important as data availability from a wide range of sources makes it possible to consider more complex relationships. As a further challenge, the discovery of interaction effects is difficult, often forcing model builders to rely on summary indicators (e.g., a credit score rather than the underlying census and credit information used to construct it). In addition, missing data from external sources poses difficulties to the modeler, as does the issue of uncertainty in sparse segments when estimating GLMs.

The machine-learning methods applied to insurance data covered here include tree-based methods and regularization methods, such as the LASSO and Bayesian variable selection methods. While reviewing methods, we also note evidence where available on other criteria, such as interpretability, resource requirements (difficulty, speed, scale), stability and prediction performance.

Before proceeding to a review of the methods, we briefly introduce machine learning.

Machine Learning

Machine learning is a field that absorbs techniques from a wide range of disciplines with the objective of prediction based on data. There are two broad categories of machine learning: supervised and unsupervised. In this review, we report comparisons of methods classified as supervised learning, though we also document unsupervised methods that have been applied to insurance problems.

There are now many textbooks that describe the range of machine-learning methods. Two references that are widely used are Hastie et al. (2009) and Murphy (2012). The aim of machine-learning methods is to be able to produce generalizable rules for prediction based on patterns identified in data. The methods offer natural tools for prediction under a range of tasks (e.g., classification, clustering, regression). The discovered patterns can often be very complex. Due to their complexity, it is also usually necessary to employ regularization to avoid overfitting that would lead to poor prediction performance in unseen data. As a result, machine-learning methods tend to incorporate both estimation and model selection within the same procedure. As the tasks and loss functions vary by context, the development of machine-learning methods has been relatively more problem specific. Caruana and Niculescu-Mizil (2006) provide comparisons of a range of methods based on multiple criteria, emphasizing the need to match the method to the loss function.

Copyright © 2019 Society of Actuaries

Two articles that discuss the potential for machine-learning methods in econometrics are Varian (2014) and Mullainathan and Spiess (2017). They provide a useful introduction to machine-learning methods applied to common problems such as the estimation of treatment effects and the prediction of outcomes based on individual characteristics.

Actuaries have recognized the potential of machine-learning methods as part of their focus on predictive analytics, suggesting several applications of the methods. An early article summarizing a selection of machine-learning methods with actuarial science applications was Shapiro (2000), in which he explained some optimization methods as well as the use of neural networks and classification algorithms.

These resources provide both an overview of machine-learning methods and detailed reference on implementation. In the sections that follow, we provide brief summaries of the most popular approaches related to GLMs and machine learning, listing some of the existing findings in the literature.

We also present comparative estimates on two data sets that we will describe in the next subsection. Accompanying this document are additional files produced in the Jupyter Notebook format that include the R code and results from the prediction comparisons we have carried out. These files can be read in a browser, and the R code copied to replicate the results. Jupyter Notebook is open-source software that integrates several languages. For those interested in working within the Jupyter format to modify the R code directly and produce new results, we have also made available the Jupyter files with a brief description and a link to the resource website for the software. The data used for the two illustrations is provided in accompanying CSV files.

Data

The data sets used for the empirical illustrations are publicly available from the Society of Actuaries. The first is the group long-term disability (GLTD) data kindly provided by Mervyn Kopinsky based on his report on using tree models to predict recovery and mortality rates (Kopinsky 2017). The second is the Society of Actuaries Long-term Care Intercompany Experience Study (Bodnar et al. 2015).

The GLTD recovery rate data consists of more than 500,000 observations of seven predictors that summarize up to 46 million records. Details of the data processing can be found in Kopinsky (2017). Summaries of the data are provided in the accompanying Jupyter file.

The long-term care claim incidence (LTCI) data is composed of more than 10 predictors and more than 1 million observations. We expand the pivot table provided and then collapse some of the categories with no actual claims. Once again, details of the data and empirical work are provided in the accompanying Jupyter file.

An essential feature of building good data-driven models is the data itself. The data we use in this report is already cleaned and prepared, requiring relatively minor adjustments. While we do not focus in this survey on the preprocessing of data, it should be noted that each paper we report here spends a considerable amount of attention describing the data and discussing its quality and relevance.

When comparing the models, we consider the mean-squared error (MSE) when estimating the GLTD data and the area under the curve (AUC) when estimating the LTCI data. The AUC refers to the area under the receiver operating characteristic curve, which is produced by plotting the true positive rate against the false positive rate at various points (quantiles). When classification is perfect, the AUC equals 1.0 (the area of the square plot), whereas when it is random, the AUC moves toward 0.5.

Performance criteria play an important role in machine-learning methods because they are directly aligned with the objective function. As a result, model selection and estimation apply the same objective function.

Generalized Linear Model

GLMs are widely used for pricing in the insurance industry. The classic textbook on GLMs is McCullagh and Nelder (1989). Applications of GLMs to insurance problems can be found in Brockman and Wright (1992), Anderson et al. (2007), De Jong and Heller (2008), Ohlsson and Johansson (2010) and Dean (2014).

GLMs use a linear regression via a link function to predict variables that have potentially non-normal distributions, as is often the case in actuarial science. In other words, a GLM is specified as:

$$g(E[y]) = X\beta,$$

where $g(\cdot)$ is a differentiable and strictly monotonic function, y is the dependent variable, X is a matrix of predictors and β is the parameter vector. The data is assumed to be from the exponential family of distributions (including common distributions such as the binomial, Poisson and Gaussian), and the model can be estimated using maximum likelihood or Bayesian methods.

Issues With GLMs in Insurance Modeling

Some of the criticisms of GLMs include the following:

- Either zero or full credibility is given to the data, and there is no way to do blending.
- Prediction of a risk depends on data in other, potentially different segments.
- Model predictions depend on the mixture of rating factors in the data.
- Maximum-likelihood estimate of prediction is lower than the mean of the prediction distribution.
- Link function could bias the model prediction and significantly change the lower and upper bound of prediction.
- Model diagnostics are relevant only in the segments where the model is used.

In this report, we will consider two main broad approaches to this problem. The first approach assumes that the structure of GLMs is correct but that there is a list of potential features. It is often ineffective to include all these variables in the model, since parameter estimates can be poor with large standard errors leading to inaccurate predictions. An effective method for choosing features that are useful for prediction can address these problems and allow consideration of a large number of potential features. This can allow more accurate predictions than careful preselection of features. This is known as ***feature selection***. The second approach assumes that the relationship between the features and the mean of the target is not appropriately modeled by a GLM. Specifically, these methods allow for nonlinear relationships and feature selection rather than the linear relationships (after a transformation through the link function) and feature selection in the first approach. The second approach is more flexible but can more easily overfit the data, leading to poor predictions. The first approach offers models that are often easier to interpret but can predict poorly if the linearity assumption is not appropriate for the data.

For the first approach, a number of methods are available under the category of regularization methods. Next we discuss general regularization methods that are widely applied to most estimation methods. Other methods will be introduced in later sections.

Regularization

The essential objective with regularization is to fit regression models with large numbers of variables while avoiding overfitting within the training data. Two broad groups of approaches are described next.

LASSO, Elastic Net and Ridge Regression

Regularization methods have become increasingly popular following the seminal work of Tibshirani (1996), where the LASSO estimator for linear regression models was developed. The idea is to replace the maximum-likelihood estimator of the regression coefficients (and any other model parameters) with a penalized maximum-likelihood estimator. Suppose we have regression coefficients β (and potentially other model parameters θ); then the penalized maximum likelihood estimator of β and θ is

$$\text{argmax}_{\beta,\theta}\,(L(\beta,\theta) - p(\beta)),$$

where $L(\beta,\ \theta)$ is the log-likelihood function for a GLM and $p(\beta)$ is a pre-specified penalty function. The penalty function penalizes β_i, whose absolute value is unrealistically large and leads to estimates for which the overall magnitude of β is smaller than the maximum-likelihood estimate. The following choices for $p(\beta)$ are popular:

- LASSO penalty function: $p(\beta) = \lambda \sum_{i=1}^{p} |\beta_i|$ where we have p features
- Elastic net penalty function: $p(\beta) = \lambda_1 \sum_{i=1}^{p}|\beta_i| + \lambda_2 \sum_{i=1}^{p} \beta_i^2$
- Ridge penalty function: $p(\beta) = \lambda \sum_{i=1}^{p} \beta_i^2$

An important property for feature selection is that the penalized maximum-likelihood estimates for the effect of some features can be exactly zero. This allows some variables to be removed from the model. The ability to generate zeros is controlled by the choice of $p(\beta)$. The LASSO and elastic net penalty functions are able to generate zeros, whereas the ridge penalty cannot. The elastic net penalty reduces to the LASSO penalty if $\lambda_2 = 0$ and reduces to the ridge penalty if $\lambda_1 = 0$. Therefore, the elastic net is seen as a compromise between the LASSO penalty and the ridge penalty. There has been a substantial amount of work on the properties of these procedures, which is reviewed in Hastie et al. (2015). A succinct review for actuaries is provided in Niemerg (2016). One important practical consideration is the way that the feature selection method behaves when some features are highly correlated. If a group of highly correlated features contains important features, the LASSO penalty will tend to set the coefficient for only one feature to be nonzero, whereas the elastic net will tend to set more of these coefficients to be nonzero.

A practical problem in the use of regularization methods is the choice of the penalty parameters: λ for the LASSO and ridge penalty functions, and λ_1 and λ_2 for the elastic net penalty. These are usually chosen using the idea of separating the data into different sections for validation purposes. The basic approach is to have in-sample data or a training sample to estimate the model and out-of-sample data or a test sample to evaluate performance. The estimation step itself requires regularization though methods such as ***cross-validation***. In cross-validation, the data is separated into a number of ***folds***, and then estimation takes place by sequentially leaving out one or more folds for evaluation of the predictive fit of the model. The process allows one to tune parameters that may, for instance, determine the complexity of the model. At the least, to avoid overfitting, we need to separate the data into a training section and a test section. Recent reports on predictive modeling from the Society of Actuaries (see, e.g., Xu et al. 2015; Ewald and Wang 2015) highlight the use of separate training and test data sets as an essential element of building a predictive model.

For both data sets that we have analyzed in this report, we set aside 30% of the sample as a test sample and use 70% for training. Our initial checks using regression and a GLM with the logit link highlight the degree of nonlinearity, something that linear models would struggle with. We are using a data set with relatively fewer variables than one might have access to when conducting a pricing exercise. As a result, there is little room for methods such as the LASSO to improve the out-of-sample performance of the model by removing overfitting or shrinking the parameters of redundant predictors.

Feature Selection: Subset Selection, Stepwise Regression and Bayesian Variable Selection

Copyright © 2019 Society of Actuaries

In regularization methods, we concentrate on estimating the coefficients of the features, the βs, and feature selection arises as a by-product of the regularization. Other feature selection methods concentrate on calculating a goodness-of-fit measure for different possible subsets of the potential features. For example, subset selection methods calculate a goodness-of-fit measure, such as AIC or BIC, for all possible subsets. The final model is chosen to be the model with the best value for the goodness-of-fit measure, such as the smallest value of the AIC or the BIC. There are 2^p possible models if there are p possible features, so this method is feasible only if p is relatively small (for example, p less than 30). In large-scale problems, iterative algorithms can be used to find an optimal model. For example, in stepwise methods, suppose we have a chosen subset of features at one iteration; the following iteration considers the effect on the goodness-of-fit measure of including any one of the features that are currently not used or removing one feature that is currently used. The algorithm would then choose the change that leads to the largest improvement in the measure of goodness of fit. Randazzo and Kinney (2015) use a logistic regression to identify likely frequent visitors to the emergency room. While they do not specify the variable selection approach, they do highlight the top 20 predictors in order of their contribution to R^2.

Many approaches to feature selection concentrate on finding a single model. Bayesian approaches offer the opportunity to combine all possible models in the analysis. Predictions can be made by combining the predictions for each model with weights determined by the fit of each model to the data. In the Bayesian approach, a prior distribution is defined for all parameters, and we are able to use the decision to include or exclude a feature as a parameter. We define γ_i to be 1 if the *i*th feature is included and 0 if it is not included, and $\gamma = (\gamma_1, \dots, \gamma_p)$ is a vector for all possible features. Then, conditional on the included feature, a prior distribution can be specified for the regression coefficients for the included features, β_γ, and other parameters, θ. Although the prior distributions can be chosen using expert information, they are usually chosen to lead to a suitably simple model. A typical setup is $\gamma_1, \dots \gamma_p$ are independent and $\pi(\gamma_i = 1) = h$. This implies that the prior expected number of included features is ph, and this can be used to choose a value of h that leads to a suitable prior mean for the data.

Once a prior distribution has been chosen, Markov chain Monte Carlo (MCMC) algorithms or variational Bayes algorithms (Carbonetto et al. 2017) can be used to calculate the posterior distribution of sets of included features. This allows a prediction to be calculated by averaging predictions from the different sets of included features weighted by their respective posterior probabilities.

Nonlinear Classification and Regression Models

Generalized linear regression models assume that the relationship between features and the response are linear (usually, after some transformation of the effect of the features through a link function). In particular, this restricts the expected response to be either increasing or decreasing as a function of each feature. Generalized linear regression models can be made more flexible by including powers of features or interactions between features. However, if the number of features is large, this approach can lead to models with huge numbers of parameters, which makes estimation challenging. These difficulties with generalized linear models have led to interest in nonlinear classification and regression models, which allow more general relationships between the features and the expected response, including interactions. In this section, we will first review the popular classification and regression tree (CART) approach before considering development of this approach to so-called ensemble methods that include random forests and a Bayesian alternative, Bayesian additive regression trees (BART).

Classification and Regression Trees (CART)

In GLMs, a single model is used for all values of the feature. In contrast, a popular approach to nonlinear models assumes that different models are used for different combinations of the features. For example, if we have two features, age and sex, we might be able to build separate models for men over 60, men under 60, women over 50 and women under 50. All observations will fall within one of these categories, and we say we have partitioned the feature space into the different categories. Note that the male and female populations have been split at different ages. Figure 1 illustrates the idea.

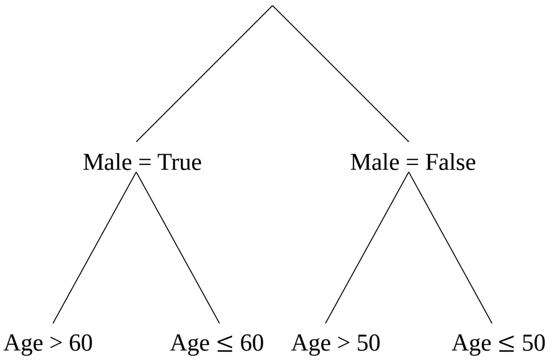


Figure 1: A Simple Tree Structure

To use the approach, we need a way to describe the partition of feature spaces (i.e., the different categories), the model for the categories, and a way to infer the partitions and models from data. In CART, the partitions are described by a tree. This structure has benefits for inferring the partitions which can be learned sequentially. In each category, we usually assume a constant level for the response for continuous data and assume a constant probability of positive response in a classification problem. If we use $R_1, \dots, R_m$ to represent the partitions, then, in the continuous case, the regression function is

$$f(X) = \sum_{k=1}^{m} m_k \, I(X \in R_k),$$

where m_k is the level for the *k*th partition.

The advantage of using a tree structure is that the structure of the tree can be learned sequentially. We start by considering the partition $R_1 = \{X|X_j \leq s\}$ and $R_2 = \{X|X_j > s\}$, where the variable to be split *j* and the split point *s* need to be chosen. These parameters can be chosen by finding the values that minimize

$$\sum_{x_i \in R_1} (y_i - \widehat{m}_1)^2 + \sum_{x_i \in R_2} (y_i - \widehat{m}_2)^2,$$

where $\widehat{m}_1$is the mean of the y_is in $R_1$and $\widehat{m}_2$is the mean of the y_is in R_2**.** The minimization over the split points can be simplified by noticing that we only need to consider the splits defined by each observed value of the *j*th variable once the first split variable and split point have been found. The algorithm then subdivides $R_1$and R_2 by finding split variables and split points in each region. This process can be continued by subdividing all current regions at further steps of the algorithm. Typically, the tree is grown until the minimum number of observations in a region (such as 5) is reached. This usually leads to a large tree that will tend to overfit the data and lead to poor out-of-sample predictive performance. This is addressed by pruning the tree by removing some splits.

The method has been extensively applied in the machine-learning literature, but some limitations have been identified:

1. *Instability of the tree.* The estimated tree structure can be sensitive to sample of data. For example, dividing the data into two halves and estimating trees can lead to very different inferred trees. Therefore, CART is often considered a method with a high variance. This is a particularly important consideration for actuarial

applications, where the dependent variable (e.g., claim incidence) can be very unevenly distributed in the data. Cross-validation and ensembling help to address this problem. Ewald and Wang (2015) also suggest the use of stratified sampling.

2. *Lack of smoothness.* By its very nature, the tree allows two adjacent groups to have significantly different models. While this may be appropriate in some circumstances, it may also lead to difficulties. For instance, actuaries often interpolate or extrapolate predictions to categories where the data are insufficient (e.g., a class where there is no experience of claims in the data).

Parkes (2015) uses ensemble decision trees to incorporate numerical measurements from electronic medical records to improve prediction of care costs. He highlights that one of the advantages of using the method is its flexibility in the presence of high levels of dependence among the features and the overall nonlinearity of the relationships modeled.

Tree-related models can be studied using the rpart library in R. A careful evaluation of the use of CART is provided in Kopinsky (2017), from whom we have received the GLTD data. We confirm his finding that CART improves on GLMs for the GLTD data. However, when applied to the LTCI data, we do not find that CART outperforms GLMs.

Ensemble Methods

The previous methods have only considered the use of one model for prediction. In contrast, ensemble methods use a number of different models to make predictions. Intuitively, this is a case of "hedging our bets." If the different models provide different predictions, then sensibly combining the predictions from each model can lead to much better overall predictive performance. There are many ensemble methods, and we will consider some of the most popular ones. Before introducing the idea of a random forest, it is useful to describe the idea of bagging. We will then discuss boosting and BART.

Bagging

Bagging of trees was introduced to address the problem of the high variance of trees. In this method, *B* new data sets are created from the original data set. Each new data set has *n* observations, which are sampled with replacement from the original data set. A regression tree is estimated from each new data set, and the predictions for each tree are averaged. Therefore, the ensemble estimate of the regression function is

$$f_{\text{ave}}(X) = \frac{1}{B}\sum_{b=1}^{B} f_b(X),$$

where f_b is the regression function from fitting a regression tree to the *b*th new data set. The intuition behind the idea is that the variance of an average will be lower than the variance for each element in the average.

To implement bagging with CART, one could use the Rborist package in R. When implementing bagging, we need to choose the number of trees, B. In practice, the prediction error on test data, which is often called the "bagging error", tends to stabilize as the number of trees increases, so we only need to find the point at which stabilization begins. Figure 2 shows the results for different numbers of trees with the LTCI data. We find that the error rate stabilized beyond approximately 250 trees, with values in the range 0.0011615 to 0.0011625.

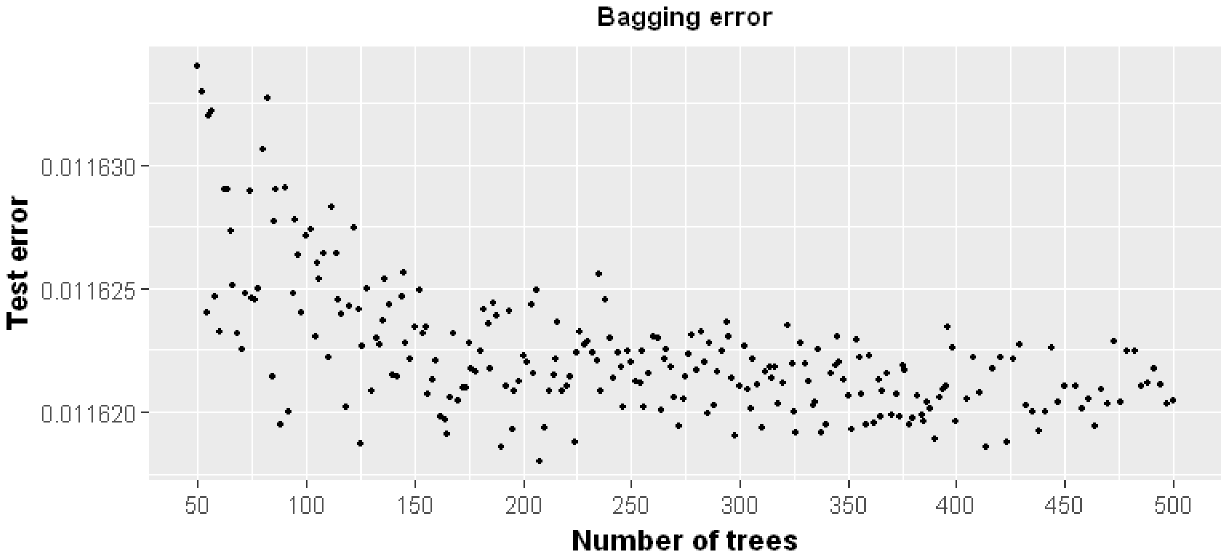


Figure 2: Mean-Squared Error for Different Numbers of Trees: Held-Out Testing Sample With GLTD Data Set

When we applied bagging in the context of CART, we found an improvement in the MSE for the GLTD data set but no improvement for the LTCI data set. This may be, in part, due to the differences in the incidence rates in the two data sets, as the LTCI data set has very low incidence rates, leading to significant variation across small subsamples.

Random Forests

Breiman (2001) introduced the random forest as a variation on the idea of bagging trees. Again, B new data sets are created by sampling with replacement, but the way the regression tree is constructed is restricted. In regression trees, we are allowed to use any variable to make the split at any point in the tree. In a random forest, at any point in the tree, a candidate set of c features are selected at random (without replacement) from the list of features, and only these features can be used to define a split. The value of c is much smaller than the number of predictors p and is typically chosen to be $c = \sqrt{p}$. To understand why this method might work, consider the argument for bagging trees, i.e., that the variance of an average is much smaller than the variance of elements of the average. This is clearly true if the elements are independent (which is essentially the central limit theorem), but the reduction will be small if the elements are highly correlated. Therefore, reduction in the variance from bagging depends on the predictions from the trees estimated from each new data set not being highly correlated. The extra step in random forests encourages greater diversity in trees than bagging, since some variables are excluded as split variables at each step of the tree-fitting algorithm.

One of the early applications of tree ensembles that supports their potential for insurance applications was Derrig and Francis (2008). They showed improvements by random forest and TreeNet (an alternative algorithm) over logistic regression in predicting fraud in vehicle claims.

Shehadeh et al. (2016) applied random forests to data on 130,000 applications for life insurance. They found that stratified sampling was essential to exploit the performance of the random-forest algorithm, as the data was heavily unbalanced and claims represented less than 5% of the data set.

The Rborist package also works for random forests. The user must choose hyperparameters, such as the number of trees, the number of predictors chosen at each split, and the minimum size of a node (pruning choice). We used 250 trees following the results using bagging (although this could again be chosen using the MSE on a test set) and compared different minimum sizes of nodes and the number of predictors chosen at each split.

Copyright © 2019 Society of Actuaries

Figure 3 shows the MSE for the test data after varying these two hyperparameters for the GLTD data. The test errors are reduced by allowing three variables to be tried at each split (gold line), rather than two variables (red line), but increases as we use four, five, six or seven variables. This is typical of random forests with an optimal number of variables to be allowed at each split, providing the lowest MSE. In contrast, bagging allows all variables to be considered at each split, so random forests usually provide a lower MSE. The plot also show that imposing a minimum size of nodes leads to improved MSE on the data test. If we restrict to three possible variables at each split, the MSE decreases as the minimum size of the node increases to 150 but then increases for larger minimum size. In our applications, we find that random forests perform best of all models for the GLTD data but do not offer much improvement for the LTCI data over the CART model.

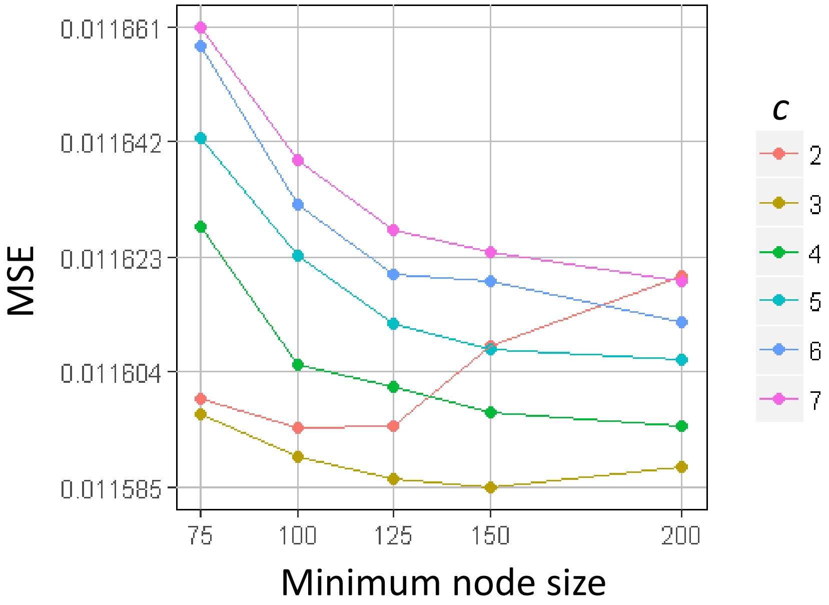


Figure 3: Mean-Squared Error (MSE) for Different Parameters of the Random Forest Algorithm: Held-Out Testing Sample With GLTD Data Set

Note: The MSE is reported for minimum node sizes ranging from 75 to 200, with *c* representing the number of predictors chosen at each split.

A challenge for random forests is their complexity, as the interpretation becomes much more complex than for the underlying trees. One way of addressing this is to provide measures of relative variable importance. In fitting a CART model, the split variable is chosen by looking at the improvement of the fit to the data. A natural way to measure the relative importance of a particular variable is to sum the improvements due to that variable. In a random forest, we further sum across all these measures across all trees in the ensemble and weight according to the contribution of each tree to the overall fit to the data. Figure 4 shows the relative importance for the variables in the LTCI data set. This identifies a few variables as very important to the fit. Unfortunately, in contrast to the fitted random forest, this method cannot describe the interactions between variables.

Copyright © 2019 Society of Actuaries

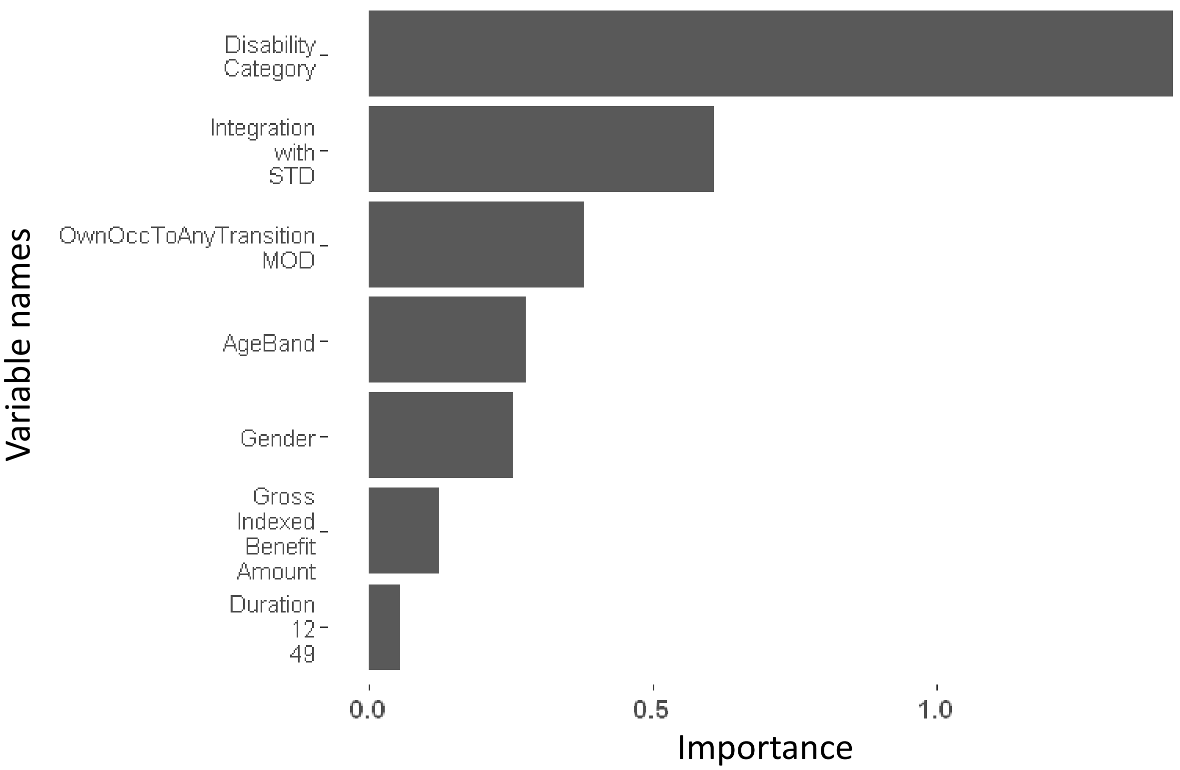


Figure 4: Relative Importance of Variables: Random Forest Algorithm With GLTD Data Set

Boosting and Gradient Boosting

Boosting is an alternative method for building an ensemble of models. Unlike bagging, which fits models to new data sets created by resampling the original data set, boosting works sequentially and fits a model based on the output from the previous models. The approach is most easily understood as an algorithm. We initialize the regression function $f_0(x) = 0$ and residuals $r_i = y_i$ for all observations. The algorithm then repeats the following steps for $b = 1, \ldots, B$:

1. Fit a tree model to the features *X* and responses *r*. The regression function for this tree is g_b.
2. Set $f_b(x) = f_{b-1}(x) + \lambda g_b(x)$.
3. Calculate new residuals $r_i = r_i - \lambda g_b(x_i)$.

The final fitted model is $f_B(x) = \lambda \sum_{b=1}^{B} g_b(x)$. Intuitively, this approach allows the model to adjust the regression function in several iterations rather than in one simple fit, as in the basic CART model. By fitting the residuals from previous fits, the model is able to concentrate on fitting the data in areas where previous models have performed more poorly. Many applications of these methods have shown that boosted methods outperform standard approaches.

Gradient boosting is a generalization of boosting to allowing the estimation problem to be defined by a loss function. For example, if we have a continuous response variable, then we would want to learn the regression function $f(X)$, which is the mean of the response variable for a particular value of the features X. In this case, a suitable loss function is $L(y,\ f(X)) = (y - f(X))^2$, since it is minimized when $f(X) = y$. However, we could also consider another loss

Copyright © 2019 Society of Actuaries

function such as $L\big(y,\ f(X)\big) = |y - f(X)|$, which might be more robust to a heavy-tailed distribution of regression errors. The gradient-boosting algorithm starts with an initial regression function $f_0(x) = 0$ and then involves the following steps for $b = 1, \dots, B$:

1. Calculate pseudo-residuals $r_i^{(b)} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{b-1}}$.

2. Fit a model to the features *X* and responses *r*. The fitted regression function is g_b.

3. Find λ_b that minimizes $\sum_{i=1}^{n} L(y_i, f_{m-1}(x_i) + \lambda_b g_b(x_i)$.

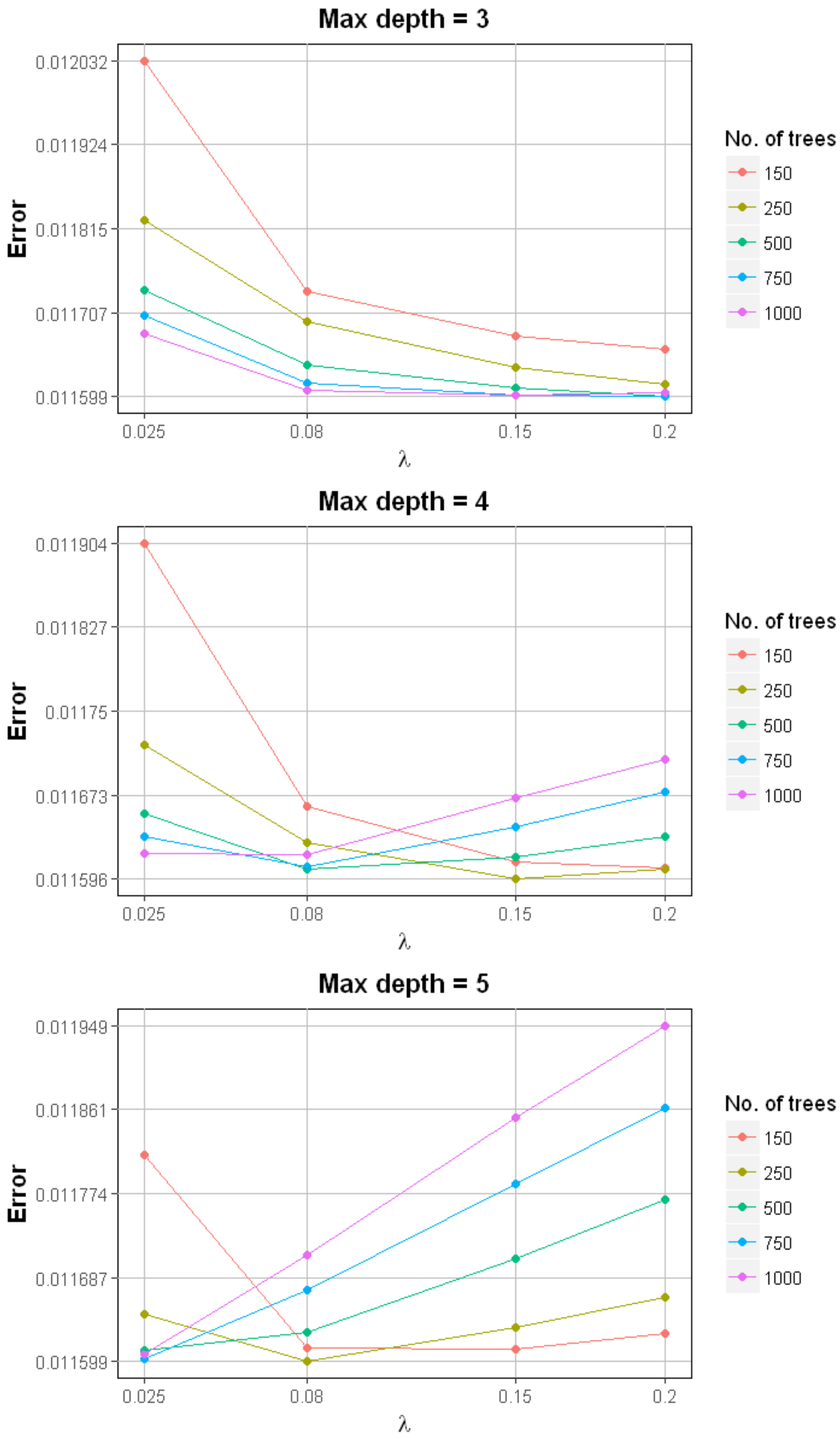
4. Set $f_b(x) = f_{b-1}(x) + \lambda_b g_b(x)$.

The final fitted model is $f_B(x)$.

Using the gradient-boosting example, we highlight the role of another hyperparameter important for the pruning of trees: the maximum tree depth. In addition, we try different values of the shrinkage parameter around the default. To implement boosting in R, we use the xgboost package. Several hyperparameters need to be chosen, and we initially concentrate on three: the number of trees, the maximum depth of each tree and the λ_b parameter in the learning algorithm. The MSEs are shown in Figure 5. Varying λ_b for a fixed number of trees and maximum depth leads to a similar shape, where the MSE is highest for small and large values with an optimum in between. The optimum value of λ_b tends to get smaller as the maximum depth and the number of trees increase. We find that the optimum MSE over these parameters is given by the choices of max depth equals 4, number of trees equals 250, and $\lambda_b = 0.15$.

Many other parameters can be adjusted in the gradient-boosting algorithm, and results of varying these hyperparameters are given in the accompanying HTML document. We find that these parameters have a much smaller effect on MSE for the problems that we consider.

Copyright © 2019 Society of Actuaries



Max depth = 3

Max depth = 4

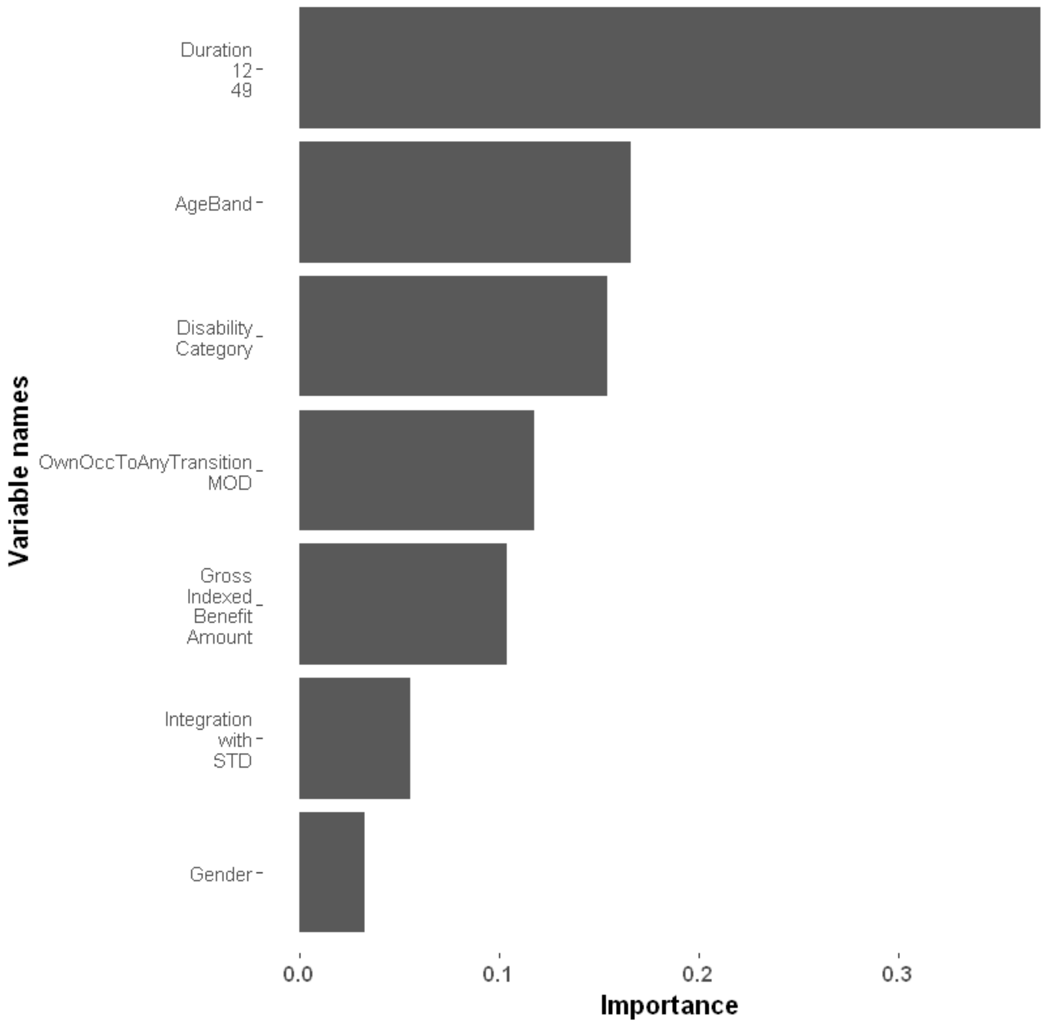
Max depth = 5

Figure 5: MSEs for Different Choices of Hyperparameters in Gradient-Boosting Algorithm With GLTD Data

Note: We varied λ and the number of trees, with cases of maximum tree depth of 3, 4 and 5.

Copyright © 2019 Society of Actuaries

The relative importance of variables in gradient boosting can be measured in a similar way to the random forest. The importance of a variable within each tree can be measured by summing the improvements in overall fit due to split in that variable. In gradient boosting, these measures are combined across different trees in the ensemble by weighting the bth tree by λ_b. Figure 6 shows the results for the LTCI data.



Alternative scaling functions for the errors also can be used in the boosting iterations. Lee et al. (2015) and Lee and

Figure 6: The Importance of Different Variables for the Gradient-Boosting Algorithm

Lin (2018) have proposed a modified model to take account of the fact that actuarial adjustments are often multiplicative. In this approach, in the non-Gaussian cases, the residuals to be fitted are ratios rather than differences.

Bayesian Additive Regression Trees (BART)

Bayesian Additive Regression Trees (BART) are a Bayesian ensemble method. In the case of a continuous response, the additive regression model is

$$y_i = \sum_{b=1}^{B} g_b(x) + \varepsilon_i,$$

Copyright © 2019 Society of Actuaries

where $\varepsilon_i \sim N(0,\, \sigma^2)$ and $g_1(x), \ldots, g_B(x)$ are *B* tree models with different parameters (split variables and split positions). The intuition is that each $g_b(x)$ function can estimate different aspects of the relationship between *x* and *y*. The depth of a tree is the maximum number of splits between the root node, partition with no splits and the leaves, the partitions at the end of the tree. Boosting approaches to estimation with trees will limit the depth of each tree to avoid overfitting the data (to create weak learners). In this Bayesian modeling approach, a prior is placed on the depth of the tree to avoid each tree becoming "too complicated" and leading to an ensemble of weak learners. This prior that a node at depth *d* is terminal has the simple form

$$p(d) = 1 - \alpha(1 + d)^{-\beta},$$

where $0 < \alpha < 1$ and $\beta \geq 0.$ Increasing the parameter α for fixed β leads to a lower probability of terminating at depth *d* and so encourages a larger tree. Increasing the parameter β for fixed α leads to a higher probability of terminating at depth *d* and so encourages a smaller tree. The authors suggest using the values $\alpha = 0.95$ and $\beta = 2$ as a default, which encourages trees with depth of 2 or 3. Inference can be made in the model using Bayesian backfitting MCMC (Hastie and Tibshirani 2000).

Other Machine-Learning Methods and Topics

Multivariate Adaptive Regression Splines (MARS)

MARS was introduced by Friedman (1991) as a nonparametric regression technique that essentially splits the regression line into m shorter segments:

$$f(x) = \beta_0 + \sum_{i=1}^{m} \beta_i \, h(x_i),$$

where $h(x_i)$, known as basis functions, are of the form $\max(x_i - c,\ 0)$, $\max(c - x_i,\ 0)$ or a product of these expressions. These effectively introduce kinks in the regression function. The model is estimated as a sequence of forward regressions that are likely to lead to overfitting. Then terms are dropped by carrying out a backward regression. While this is easier to visualize in a univariate context, it should be highlighted that MARS is meant to be particularly useful in the multivariate context, because it explores the set of interactions.

To implement MARS in R, we use the package earth. On both our data sets, we find that MARS improves on GLMs, suggesting the considerable nonlinearity required of our model. However, it is outperformed by other methods.

Francis (2003) compared MARS with neural networks on a closed-claims database for personal-injury protection, a database used to predict fraudulent claims or suspicion of fraudulent claims. Her results were not conclusive, due to the relatively small sample size, but she argues for considering MARS because of its relative interpretability.

Neural Nets and Deep Learning

One of the most widely known methods from machine learning is that of neural networks. The earlier literature has focused extensively on "neural nets," and they are now a standard part of the estimation tool kit in many applications. However, for insurance, they typically suffer from the criticism of being difficult to interpret. The neural net is also the underlying component of currently popular classification approaches such as deep learning. However, the issue of interpretability still remains a key one for insurance applications.

We tried a simple neural net for the GLTD data and found that it underperformed other methods, so we have continued to leave it outside the scope of this review. Francis (2003) compares MARS against neural networks for predicting fraud and finds that neural networks are marginally superior on a number of criteria.

Unsupervised Learning

Unsupervised learning is different from the approaches we cover in this review because the objective is to discover patterns in the data without any guidance—in particular, no predicted or dependent variable or model. Francis (2014) provides an overview of the potential for use of unsupervised learning in insurance. Many of the techniques used for supervised learning may also be used for unsupervised learning. For instance, Francis (2016) uses PRIDIT (see Ai et al. 2009) and random forests to help identify suspicious automobile insurance claims. Unsupervised-learning approaches have been used in conjunction with supervised learning, often as a data-processing step. The most popular use of such approaches is to group data into clusters, which we briefly describe in this section.

Cluster Analysis

Clustering is the process of grouping a set of data objects into clusters so that data objects within a cluster have high similarity in comparison to one another but are dissimilar to objects in other clusters. Usually a similarity measure is defined, and the clustering procedure is to optimize this measure locally or globally.

Clustering analysis is widely used in the pre-modeling phase to group granular data into a more manageable number of levels for modeling. It enables a better understanding of data, as features of data may become clearer and more meaningful after the data are grouped into clusters. It also reduces the volatility of data and may help attain more stable rates over time. The result is that the modeler may reduce the number of levels in a rating factor, thereby making it more likely that a model converges and produces statistically significant results. Yao (2016) provides illustrations of various clustering methods with the help of auto insurance claims data in the UK. He also proposes an exposure adjusted hybrid (EAH) method that relies on clustering for modeling claim risk.

There are various clustering methods, including the following:

- Partitioning methods: *k*-means method, *k*-medoids method, and expectation maximization
- Hierarchical methods: agglomerative nesting (AGNES), Divisia analysis (DIANA), balanced iterative reducing and clustering using hierarchies (BIRCH), clustering using representatives (CURE) and Chameleon
- Density-based methods: density-based spatial clustering of application with noise (DBSCAN), ordering points to identify the clustering structure (OPTICS), and density-based clustering (Denclue).
- Grid-based methods
- Kernel and spectral methods

Additional Methods and Combinations

As predictive analytics develops, more and more studies rely on a combination of the core machine learning and statistical methodologies to carry out their analyses. Apart from data processing or preprocessing, previously discussed, other combinations are regularly attempted in insurance. For instance, Kolyshkina et al. (2005) use MARS as a preprocessing step to speed up the estimation of a GLM on car claims data by a significant multiple. Niemerg (2016) evaluates the random GLM on a range of data sets and finds that, although it offers greater interpretability, its prediction ability is not superior to random forests. Kunce and Chatterjee (2017) also propose a "machine learning approach to parameter estimation” that combines a sequence of techniques including *k*-nearest neighbors, kernel regression and relevance vector machines.

Existing Comparative Studies

There are not many large-scale studies comparing a wide range of models for specific insurance applications. Dugas et al. (2003) carried out a rate-making study for a nationwide U.S. automobile insurer and compared a range of models with an emphasis on neural networks. They provide a detailed explanation of the effect of the advantages of neural networks to capture the nonlinearities in insurance problems. They compare a number of models (including multiway tables, a GLM, GLMs with regularization, MARS, neural networks and mixture models). They find that mixture models perform the best overall on criteria such as fairness (reasonable pricing for each subcategory of risk class) and predictive-error minimization.

Duncan et al. (2016) compare the linear model for health care claim cost prediction against a variety of alternatives, including LASSO, MARS, random forests, gradient boosting machines (GBM) and M5 decision trees (a variant of CART). They present comparisons on criteria including the R^2, mean absolute error, quantile truncated mean absolute deviation, and the correlation between actual and predicted values out of sample. They also show that, while for some parameters or features the various models offer similar predictions, this is not the case for all parameters. Using the example of age (among male non-claimants in the current year) as a predictor of next year’s claim costs, they find a wide range of results. While the linear model is the most restricted, the nonlinear models produce curves that are difficult to interpret, highlighting the trade-offs documented in the literature.

Applications to Wider Problems for Actuaries

This review has focused on studies related to pricing but has also touched upon the application of machine learning to other areas of actuarial practice, such as the identification of fraudulent claims. There is a wide range of applications for which machine learning may be useful, particularly related to health care. In addition, the business management areas such as marketing also benefit from the use of machine learning in, for instance, identifying appropriate products for consumers. Studies investigating such management applications have not been covered in this report.

Estimation Packages

Most of the methods described in this review are readily applied using packages available in the open-source software R. Those used in the empirical study have been listed in the relevant sections, with the code to call them provided in the reports accompanying this document. Other packages in R are listed at the Comprehensive R Archive Network website, https://cran.r-project.org/web/views/MachineLearning.html.

Due to the computational intensiveness and memory required to apply machine learning to large data sets, code is often developed in other languages, including Python. We are not aware of a collated source for such code, so one would need to search for the appropriate code on a case-by-case basis. Most commercial providers of statistical software (e.g., MATLAB and SAS) also have well-developed machine-learning toolboxes.

Copyright © 2019 Society of Actuaries

Challenges

Many comparative studies have highlighted the advantages of specific machine-learning methods, and few have compared their performance in different dimensions. There is a plethora of methods available, and researchers are increasingly using them in various combinations to address particular problems. Ali (2016) discusses a classification of the methods and possible applications. Guo (2003) offers an example of a study where the performance of machine-learning methods is underwhelming, partly because the data used in the study is artificial.

An important judgment to make with respect to machine-learning methods is when it is appropriate to use them. For this, the goals of the study and an exploratory analysis of the data are essential. One of the first questions to consider is how interpretable we want our model to be. In terms of data, the common considerations are nonlinearity, interactions and sparse or missing data.

Interpretation

Brockett and Golden (2007) and Golden et al. (2016) highlight the importance of interpretation by discussing social, psychological and biomedical characteristics as the basis behind the use of over 30 credit-score-related variables to predict claims. While credit scores have been shown to predict both incidence and severity of claims, some states have legislated not to allow their use in setting insurance premiums. This highlights the potential that, while machine learning might identify useful predictive features in the data, the incorporation of new variables from the growing sets of big data will need to be justified on economic and legal grounds.

Discovering Nonlinear Relationships, Dealing With Large Numbers of Variables and Sparse Data

Very often, the problem we face is feature selection. Approaches to feature selection go a long way toward addressing the overfitting problems in most models, particularly the nonlinear ones. They also help with the standard GLMs approach. However, actuaries have been practicing other ways of dealing with large numbers of variables, including feature design, which is currently of great interest to the machine-learning community. For instance, the grouping of various credit report variables to assign a credit-related score would count as feature design.

Conclusion

There is a large and growing literature on machine-learning methods in general, but only limited evidence of their use in insurance problems. The potential for their use is significant, given the complex classification problems and nonlinear regression relationships seen in insurance data. An increase in the availability of larger data sets also offers opportunities to adopt machine-learning methods suited to dealing with differently shaped data (e.g., tall, fat, sparse).

We have reviewed the existing literature on the application of machine-learning methods to problems in insurance claim prediction. We have also provided two illustrative exercises using data publicly available from the SOA. Although the prediction problems from the two data sets are similar, we find that different approaches work better, reflecting the degree of nonlinearity in the relationship among other factors. In both cases, however, GLMs are outperformed by the other models within the limited interpretability of the illustrations. This is despite the fact that regularization of a GLM offers some improvement in the case of the LTCI data. Thus, it is difficult to conclude that a particular method is superior in general, more so given the limited transparency of some of the methods. Achievement of reliable model comparisons will require more robustness analyses on a variety of data sets.

One of the observations that arise from our survey is that only a limited amount of data is available to researchers for robust comparisons of the various methods applicable to insurance problems. Data sets with large numbers of

Copyright © 2019 Society of Actuaries

variables and large numbers of records are regularly used in the insurance industry. It would be beneficial to develop some standard shared databases that can be used to test and compare model predictions.

Copyright © 2019 Society of Actuaries

References

Ai, J., P. L. Brockett and L. L. Golden. 2009. Assessing Consumer Fraud Risk in Insurance Claims With Discrete and Continuous Data. *North American Actuarial Journal* 13: 438–458.

Ali, S.D. 2016. Machine Learning: An Analytical Invitation to Actuaries. *Predictive Analytics and Futurism* 14: 24-27.

Anderson, D., S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher T. and Neeza. 2007. *Practitioner's Guide to Generalized Linear Models.* Casualty Actuarial Society (CAS), Syllabus Year: 2010, Exam Number: 9, 1–116.

Bodnar, V., M. Morton, B. Williams and T. Wood. 2015. *Long Term Care Experience Basic Table Development*. Society of Actuaries. URL: https://www.soa.org/experience-studies/2015/2000-2011-ltc-experience-basic-table-dev/

Breiman, L. 2001. Random forests. *Machine learning* 45: 5-32.

Brockett, P. L., and L. L. Golden. 2007. Biological and Psychobehavioral Correlates of Credit Scores and Automobile Insurance Losses: Toward an Explication of Why Credit Scoring Works. *Journal of Risk and Insurance* 74: 23–63.

Brockman, M., and T. Wright. 1992. Statistical Motor Rating: Making Effective Use of Your Data. *Journal of the Institute of Actuaries* 119: 457–543.

Carbonetto, P., X. Zhou and M. Stephens. 2017. varbvs: Fast Variable Selection for Large-Scale Regressions. arXiv: 1709.06597.

Caruana, R., and A. Niculescu-Mizil. 2006. An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning*: 161–168.

Chipman, H. A., E. I. George and R. E. McCulloch. 2010. BART: Bayesian Additive Regression Trees. *Annals of Applied Statistics* 4: 266–298.

De Jong, P., and G. Z. Heller. 2008. *Generalized Linear Models for Insurance Data.* Cambridge: Cambridge University Press.

Dean, C. G. 2014. Generalized Linear Models. In Frees et al., *Predictive Modeling Applications in Actuarial Science*, vol. 1. Cambridge: Cambridge University Press.

Derrig, R. A., and L. Francis. 2008. Distinguishing the Forest from the TREES: A Comparison of Tree Based Data Mining Methods. *Variance* 2: 184–208.

Dugas, C., Y. Bengio, N. Chapados, P. Vincent, G. Denoncourt and C. Fournier. 2003. Statistical Learning Algorithms Applied to Automobile Insurance Ratemaking. In Shapiro and Jain, *Intelligent and Other Computational Techniques in Insurance*, Singapore: World Scientific

Duncan, I., M. Loginov and M. Ludkovski. 2016. Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs. *North American Actuarial Journal* 20, no. 1: 65–87, DOI: 10.1080/10920277.2015.1110491.

Ewald, M., and Q. Wang. 2015. Predictive Modeling: A Modeler’s Introspection. Society of Actuaries. URL: https://www.soa.org/research-reports/2015/2015-predictive-modeling/

Francis, L. A. 2003. Martian Chronicles: Is MARS better than Neural Networks? *Casualty Actuarial Society Forum*, Winter, 253–320.

———. 2014. Unsupervised Learning. In Frees et al., *Predictive Modeling Applications in Actuarial Science*, vol. 1. Cambridge: Cambridge University Press.

———. 2016. Application of Two Unsupervised Learning Techniques to Questionable Claims: PRIDIT and Random Forest. In Frees et al., *Predictive Modeling Applications in Actuarial Science*, vol. 2 Cambridge: Cambridge University Press.

Frees, E. W., R. A. Derrig and G. Meyers, eds. 2014. *Predictive Modeling Applications in Actuarial Science*, vol. 1, *Predictive Modeling Techniques.* Cambridge: Cambridge University Press.

———, eds. 2016. *Predictive Modeling Applications in Actuarial Science*, vol. 2, *Case Studies in Insurance.* Cambridge: Cambridge University Press.

Frees, E. W. J., G. Meyers and A. D. Cummings. 2013. Insurance Ratemaking and a Gini Index. *Journal of Risk and Insurance* 81: 335–366.

Friedman, J. H. 1991. Multivariate Adaptive Regression Splines. *Annals of Statistics* 19: 1.

Golden, L. L., P. L. Brockett, J. Ai and B. Kellison. 2016. Empirical Evidence on the Use of Credit Scoring for Predicting Insurance Losses With Psychosocial and Biochemical Explanations. *North American Actuarial Journal* 20, no. 3: 233–251, DOI:10.1080/10920277.2016.1209118.

Guelman, L. 2012. Gradient Boosting Trees for Auto Insurance Loss Cost Modeling and Prediction. *Expert Systems With Applications* 39, no. 3: 3659–3667.

Guo, L. 2003. Applying Data Mining Techniques in Property Casualty Insurance. Casualty Actuarial Society, Data Management, Quality, and Technology Call Papers and Ratemaking Discussion Papers, Winter Forum. URL: https://www.casact.org/pubs/forum/03wforum/03wf001.pdf

Hastie, T., and R. Tibshirani. 2000. Bayesian Backfitting (With Comments and a Rejoinder by the Authors). *Statistical Science* 15, no. 3: 196–223.

Hastie, T., R. Tibshirani and J. Friedman. 2009. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction.* Springer Series in Statistics. New York: Springer.

Hastie, T., R. Tibshirani and J. Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations.* Monographs on Statistics and Applied Probability 143. Boca Raton, FL: CRC Press.

Khandani, A. E., A. J. Kim and A. W. Lo. 2010. Consumer Credit-Risk Models via Machine-Learning Algorithms. *Journal of Banking and Finance* 34, no. 11: 2767–2787.

Kolyshkina, I., S. Wong and S. Lim. 2005. Enhancing Generalised Linear Models With Data Mining. Casualty Actuarial Society Discussion Paper Program. URL: https://www.casact.org/pubs/dpp/dpp04/04dpp279.pdf

Kopinsky, M. 2017. Predicting Group Long Term Disability Recovery and Mortality Rates Using Tree Models. Society of Actuaries. URL: https://www.soa.org/experience-studies/2017/2017-gltd-recovery-mortality-tree/

Kunce, J., and S. Chatterjee. 2017. A Machine-Learning Approach to Parameter Estimation. CAS Monograph Series no. 6. Arlington: Casualty Actuarial Society.

Copyright © 2019 Society of Actuaries

Lee, S., S. Lin and K. Antonio. 2015. Delta Boosting Machine and Its Application in Actuarial Modeling. Paper presented at the ASTIN, AFIR/ERM and IACA Colloquia of the International Actuarial Association, August 23–27, 2015, Sydney.

Lee, S. C. K., and S. Lin. 2018. Delta Boosting Machine With Application to General Insurance. *North American Actuarial Journal* 22, no. 3: 405–425.

McCullagh, P., and J. A. Nelder. 1989. *Genaralized Linear Models*. 2nd ed. Monographs on Statistics and Applied Probability 37. Boca Raton, FL: Chapman & Hall/CRC.

Miller, H. 2015. A Discussion on Credibility and Penalized Regression, With Implications for Actuarial Work. Paper presented at the ASTIN, AFIR/ERM and IACA Colloquia of the International Actuarial Association, August 23–27, 2015, Sydney.

Mullainathan, S., and J. Spiess. 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* 31, no. 2: 87–106.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Niemerg, M. 2016. Beyond Multiple Regression. *Predictive Analytics and Futurism*, no. 13 (July): 15-17.

Ohlsson, E., and B. Johansson. 2010. *Non-Life Insurance Pricing With Generalized Linear Models.* Berlin: Springer.

Parkes, S. K. 2015. Producing Actionable Insights From Predictive Models Built Upon Condensed Electronic Medical Records. Society of Actuaries 2014 Call for Articles, pp. 7–11. URL: https://www.soa.org/essays-monographs/research-2015-predictive-analytics.pdf

Randazzo, J., and J. P. Kinney. 2015. Predicting Emergency Room Frequent Flyers. Society of Actuaries 2014 Call for Articles, pp. 3–6. URL: https://www.soa.org/essays-monographs/research-2015-predictive-analytics.pdf

Rempala, G. A., and R. A. Derrig. 2005. Modeling Hidden Exposures in Claim Severity via the EM Algorithm. *North American Actuarial Journal* 9, no. 2: 108–128.

Shapiro, A. F. 2000. A Hitchhiker's Guide to the Techniques of Adaptive Nonlinear Models. *Insurance: Mathematics and Economics* 26, nos. 2–3: 119–132.

Shehadeh , M., R. Kokes and G. Hu. 2016. Variable Selection Using Parallel Random Forest for Mortality Prediction in Highly Imbalanced Data. Society of Actuaries Predictive Analytics 2016 Call for Essays, pp. 13–16. URL: https://www.soa.org/essays-monographs/research-2016-predictive-analytics-call-essays.pdf

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)* 58, no. 1: 267–288.

Varian, H. R. (2014). Big data: New tricks for econometrics. The Journal of Economic Perspectives 28 (2), 3-27.

Wu, C. S. P., and J. C. Guszcza. 2003. Does Credit Score Really Explain Insurance Losses? Multivariate Analysis From a Data Mining Point of View. In *Proceedings of the Casualty Actuarial Society* (March): 113–138.

Copyright © 2019 Society of Actuaries

Xu, R., D. Lai, M. Cao, S. Rushing and T. Rozar. 2015. Lapse Modeling for the Post-Level Period: A Practical Application of Predictive Modeling. Society of Actuaries. URL: https://www.soa.org/research-reports/2015/lapse-2015-modeling-post-level

Yao, J. 2016. Clustering in General Insurance Pricing. In Frees et al., *Predictive Modeling Applications in Actuarial Science*, vol. 2. Cambridge: Cambridge University Press.

Copyright © 2019 Society of Actuaries

About The Society of Actuaries

The Society of Actuaries (SOA), formed in 1949, is one of the largest actuarial professional organizations in the world, dedicated to serving 32,000 actuarial members and the public in the United States, Canada and worldwide. In line with the SOA Vision Statement, actuaries act as business leaders who develop and use mathematical models to measure and manage risk in support of financial security for individuals, organizations and the public.

The SOA supports actuaries and advances knowledge through research and education. As part of its work, the SOA seeks to inform public-policy development and public understanding through research. The SOA aspires to be a trusted source of objective, data-driven research and analysis with an actuarial perspective for its members, industry, policy makers and the public. This distinct perspective comes from the SOA as an association of actuaries, who have a rigorous formal education and direct experience as practitioners as they perform applied research. The SOA also welcomes the opportunity to partner with other organizations in our work where appropriate.

The SOA has a history of working with public-policy makers and regulators in developing historical experience studies and projection techniques as well as individual reports on health care, retirement and other topics. The SOA's research is intended to aid the work of policy makers and regulators and follow certain core principles:

Objectivity: The SOA's research informs and provides analysis that can be relied upon by other individuals or organizations involved in public-policy discussions. The SOA does not take advocacy positions or lobby specific policy proposals.

Quality: The SOA aspires to the highest ethical and quality standards in all of its research and analysis. Our research process is overseen by experienced actuaries and nonactuaries from a range of industry sectors and organizations. A rigorous peer-review process ensures the quality and integrity of our work.

Relevance: The SOA provides timely research on public-policy issues. Our research advances actuarial knowledge while providing critical insights on key policy issues, and thereby provides value to stakeholders and decision makers.

Quantification: The SOA leverages the diverse skill sets of actuaries to provide research and findings that are driven by the best available data and methods. Actuaries use detailed modeling to analyze financial risk and provide distinct insight and quantification. Further, actuarial standards require transparency and the disclosure of the assumptions and analytic approach underlying the work.

Society of Actuaries
475 N. Martingale Road, Suite 600
Schaumburg, Illinois 60173
www.SOA.org

Copyright © 2019 Society of Actuaries