*Astin Bulletin* (2023), **53**, pp. 580–595
doi:10.1017/asb.2023.26

International Actuarial Association
Association Actuarielle Internationale

RESEARCH ARTICLE

# A hybrid data mining framework for variable annuity portfolio valuation

Hyukjun Gweon and Shu Li

Department of Statistical and Actuarial Sciences, Western University, London, ON, Canada
**Corresponding author:** Hyukjun Gweon; Email: hgweon@uwo.ca

**Received:** 26 August 2022; **Revised:** 2 May 2023; **Accepted:** 27 June 2023; **First published online:** 28 July 2023

**Keywords:** Variable annuity; portfolio valuation; bootstrap aggregating; random forest; hybrid framework

## Abstract

A variable annuity is a modern life insurance product that offers its policyholders participation in investment with various guarantees. To address the computational challenge of valuing large portfolios of variable annuity contracts, several data mining frameworks based on statistical learning have been proposed in the past decade. Existing methods utilize regression modeling to predict the market value of most contracts. Despite the efficiency of those methods, a regression model fitted to a small amount of data produces substantial prediction errors, and thus, it is challenging to rely on existing frameworks when highly accurate valuation results are desired or required. In this paper, we propose a novel hybrid framework that effectively chooses and assesses easy-to-predict contracts using the random forest model while leaving hard-to-predict contracts for the Monte Carlo simulation. The effectiveness of the hybrid approach is illustrated with an experimental study.

## 1. Introduction

Variable annuity (VA) is a modern long-term life insurance product designed as an investment vehicle for the purposes of retirement planning (Hardy, 2003). As a protection against the fluctuation (generally the downside risk) of the investment, VA provides certain guaranteed minimum death and living benefits regardless of fund performance. For instance, the Guaranteed Minimum Death Benefit (GMDB) offers a policyholder the greater of a guaranteed minimum amount and the balance of the investment account upon the death of the policyholder, while a Guaranteed Minimum Maturity Benefit (GMMB) offers the same upon the maturity of the contract. The Guaranteed Minimum Accumulation Benefit (GMAB) will reset the minimum guarantee amount at renewal times. The Guaranteed Minimum Income Benefit (GMIB) promises the minimum income streams when annuitized at the payout phase (e.g., after retirement), whereas the Guaranteed Minimum Withdrawal Benefit (GMWB) allows for systematic withdrawals without penalty. In addition, the popularity of VA is also partially due to its eligibility for tax deferral advantages, since the majority of sales on the US market relate to the retirement savings plans. According to the Secure Retirement Institute U.S. Individual Annuities Sales Survey, the sales of variable annuities amounted to \$125.6 billions in the year of 2021, which was 27% higher than the previous year, despite the circumstances of the pandemic. For the fair market valuation of the guarantees embedded in a single variable annuity, interested readers are referred to Feng et al. (2022) and the reference therein about the stochastic modeling of embedded guarantees and its valuation via different actuarial approaches.

Insurance companies are exposed to the investment risk through the minimum guaranteed benefits embedded in variable annuity contracts, and thus, one important risk management strategy in practice is dynamic hedging, which requires the efficient valuation of the large portfolio of variable annuity contracts. In addition to the financial risk, VAs also carry the interest rate risk and policyholder lapse or surrender risk due to its long-term nature, as well as the mortality and longevity risk as a life insurance product. Therefore, the closed-form solution for the fair market valuation of VA is not available for most cases. To integrate all variations into the valuation, insurance companies rely on the Monte Carlo simulation in practice. However, the Monte Carlo simulation method is time consuming and computationally intensive; see, for example, Gan and Valdez (2018). Considering the complexity of the product design and the requirement of valuation and dynamic hedging, the workload of computation (dealing with hundreds of thousands of VA contracts) through Monte Carlo simulation grows extensively.

Recent progress in valuation of large VA portfolios uses modern data mining techniques focused on predictive analytics. Existing data mining frameworks include metamodeling (Gan, 2013, 2022) and active learning (Gweon and Li, 2021). In such a framework, the final assessment of a VA portfolio is obtained by a machine learning model that has been trained on a set of example data. Many predictive modeling algorithms have been examined for effective valuations of a large VA portfolio (see Section 2 for literature review).

In this paper, our primary target is to address situations where a highly accurate valuation of a large VA portfolio is required (e.g., $R^2$ of 0.99 is desired), which, to our best knowledge, has not been discussed in the previous literature. More specifically, in existing frameworks including metamodeling and active learning with some chosen machine learning methods, the improvement of the overall quality of a large VA portfolio assessment requires more example data that are fed to the predictive model. Increasing the data size arises two challenges: (1) the computation time required for constructing the predictive model increases, and (2) it is unclear how to determine the size of the training data to achieve the desired predictive performance. Neither the metamodeling nor active learning approach can address the two challenges simultaneously. As such, we propose a hybrid data mining framework that can achieve highly accurate prediction results by selectively using the predictive model for the assessment of "easy-to-predict" contracts and the Monte Carlo engine for the assessment of "hard-to-predict" contracts. Prediction uncertainty metrics are designed to effectively divide easy/hard-to-predict groups. Also, our proposed approach is informative in terms of estimating the target accuracy of the portfolio assessment. Therefore, under the hybrid framework, the two aforementioned practical challenges can be addressed while keeping the size of data for model training small. The empirical results demonstrate that the proposed hybrid approach contributes to a substantial computational cost saving with the minimized prediction errors. Comparing to the existing metamodeling approach, the advantages of our hybrid approach are seen in terms of both predictive performance and runtime.

The main contributions of this paper are in three-folds. First, we develop a novel hybrid data mining framework based on random forest, which complements the existing ones (such as metamodeling and active learning frameworks) with the applications in the insurance field. Second, we design a metric that provides the expected performance of the hybrid approach at any fraction of regression-based prediction. Our proposed approach helps to address the two aforementioned practical challenges without expanding the size of data for predictive model training while achieving the desired accuracy. Third, our empirical results show that the proposed hybrid approach is effective for valuing a large VA contract portfolio such that the targeted prediction error is reached with a vast reduction in Monte Carlo simulation.

The rest of the paper is organized as follows. In Section 2, we provide the literature review on the data mining frameworks for effective VA valuation, as well as the concept of semi-automated classification which links to the proposed hybrid approach. Section 3 presents the details of the proposed hybrid data mining framework, In particular, we discuss the measurement for prediction uncertainty and the expected model performance. In Section 4, we demonstrate the effectiveness of the hybrid approach using a synthetic VA dataset and further make comparisons with the metamodeling framework. The final section concludes the paper.

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press

## 2. Literature review

This section provides a brief review of (1) existing data mining frameworks for dealing with the computational challenges associated with the valuation of large VA portfolios, and (2) semi-automated classification related to the proposed hybrid approach.

(1) Data mining methods using statistical models aim to dramatically reduce the number of VA contracts valued by Monte Carlo simulation. A common approach is the metamodeling framework that has four sequential modeling stages (Barton, 2015): (a) (sampling stage) choosing a subset of the portfolio; (b) (labeling stage) running the Monte Carlo simulation to compute fair market values (FMVs) of the chosen VA contracts; (c) (regression stage) fitting a predictive regression model to the VA contracts in the subset; and (d) (prediction stage) predicting FMVs for the rest of the contracts in the portfolio using the fitted regression model. The purpose of the sampling stage (a) is to efficiently divide a large dataset into many clusters or groups from which representative contracts are chosen. Several unsupervised learning algorithms have been proposed including the truncated fuzzy c-means algorithm (Gan and Huang, 2017), conditional Latin hypercube sampling (Gan and Valdez, 2018), and hierarchical $k$-means clustering (Gan and Valdez, 2019). Popular supervised learning algorithms, such as (Gan, 2013), GB2 (Gan and Valdez, 2018), group LASSO (Gan, 2018), and tree-based approaches (Xu et al., 2018; Gweon et al., 2020; Quan et al., 2021), have been examined for use in the regression stage (c). In metamodeling, the FMVs of most contracts in the portfolio are estimated using the regression model. A simple variation of the metamodeling approach is model points (Goffard and Guerrault, 2015) that divide policies into non-overlapping groups based on an unsupervised learning algorithm and assign a representative prediction value (e.g., the sample mean) to each group.

Recently, Gweon and Li (2021) proposed another data mining framework based on active learning (Cohn et al., 1994; Settles, 2010). The goal of active learning is to achieve the highest prediction accuracy within a limited budget for the labeled data. To achieve this goal, a regression model is initially fitted to a small number of labeled representative contracts. The fitted model is then used to iteratively and adaptively select a batch of informative contacts from the remaining unlabeled contracts. The selected contracts are assessed using Monte Carlo simulation and added to the labeled data so that the regression model is updated with the augmented labeled data. Unlike metamodeling, the active learning framework allows the regression model to actively choose and learn from contracts for which the current model does not perform well.

(2) The fundamental idea of the hybrid approach proposed in Section 4 is inspired by semi-automatic classification (Schonlau and Couper, 2016) that has been studied in survey data classification. Text answers in surveys are difficult to analyze and therefore are often manually classified into different classes or categories. With a large amount of data, manual classification becomes time consuming and expensive as it requires professional experienced human labelers. While the use of statistical learning methods reduces the total cost of coding, fully automated classification of text answers to open-ended questions remains challenging. This is a problem for researchers and survey practitioners who value accuracy over low cost. To address this problem, semi-automated classification uses statistical approaches to perform partially automated classification. In this way, easy-to-classify answers are automatically categorized and hard-to-classify answers are manually categorized. The idea of semi-automated procedure has been applied to single-labeled survey data (Schonlau and Couper, 2016; Gweon et al., 2017) and multi-labeled survey data (Gweon and Wenemark, 2020).

## 3. The hybrid framework for valuing large VA portfolios

### 3.1. The hybrid framework

Our goal is to achieve highly accurate prediction of the fair market values (FMVs) of the large portfolio of VAs via a combination of the predictive regression model and Monte Carlo simulation engine. The proposed hybrid valuation framework is summarized in the following four steps:

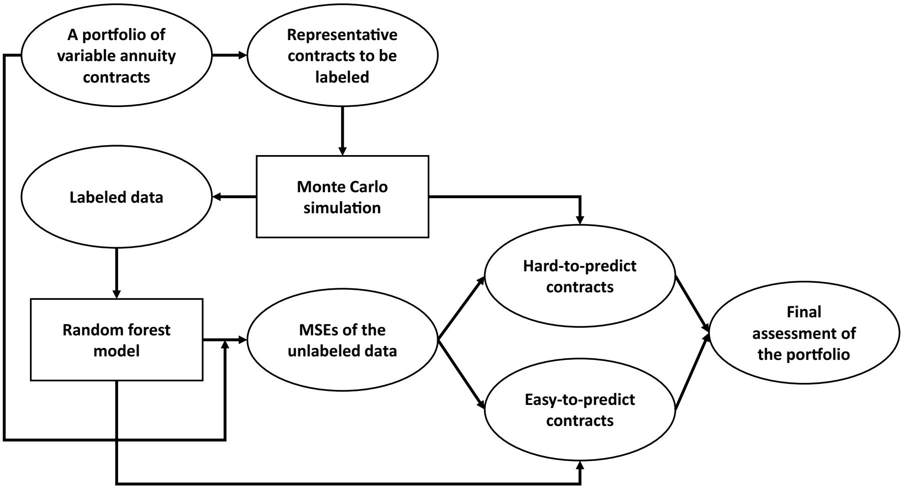
A portfolio of variable annuity contracts
Representative contracts to be labeled
Labeled data
Monte Carlo simulation
Hard-to-predict contracts
Random forest model
MSEs of the unlabeled data
Final assessment of the portfolio
Easy-to-predict contracts

***Figure 1.*** *An illustration of the hybrid data mining framework.*

1. Select a set of representative VA contracts from the portfolio. An unsupervised learning algorithm can be employed for the selection task.
2. Calculate the FMVs of the guarantees for the representative contracts using the Monte Carlo simulation. The resulting labeled data become the training data.
3. Build a regression model (e.g., random forest) using the training data.
4. Use the regression model to predict the FMVs of $100\alpha\%$ of the contracts (for $\alpha \in [0, 1]$) and employ the Monte Carlo simulation for valuing the remaining contracts.

Note that the key difference between the hybrid framework and metamodeling framework is in Step 4, where we introduce the parameter $\alpha \in [0, 1]$. Having $\alpha = 0$ means that all VA contracts in the portfolio are assessed using the Monte Carlo simulation, which reduces to the simulation approach (only), while $\alpha = 1$ corresponds to using the regression-based prediction for all contracts (except the small set of representative contracts in steps 1 and 2). Therefore, the existing metamodeling framework can be viewed as a special case of the hybrid framework at $\alpha = 1$. For $0 < \alpha < 1$, the hybrid approach employs a combination of both approaches. As $\alpha$ changes, there exists a trade-off between computational cost and valuation accuracy. Increasing $\alpha$ results in more contracts being assessed by the regression model whose predictions are fast but come with inevitable errors. Despite the low computational cost, the metamodeling approach ($\alpha = 1$) provides no practical strategy for an effective trade-off between computational cost and valuation accuracy. As $\alpha$ decreases, the overall valuation accuracy can increase at the cost of the increased amount of computing time required for running the Monte Carlo simulation. Hence, two crucial components of the proposed hybrid approach are: the selection of an appropriate value of $\alpha$ and how to determinate the two sub-groups (for regression-based predictions and Monte Carlo valuation). As such, we further specify Step 4 in the following two parts:

4(a) Decide the fraction $0 \leq \alpha \leq 1$ for the regression-based prediction. The choice of parameter $\alpha$ should take into consideration the expected accuracy. Here, by "accuracy" we mean the $R^2$ of the portfolio.

4(b) Use the regression model to predict the FMVs of $100\alpha\%$ of the contracts with the smallest prediction uncertainty (referred to as "easy-to-predict" contracts). We refer to the remaining contracts as "hard-to-predict" contracts, and employ the Monte Carlo simulation for the evaluation.

See Figure 1 for illustration of the proposed hybrid data mining framework.

It is worth discussing the similarity and difference between the proposed framework with a given $\alpha$ and the metamodeling framework that uses $(1-\alpha)100\%$ of the contracts as training data. In both frameworks, $(1-\alpha)100\%$ of the portfolio is evaluated by the MC simulation and the other $100\alpha\%$ by a trained predictive model. That is, both frameworks require the same computational cost for running the MC simulation engine. In metamodeling, splitting the portfolio into the labeled data ($(1-\alpha)100\%$) and remaining data ($100\alpha\%$) is conducted at the first step of the framework. For a split, metamodeling relies on a data clustering or unsupervised learning algorithm that creates a set of representative data using the feature information only. On the other hand, the proposed hybrid method makes the final split (based on $\alpha$) after a predictive model is trained and this allows the trained model to actively identify and assign hard-to-predict contracts to the MC engine (equivalently, assign easy-to-predict contracts to the predictive model). Due to this difference, at a moderate value of $\alpha$, metamodeling requires a much more computing time for training a predictive model compared to the hybrid approach. This is investigated further in Section 4.4.2.

In what follows, we address the two crucial components using random forest as a predictive regression model. More specifically, we will explain the measurement for prediction uncertainty in order to label the "easy-to-predict" and "hard-to-predict" contracts and construct a functional relationship between the parameter $\alpha$ and the expected accuracy of the portfolio valuation which, in turn, becomes useful for the choice of $\alpha$.

### 3.2. Random forests and measuring prediction uncertainty

Consider a portfolio of $N$ VA contracts, $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ where $\mathbf{x}_i \in \mathbb{R}^p$ contains the feature attributes associated with its VA contact. Also, let $Y_i$ be the FMV of the contract $\mathbf{x}_i$. The random forest model assumes the general model form:

$$Y_i = f(\mathbf{x}_i) + \epsilon_i,$$

where $f(\cdot)$ is the underlying regression model and $\epsilon$ is the random error. To estimate the regression function, we use random forest (Breiman, 2001) with regression trees (Breiman, 1984) as the base model. Details about the use of regression trees for variable annuity application are found in Gweon et al. (2020) and Quan et al. (2021).

Let $L$ be the labeled training data of size $n$, $L_b$ be the $b$th bootstrap sample of $L$ and $\widehat{f_{L_b}}(\mathbf{x})$ be a regression tree fitted to $L_b$, for $b = 1, ..., B$. For any unlabeled contract $\mathbf{x}$, the prediction is obtained by averaging all of the $B$ regression trees:

$$\hat{f}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f_{L_b}}(\mathbf{x}).$$

Despite its simplicity, random forest demonstrates promising predictive performance in valuing contracts (Quan et al., 2021; Gweon et al., 2020).

In the hybrid approach, we propose to label the "easy-to-predict" and "hard-to-predict" via prediction uncertainty. A common measurement for uncertainty is the mean square error (MSE) for the underlying function that is defined as

$$\mathrm{MSE}(\hat{f}(\mathbf{x})) = E\left((\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2\right). \tag{3.1}$$

By the bias-variance decomposition, we have

$$\begin{aligned}\mathrm{MSE}(\hat{f}(\mathbf{x})) &= E\left((\hat{f}(\mathbf{x}) - E(\hat{f}(\mathbf{x})) + E(\hat{f}(\mathbf{x})) - f(\mathbf{x}))^2\right)\\ &= \left(E(\hat{f}(\mathbf{x})) - f(\mathbf{x})\right)^2 + E\left((\hat{f}(\mathbf{x}) - E(\hat{f}(\mathbf{x})))^2\right),\end{aligned} \tag{3.2}$$

where the first and second terms are the squared bias and variance, respectively. This decomposition result provides a plug-in estimate of MSE by estimating the bias and variance separately.

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press

Gweon et al. (2020) show that the prediction bias of random forest is not negligible when applied to the VA valuation. The prediction bias can be estimated by bias-correction techniques (Breiman, 1999; Zhang and Lu, 2012; Gweon et al., 2020), where another random forest model is fitted to the out-of-bag (OOB) errors. Following Gweon et al. (2020), for the prediction vector $\mathbf{x}_i$ in the training data, the out-of-bag prediction is defined as

$$\hat{f}^{OOB}(\mathbf{x}_i) = \frac{1}{B_i}\sum_{b=1}^{B}\hat{f}_{L_b}(\mathbf{x}_i)I((\mathbf{x}_i, y_i) \notin L_b),$$

where $I(\cdot)$ is the indicator function, and $B_i$ is the number of bootstrap regression trees for which data point $(\mathbf{x}_i, y_i)$ is not used (i.e., $B_i = \sum_{b=1}^{B} I((\mathbf{x}_i, y_i) \notin L_b)$). Then, another random forest model $\hat{g}(\cdot)$ is fitted to the set of representative VA contracts where the response variable is $Bias(\mathbf{x}) = \hat{f}^{OOB}(\mathbf{x}) - Y$, instead of $Y$. The prediction obtained by the resulting model is the estimated bias. That is,

$$\widehat{Bias}(\hat{f}(\mathbf{x})) = \hat{g}(\mathbf{x}). \tag{3.3}$$

For estimating the variance of random forest, one common method is jackknife-after-bagging (Efron, 1992; Sexton and Laake, 2009) that aggregates all regression trees where the $i$th contract is not included in the construction of the trees. The estimated variance is obtained by

$$\widehat{Var}(\hat{f}(\mathbf{x})) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{f}^{OOB_{-i}}(\mathbf{x}) - \hat{f}^{OOB_*}(\mathbf{x}))^2, \tag{3.4}$$

where

$$\hat{f}^{OOB_{-i}}(\mathbf{x}) = \frac{1}{B_i}\sum_{b=1}^{B}\hat{f}_{L_b}(\mathbf{x})I((\mathbf{x}_i, y_i) \notin L_b),$$

and

$$\hat{f}^{OOB_*}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}^{OOB_{-i}}(\mathbf{x}).$$

The sampling variability of random forests has also been analyzed in, for example, Lin and Jeon (2006), Wager and Efron (2014), and Mentch and Hooker (2016).

### 3.3. Determining $\alpha$ and expected $R^2$

For a VA portfolio with $N$ contracts, denote $S_{RF}$ and $S_{MC}$ as the sets containing contacts with valuations obtained by the regression model and the Monte Carlo simulation, respectively. We use the notation $f(\mathbf{x}_i)$ for the FMV of a contract computed using the Monte Carlo simulation[1], and the notation $\hat{f}(\mathbf{x}_i)$ for the FMV of a contract predicted by the regression model. The $R^2$ of the portfolio, denoted by $R^2_{S_{RF}}$ to be more precise, is obtained by

$$\begin{aligned} R^2_{S_{RF}} &= 1 - \frac{\sum_i (\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i))^2}{\sum_i (f(\mathbf{x}_i) - \bar{f}(\mathbf{x}))^2} \\ &= 1 - \frac{\sum_{\mathbf{x}_i \in S_{RF}} (\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i))^2 + \sum_{\mathbf{x}_i \in S_{MC}} (f(\mathbf{x}_i) - f(\mathbf{x}_i))^2}{\sum_i (f(\mathbf{x}_i) - \bar{f}(\mathbf{x}))^2} \\ &= 1 - \frac{\sum_{\mathbf{x}_i \in S_{RF}} (\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i))^2}{c}. \end{aligned}$$

[1] The FMVs of variable contracts with identical features are assumed to be equal without unexplained errors.

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press

where $c=\sum_i (f(\mathbf{x}_i)-\bar{f}(\mathbf{x}))^2$ is a constant and $\bar{f}(\mathbf{x})=N^{-1}\sum_i f(\mathbf{x}_i)$. Taking the expectation gives

$$E\left(R^2_{S_{RF}}\right)=1-\frac{\sum_{\mathbf{x}_i\in S_{RF}} E\left[(\hat{f}(\mathbf{x}_i)-f(\mathbf{x}_i))^2\right]}{c},$$

where $E\left[(\hat{f}(\mathbf{x}_i)-f(\mathbf{x}_i))^2\right]$ refers to the MSE of the prediction $\hat{f}(\mathbf{x}_i)$ with respect to $f(\mathbf{x}_i)$ by Equation (3.1). Notice that $E\left(R^2_{S_{RF}}\right)$ monotonically decreases as more contracts are assessed by the regression model (i.e., the size of $S_{RF}$ increases). In order to maximize $E\left(R^2_{S_{RF}}\right)$, we seek an optimal set $S^*_{RF}$ with a constraint on its size, that is,

$$\begin{aligned} S^*_{RF} &= \underset{S_{RF}}{\operatorname{argmax}}\ E\left(R^2_{S_{RF}}\right) \\ &= \underset{S_{RF}}{\operatorname{argmax}} \sum_{\mathbf{x}_i\in S_{RF}} E\left[(\hat{f}(\mathbf{x}_i)-f(\mathbf{x}_i))^2\right], \\ &\qquad\qquad\qquad\qquad \text{subject to } |S_{RF}|=\alpha N \text{ for a given } \alpha, \end{aligned}$$

where $|S_{RF}|$ represents the size of the set $S_{RF}$, that is, the number of VA contracts in the set. By selecting the contracts with the smallest MSE values, optimization is achieved over all possible subsets that form the set $S_{RF}$ of a certain size. This provides a crucial rationale for the proposed hybrid approach to select the least uncertain contracts for the random forest-based (RF-based) prediction.

Recall that from the bias-variance decomposition result, we have

$$E\left[(\hat{f}(\mathbf{x}_i)-f(\mathbf{x}_i))^2\right]=Var(\hat{f}(\mathbf{x}_i))+(Bias(\hat{f}(\mathbf{x}_i)))^2.$$

Using random forest, the variance and bias can be separately estimated using the methods described in Section 3; see Equations (3.3) and (3.4). The constant $c=\sum_i (f(\mathbf{x}_i)-\bar{f}(\mathbf{x}))^2$ can be estimated by

$$\hat{c}=N/n\sum_i (f(\mathbf{x}_i)-\bar{f}(\mathbf{x}))^2 I((\mathbf{x}_i,f(\mathbf{x}_i))\notin L). \tag{3.5}$$

Combining those individual estimators in Equations (3.3), (3.4), and (3.5), a plug-in estimate of $R^2_{S_{RF}}$ is

$$\widehat{R}^2_{S_{RF}}=1-\frac{\sum_{\mathbf{x}_i\in S_{RF}}\left[\widehat{Var}(\hat{f}(\mathbf{x}_i))+(\widehat{Bias}(\hat{f}(\mathbf{x}_i)))^2\right]}{\hat{c}}.$$

In addition to the variance of random forest, one may also consider the sample variance of the individual tree predictions, denoted as $Var(\hat{f}^b(\mathbf{x}_i))$. Assuming the pairwise correlation ($\rho_{\mathbf{x}}$) between two regression trees is non-negative, it is known that

$$Var(\hat{f}(\mathbf{x}_i))=\left(\frac{(B-1)\rho_{\mathbf{x}}+1}{B}\right)Var(\hat{f}^b(\mathbf{x}_i))\le Var(\hat{f}^b(\mathbf{x}_i)).$$

Hence, an replacement of $Var(\hat{f}(\mathbf{x}_i))$ with $Var(\hat{f}^b(\mathbf{x}_i))$ results in

$$E\left(R^2_{S_{RF}}\right)\ge 1-\frac{\sum_{\mathbf{x}_i\in S_{RF}}\left[Var(\hat{f}^b(\mathbf{x}_i))+(Bias(\hat{f}(\mathbf{x}_i)))^2\right]}{c}:=E\left(\underline{R}^2_{S_{RF}}\right).$$

As such, $E\left(\underline{R}^2_{S_{RF}}\right)$ serves as a lower bound for the expected $R^2$ of the hybrid approach for the portfolio. An estimate of $E\left(\underline{R}^2_{S_{RF}}\right)$ is

$$\widehat{\underline{R}}^2_{S_{RF}}=1-\frac{\sum_{\mathbf{x}_i\in S_{RF}}\left[\widehat{Var}(\hat{f}^b(\mathbf{x}_i))+(\widehat{Bias}(\hat{f}(\mathbf{x}_i)))^2\right]}{\hat{c}}$$

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press

***Table 1.*** *Summary statistics of the continuous feature variables in the dataset.*

| Variable | Description | Minimum | Mean | Maximum |
|---|---|---|---|---|
| gmwbBalance | GMWB balance | 0 | 35,611.54 | 499,708.73 |
| gbAmt | Guaranteed benefit amount | 0 | 326,834.59 | 1,105,731.57 |
| FundValue1 | Account value of the 1st investment fund | 0 | 33,433.87 | 1,099,204.71 |
| FundValue2 | Account value of the 2nd investment fund | 0 | 38,542.81 | 1,136,895.87 |
| FundValue3 | Account value of the 3rd investment fund | 0 | 26,740.18 | 752,945.34 |
| FundValue4 | Account value of the 4th investment fund | 0 | 26,141.80 | 610,579.68 |
| FundValue5 | Account value of the 5th investment fund | 0 | 23,026.50 | 498,479.36 |
| FundValue6 | Account value of the 6th investment fund | 0 | 35,575.67 | 1,091,155.87 |
| FundValue7 | Account value of the 7th investment fund | 0 | 29,973.25 | 834,253.63 |
| FundValue8 | Account value of the 8th investment fund | 0 | 30,212.11 | 725,744.64 |
| FundValue9 | Account value of the 9th investment fund | 0 | 29,958.29 | 927,513.49 |
| FundValue10 | Account value of the 10th investment fund | 0 | 29,862.24 | 785,978.60 |
| age | Age of the policyholder | 34.52 | 49.49 | 64.46 |
| ttm | Time to maturity in years | 0.59 | 14.54 | 28.52 |

where

$$\widehat{Var}(\hat{f}^b(\mathbf{x}_i)) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{f}_{L_b}(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\right)^2.$$

To conclude, since $E\left(R^2_{S^*_{RF}}\right)$ has a functional relationship with the fraction $\alpha$, either $\widehat{R}^2_{S^*_{RF}}$ $\left(\text{or } \underline{\widehat{R}}^2_{S^*_{RF}}\right)$ or $\alpha$ can be set at a target value which determines the second measure, that is:

- if one targets the model performance of $R^2$, say at least 99% for the portfolio, the hybrid algorithm will determine $\alpha$ and thus the set $S^*_{RF}$ such that $\underline{\widehat{R}}^2_{S^*_{RF}} = 0.99$ in a conservative manner;
- on the other hand, if the parameter $\alpha$ is fixed (for instance, when the budget for the computational cost is limited), the hybrid approach will examine the expected model performance with the optimal set $S^*_{RF}$ through either $\widehat{R}^2_{S^*_{RF}}$ or $\underline{\widehat{R}}^2_{S^*_{RF}}$ to maximize the prediction accuracy.

## 4. Application in variable annuity valuation

### *4.1. A synthetic portfolio*

We examined the proposed hybrid approach using a synthetic VA dataset in Gan and Valdez (2017). The dataset consists of 190,000 VA contracts with 16 feature variables, after removing variables that were identical for all contracts (Gan et al., 2018). The continuous feature variables used for our analysis are summarized in Table 1.

The dataset has two categorical features: gender and product type. The gender ratio is female:male = 40%:60%. There are 19 product types (e.g., variants of GMAB and GMIB), and the dataset contains 10,000 contracts for each product type; see Gan and Valdez (2017) for more details.

Our target response variable is FMV, the difference between the guarantee benefit payoff and the risk charge. Details of how the FMV value of each guarantee is obtained by the MC simulation are found in Gan and Valdez (2017). Figure 2 shows a highly skewed distribution of the FMVs of the 190,000 VA contracts in the portfolio. The skewness is due to the guaranteed payoff being much greater than the charged guarantee fee for many contracts (Gan and Valdez, 2018).

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press

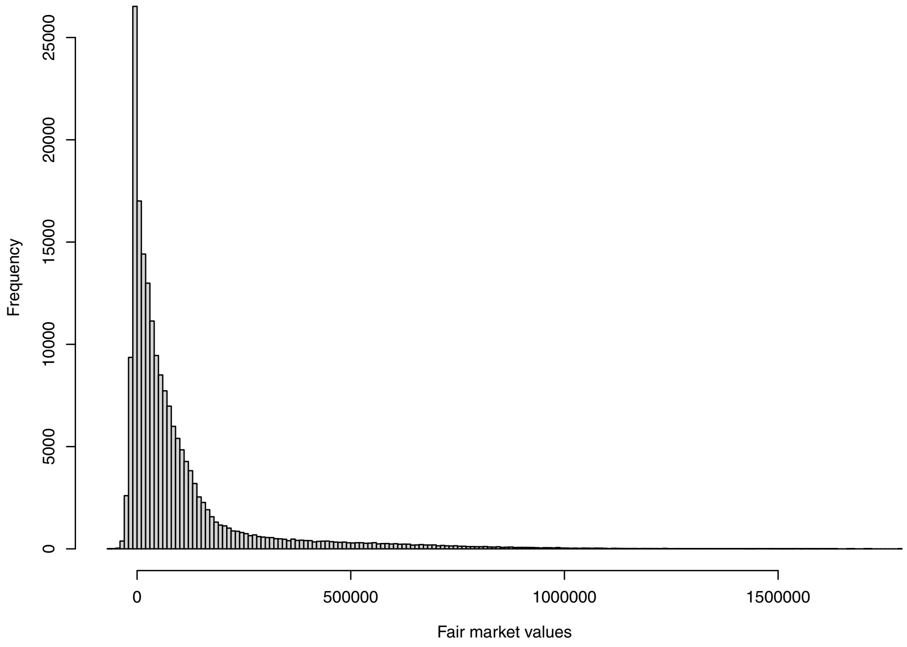

***Figure 2.*** *Histogram of the FMVs of 190,000 VA contracts.*

### *4.2. Experimental setting*

As with any other data mining approach, the hybrid approach requires a set of representative contracts for fitting the regression model. We used the conditional Latin hypercube sampling (Minasny and McBratney, 2006) because it produces reliable results as compared to other unsupervised approaches (Gan and Valdez, 2016). The conditional Latin hypercube sampling algorithm heuristically chooses a subset of the portfolio such that the distribution of the portfolio is maximally stratified. We used the R package clhs (Roudier, 2011) for the implementation in R.

For random forest, we use 300 regression trees that are large enough to reach stable model performance (Gweon et al., 2020; Quan et al., 2021). In addition, we consider all features at each binary split in the tree construction because it achieves the lowest prediction error for the dataset (Quan et al., 2021). As described in (Gweon et al., 2020), prediction biases are estimated by another random forest model with 300 regression trees fitted to the out-of-bag prediction. We use the jackknife-after-random forest estimate (Sexton and Laake, 2009) to estimate the variance of random forest.

The model performance could be affected by some random effects. To mitigate the impact of possible random effects, we ran the experiment 10 times with different seeds.

### *4.3. Evaluation measures*

To measure predictive performance, we consider $R^2$, mean absolute error (MAE), and percentage error (PE):

$$R^2 = 1 - \frac{\sum_i (\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i))^2}{\sum_i (f(\mathbf{x}_i) - \bar{f}(\mathbf{x}))^2},$$

$$\text{MAE} = \frac{1}{N} \sum_i |\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)|,$$

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press

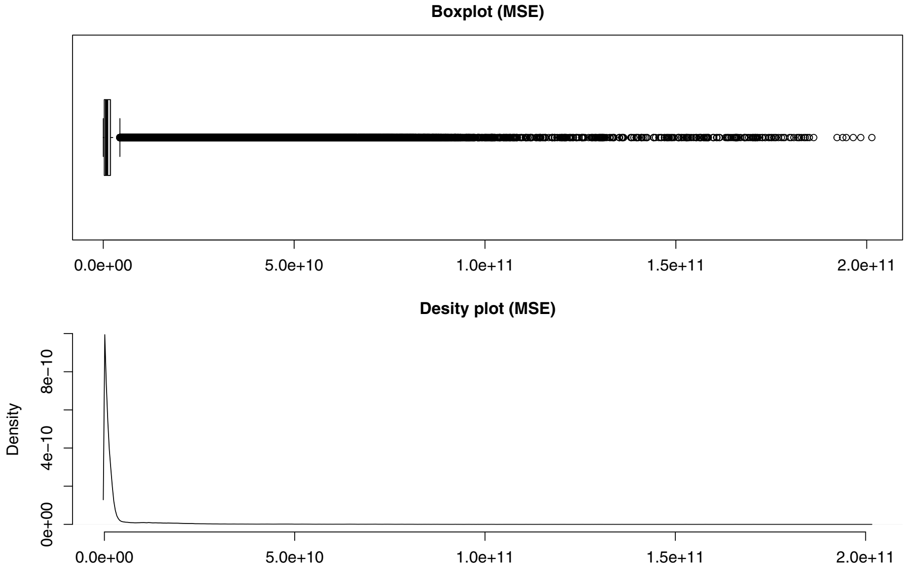

***Figure 3.*** *The boxplot (top) and density plot (bottom) of the estimated MSE of the unlabeled contracts.*

and

$$\mathrm{PE} = \frac{\sum_i f(\mathbf{x}_i) - \sum_i \hat{f}(\mathbf{x}_i)}{\sum_i f(\mathbf{x}_i)},$$

where $\bar{f}(\mathbf{x}) = N^{-1}\sum_i f(\mathbf{x}_i)$. $R^2$ and MAE measure the accuracy of the valuation result at the individual contract level. PE measures the aggregate accuracy of the valuation result where positive and negative prediction errors at the individual contract level offset each other. The result is considered accurate at the portfolio level if the absolute value of PE is close to zero. All evaluation results are performed on the whole portfolio.

## *4.4. Experimental results*

### *4.4.1. An empirical analysis of the hybrid approach*

Figure 3 presents the boxplot and density plot of the MSE values of the VA contracts in the portfolio estimated using the random forest model with $n = 1000$. The distribution is highly skewed with the majority of the values being small. This pattern favors the proposed hybrid approach, as the contracts with small MSE are considered to be easy-to-predict examples and, therefore, are expected to be accurately valued by the random forest model.

Figure 4 shows the performance (in terms of $R^2$) of the hybrid approach as a function of the fraction of RF-based valuation at different sizes of representative labeled data. The contracts with lower MSE estimates were valued first using the random forest model. For example, the fraction $\alpha = 0.2$ means only 20% of the remaining contracts with the lowest MSE are assessed by the random forest model and the other 80% are left for the Monte Carlo simulation. The two estimated $R^2$ values are obtained using the plug-in estimation methods.

As expected, there were trade-offs between accuracy and the fraction of RF-based prediction. We observed that $\widehat{R}^2_{S^*_{RF}}$ tends to be larger than the observed $R^2$ indicating underestimation of MSE (i.e., overestimation of $R^2$). Another observation is that $\underline{\widehat{R}}^2_{S^*_{RF}}$ effectively served as a lower bound of $E(R^2)$ as

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press

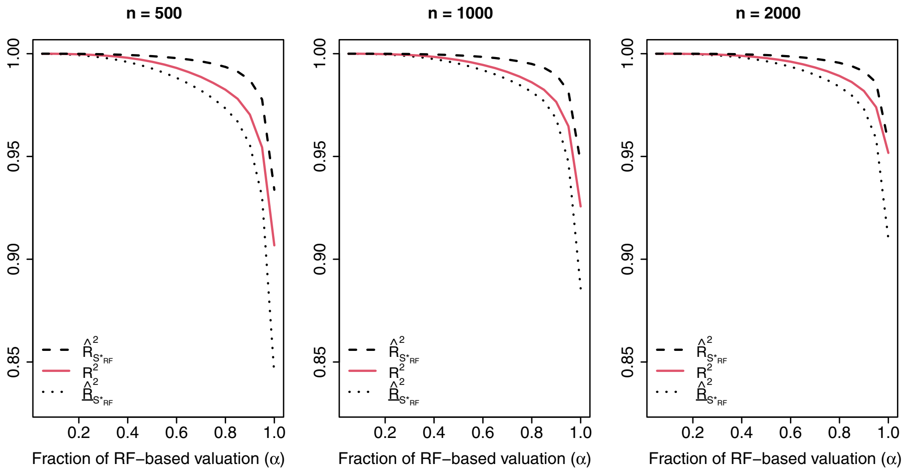

***Figure 4.*** *The estimated and observed $R^2$ values obtained by the hybrid approach.*

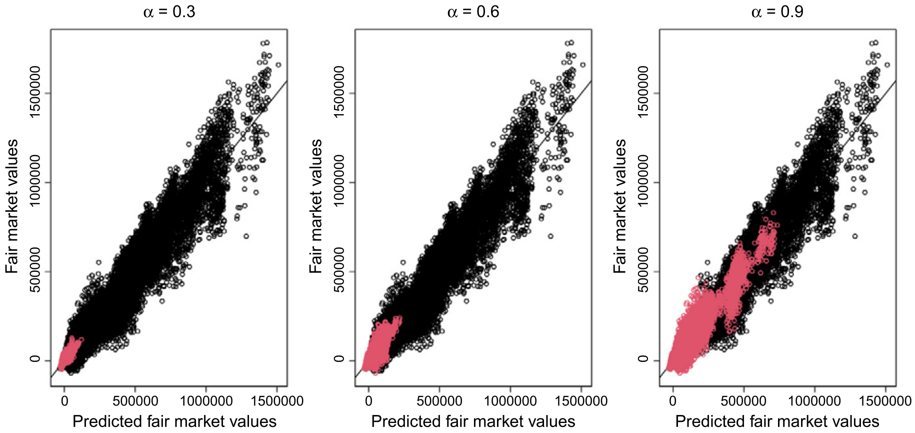

***Figure 5.*** *Scatter plots of the observed FMVs and the values predicted by random forest. The red dots represent RF-based predictions in the hybrid approach.*

$\underline{\widehat{R}}^2_{S^*_{RF}}$ was consistently lower than (and close to) the observed $R^2$. The difference between the observed $R^2$ and $\underline{\widehat{R}}^2_{S^*_{RF}}$ was particularly small when the fraction of RF-based valuation was lower than 0.8.

The slopes of the performance curves became steeper as more contracts were evaluated by random forest. This coincides with our intuition, as the hybrid method prioritized contracts with small expected errors for RF-based valuation. The small reduction in accuracy at small to medium fractions demonstrates the particular effectiveness of the hybrid method with small fractions.

Next, we further investigated easy-to-predict contracts. As shown in Figure 5, most of these have small (predicted) FMVs. This result can be explained by the highly skewed distribution of FMVs in the portfolio (Figure 2), as in the representative labeled data. The random forest model mostly learned from representative contracts with small FMVs, and thus, the trained model was more confident in predicting the contracts similar to the representative contracts as compared to others.

Figure 6 presents the performance of the hybrid approach according to the three evaluation metrics. The model performance was generally improved as the number of representative contracts (i.e., $n$) increased. In addition, greater improvement was observed at large fractions of RF-based prediction

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press

***Table 2.*** *Summary statistics for the hybrid approach ($n = 2,000$) as a function of various thresholds $\left(\underline{\widehat{R}}^2_{S^*_{RF}}\right)$. The estimated times are in minutes.*

| $\underline{\widehat{R}}^2_{S^*_{RF}}$ | $\alpha$ | $R^2$ | MAE | PE | Estimated Time (MC) |
|---|---|---|---|---|---|
| 0.998 | 0.40 | 0.999 | 2343.022 | −0.003 | 3755.160 |
| 0.995 | 0.55 | 0.997 | 4536.604 | −0.006 | 2816.370 |
| 0.990 | 0.70 | 0.993 | 7626.603 | −0.009 | 1885.680 |
| 0.980 | 0.85 | 0.986 | 11924.067 | −0.011 | 938.790 |
| 0.965 | 0.95 | 0.974 | 16561.058 | −0.014 | 312.930 |
| 0.915 | 0.99 | 0.952 | 20980.085 | −0.014 | 62.586 |

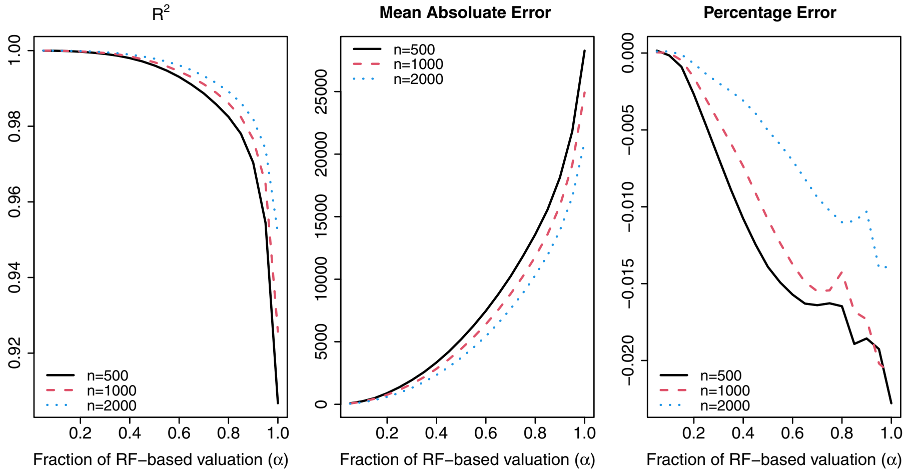

***Figure 6.*** *Performance of the hybrid approach with different sizes of representative data. For $R^2$ (left), higher is better. For mean absolute error (middle), lower is better. For percentage error (right), lower absolute value is better.*

(i.e., as $\alpha$ increases). This implies that even with a small $n$ size (e.g., $n = 500$), the random forest model performed well on easy-to-predict contracts.

In practice, the fraction parameter $\alpha$ can be determined based on the desirable lower bound of overall accuracy $\underline{\widehat{R}}^2_{S^*_{RF}}$. The performance of the hybrid approach at $n = 2,000$ is summarized in Table 2. For example, if one requires $R^2$ of at least 0.99, the hybrid approach can use the random forest model for up to 70% of the contracts and the Monte Carlo simulation for the remaining 30%. Although decreasing $\alpha$ improved the overall valuation accuracy, the price is increased computation time[2] for running the MC simulation for the $(1 - \alpha)100\%$ of the contracts. This trade-off between the prediction accuracy and time efficiency suggests that practitioners consider both the expected accuracy and computation time when determining an appropriate value of $\alpha$.

### *4.4.2. A comparison with the metamodeling approach*

To further investigate the effectiveness of the proposed hybrid approach, we compared our results with the metamodeling framework, particularly with three regression approaches namely, RF, GB2 and group LASSO (GLASSO). To make a fair comparison, for metamodeling considered here, $(1 - \alpha)100\%$ of the

[2] The computation times for the MC simulation in Table 2 were estimated based on the results of Gan and Valdez (2017) and Gan (2018) assuming the use of a single CPU. More powerful computing resources will result in less computing times.

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press

contracts were selected using the conditional Latin hypercube sampling method. The chosen contracts were used to train RF, GB2, and GLASSO models. Each of the trained models was then used to predict the FMVs of the remaining $100\alpha\%$ contracts in the portfolio. This setting allows a reasonable comparison between the metamodeling and hybrid frameworks at the fraction of RF-based valuation $\alpha$ so that both frameworks eventually require the MC simulation for the same amount ($(1-\alpha)100\%$) of the variable annuities in the portfolio. More precisely, in the metamodeling framework, the $(1-\alpha)100\%$ of the data, used for model training, are evaluated by MC simulation, whereas the hybrid approach starts from a small subset of the data of size $n$ (i.e., $n << N$) and then decides the hard-to-predict group, which contains the $(1-\alpha)100\%$ (or more precisely $(1-\alpha)N-n$ contracts) of the data to be evaluated by MC simulation.

The comparison results in terms of all performance measures and runtime at $\alpha = 0.5$ and 0.7 are presented in Table 3. For the hybrid approach, we used $n = 2,000$ and $n = 5,000$. The hybrid framework outperformed the metamodeling approaches in terms of $R^2$ and MAE. All approaches performed well in PE, with fairly small differences between all methods. For metamodeling, even though a large amount of data (e.g., when $\alpha = 0.5$, we had $n = (1-\alpha) \times N = 95,000$ representative VA contracts) were used to fit the predictive models, the trained models still produced prediction errors on the individual contracts of the unlabeled data due to hard-to-predict contracts. On the other hand, the proposed hybrid approach relied on far fewer ($n = 2,000$ or 5,000) representative contracts for the RF model and the trained model achieved high predictive accuracy for easy-to-predict contracts. The results showed the effectiveness of the hybrid framework particularly at the individual policy level. Also, the metamodeling approach required a significantly greater runtime for selecting the representative contracts and training the predictive model than the proposed approach. The hybrid approach showed a much faster and consistent runtime performance at different $\alpha$ values thanks to the fact that the size of representative data remains small in all situations. At $\alpha = 0.5$, the metamodeling approach with RF required more than 10 h to complete the RF-based prediction, whereas the hybrid approach with $n = 2,000$ only spent slightly over 1 min. Increasing $n$ from 2,000 to 5,000 improved the performance of the hybrid approach at the cost of about six additional minutes in runtime. Considering that the runtime of metamodeling for representative data selection and regression-based prediction increased with the number of representative contracts, the difference in runtime between the metamodeling and hybrid frameworks would become even larger for smaller $\alpha(<0.5)$.

## 5. Concluding remarks

In this paper, we proposed a novel hybrid data mining framework to address the practical and computational challenges associated with the valuation of large VA contracts portfolios. In the proposed hybrid framework, the FMVs of VA contracts are calculated by either the Monte Carlo simulation or a random forest model depending on the prediction uncertainty of the contracts. We also consider the expected $R^2$ of individual predictions, which help practitioners to determine the fraction of the portfolio to be assessed by the random forest model. Our numerical study on a portfolio of synthetic VA contracts shows that it is possible to use a statistical learning algorithm to achieve high accuracy and efficiency at the same time while assessing a majority of VA contracts in a portfolio. Although we use random forest for the hybrid approach, other regression methods can be employed, provided that mean square errors can be efficiently estimated.

As with other data mining approaches, the performance of the hybrid approach is generally improved as the random forest model is fed more representative data. While our numerical results show that the proposed approach can be highly effective with 2,000 $\sim$ 5,000 representative contracts, further prediction improvement is expected with a larger set of representative data.

We also examined simple random sampling (rather than conditional Latin hypercube sampling) for the creation of representative labeled data. We found that the difference between conditional Latin hypercube sampling and simple random sampling in the hybrid approach is negligible, indicating the robustness of the proposed approach to the choice of representative data selection method.

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press

*Table 3. Model performance of each approach at different values of α. Runtime represents minutes required for each approach to obtain a set of representative contracts (cLHS), train the RF model, and complete the RF-based prediction of the portfolio. Since both the metamodeling and hybrid approaches use the MC simulation for the same amount of data, the runtime required for the MC simulation is the same and thus omitted.*

| | $\alpha = 0.5$ | | | | | $\alpha = 0.7$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | metamodeling | | | hybrid | | metamodeling | | | hybrid | |
| | RF | GB2 | GLASSO | $n = 2,000$ | $n = 5,000$ | RF | GB2 | GLASSO | $n = 2,000$ | $n = 5,000$ |
| $R^2$ | 0.992 | 0.945 | 0.989 | 0.998 | 0.999 | 0.988 | 0.921 | 0.984 | 0.993 | 0.995 |
| MAE | 5760.937 | 15630.791 | 7998.983 | 3715.437 | 3036.058 | 8685.140 | 18928.962 | 11187.048 | 7626.603 | 6385.638 |
| PE | −0.005 | −0.007 | −0.001 | −0.005 | −0.003 | −0.009 | −0.017 | −0.002 | −0.009 | −0.006 |
| ***Runtime*** | | | | | | | | | | |
| cLHS | 37.042 | 37.042 | 37.042 | 0.382 | 1.707 | 14.898 | 14.898 | 14.898 | 0.382 | 1.707 |
| Training | 588.512 | 57.952 | 27.392 | 0.758 | 4.320 | 213.864 | 35.866 | 16.450 | 0.761 | 4.379 |
| Total | 625.554 | 94.994 | 64.434 | 1.140 | 6.027 | 228.762 | 87.806 | 31.348 | 1.143 | 6.086 |

In summary, the proposed procedure is preferable to other existing data mining frameworks when a highly accurate valuation (e.g., $R^2$ of over 0.99) is required in a timely manner. This innovative hybrid framework shows great potential to help practitioners in insurance industry for effective valuation and risk management.

**Acknowledgments.** This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC, grant numbers 06219 (PI: Shu Li) and 04698 (PI: Hyukjun Gweon)). Support from a start-up grant at Western University is also gratefully acknowledged by Hyukjun Gweon.

## References

Barton, R.R. (2015) Tutorial: Simulation metamodeling. In *2015 Winter Simulation Conference (WSC)*, pp. 1765–1779.

Breiman, L. (1984) *Classification and Regression Trees*. Taylor & Francis, LLC: Boca Raton, FL.

Breiman, L. (1999) Using adaptive bagging to debias regressions. Technical report, University of California at Berkeley. Technical Report.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Cohn, D., Atlas, L. and Ladner, R. (1994) Improving generalization with active learning. *Machine Learning*, **15**(2), 201–221.

Efron, B. (1992) Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society. Series B*, **54**(1), 83–127.

Feng, R., Gan, G. and Zhang, N. (2022) Variable annuity pricing, valuation, and risk management: a survey. *Scandinavian Actuarial Journal*, **2022**(10), 867–900.

Gan, G. (2013) Application of data clustering and machine learning in variable annuity valuation. *Insurance: Mathematics and Economics*, **53**(3), 795–801.

Gan, G. (2018) Valuation of large variable annuity portfolios using linear models with interactions. *Risks*, **6**(3).

Gan, G. (2022) Metamodeling for variable annuity valuation: 10 years beyond kriging. In *2022 Winter Simulation Conference (WSC)*, pp. 915–926.

Gan, G. and Huang, J.X. (2017) A data mining framework for valuing large portfolios of variable annuities. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1467–1475.

Gan, G., Quan, Z. and Valdez, E.A. (2018) Machine learning techniques for variable annuity valuation. In *2018 4th International Conference on Big Data and Information Analytics (BigDIA)*, 1–6.

Gan, G. and Valdez, E.A. (2016) An empirical comparison of some experimental designs for the valuation of large variable annuity portfolios. *Dependence Modeling*, **4**(1), 382–400.

Gan, G. and Valdez, E.A. (2017) Valuation of large variable annuity portfolios: Monte carlo simulation and synthetic datasets. *Dependence Modeling*, **5**(1), 354–374.

Gan, G. and Valdez, E.A. (2018) Regression modeling for the valuation of large variable annuity portfolios. *North American Actuarial Journal*, **22**(1), 40–54.

Gan, G. and Valdez, E.A. (2019) Data clustering with actuarial applications. *North American Actuarial Journal*, **24**(2), 168–186.

Goffard, P. and Guerrault, X. (2015) Is it optimal to group policyholders by age, gender, and seniority for BEL computations based on model points? *European Actuarial Journal*, **5**, 165–180.

Gweon, H. and Li, S. (2021) Batch mode active learning framework and its application on valuing large variable annuity portfolios. *Insurance: Mathematics and Economics*, **99**, 105–115.

Gweon, H., Li, S. and Mamon, R. (2020). An effective bias-corrected bagging method for the valuation of large variable annuity portfolios. *ASTIN Bulletin*, **50**(3), 853–871.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M. and Steiner, S. (2017) Three methods for occupation coding based on statistical learning. *Journal of Official Statistics*, **33**(1), 101–122.

Gweon, H., Schonlau, M. and Wenemark, M. (2020) Semi-automated classification for multi-label open-ended questions. *Survey Methodology*, **46**(2), 265–282.

Hardy, M. (2003) *Investment Guarantees: Modelling and Risk Management for Equity-Linked Life Insurance*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Lin, Y. and Jeon, Y. (2006) Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, **101**(474), 578–590.

Mentch, L. and Hooker, G. (2016) Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, **17**, 1–41.

Minasny, B. and McBratney, A.B. (2006) A conditioned latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, **32**(9), 1378–1388.

Quan, Z., Gan, G. and Valdez, E. (2021) Tree-based models for variable annuity valuation: parameter tuning and empirical analysis. *Annals of Actuarial Science*, pp. 1–24.

Roudier, P. (2011) CLHS: A R package for conditioned latin hypercube sampling.

Schonlau, M. and Couper, M.P. (2016) Semi-automated categorization of open-ended questions. *Survey Research Methods*, **10**(2), 143–152.

Settles, B. (2010) Active learning literature survey. Technical report.

Sexton, J. and Laake, P. (2009) Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis*, **53**(3), 801–811.

Wager, S., Hastie, T. and Efron, B. (2014) Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, **15**, 1625–1651.

Xu, W., Chen, Y., Coleman, C. and Coleman, T.F. (2018) Moment matching machine learning methods for risk management of large variable annuity portfolios. *Journal of Economic Dynamics and Control*, **87**, 1–20.

Zhang, G. and Lu, Y. (2012) Bias-corrected random forests in regression. *Journal of Applied Statistics*, **39**(1), 151–160.

https://doi.org/10.1017/asb.2023.26 Published online by Cambridge University Press