Machine Learning in Long-term Mortality Forecasting

Yang Qiao* Chou-Wen Wang† Wenjun Zhu‡

Abstract

We propose a new machine learning-based framework for long-term mortality forecasting. Based on ideas of neighbouring prediction, model ensembling, and tree boosting, this framework can significantly improve the prediction accuracy of long-term mortality. In addition, the proposed framework addresses the challenge of a shrinking pattern in long-term forecasting with information from neighbouring ages and cohorts. An extensive empirical analysis is conducted using various countries and regions in the Human Mortality Database. Results show that this framework reduces the mean absolute percentage error (MAPE) of the 20-year forecasting by almost 50% compared to classic stochastic mortality models, and it also outperforms deep learning-based benchmarks. Moreover, including mortality data from multiple populations can further enhance the long-term prediction performance of this framework.

Keywords: long-term mortality forecasting; neighbourhood effect; machine learning; ensemble models; tree boosting; longevity risk

*Department of Finance, National Sun Yat-Sen University, Taiwan.
†Department of Finance, National Sun Yat-Sen University, Taiwan.
‡Corresponding author. E-mail: wjzhu@ntu.edu.sg. Division of Banking & Finance, Nanyang Business School, Nanyang Technological University, Singapore.

1 Introduction

An accurate mortality model that is reliable for long-term forecasting is essential for seeking effective solutions to manage mortality risk and longevity risk for different stakeholders (Cairns et al., 2009; Gaille and Sherris, 2011). For example, with long-dated life-related products, life insurance companies use mortality forecasts as a fundamental factor in determining pricing, solvency reserving and other informed decisions about risk management and financial planning. Governments also rely on long-term mortality forecasts to plan for resources allocation for healthcare, pension plans, and social security. Moreover, mortality rates of multiple populations also serve as key inputs of the payoff functions of many innovative mortality- and longevity-linked securities (Blake, 1999; Blake and Burrows, 2001; Lin and Cox, 2005; Cox et al., 2006; Zhou and Li, 2020; Li et al., 2021b). Therefore, an accurate mortality predictive model is critical to facilitate the development of the Life Market, i.e., the traded market in mortality/longevity-linked assets and liabilities (Blake et al., 2013; Blake and Cairns, 2020; Zhou and Li, 2021). The objective of this paper is to propose a machine learning-based long-term mortality modelling framework which effectively improves the accuracy of mortality forecasting, particularly with a long prediction window.

Long-term mortality forecasting is a challenging task, for several reasons. First, the mortality is improving around the globe. For example, over the last two centuries, life expectancy at birth has more than doubled worldwide, and improved even more in developed regions (Riley, 2001; Oeppen and Vaupel, 2002). Particularly over the past decades, life expectancy increased from a global average of 52.57 years in 1960 to 73 years in 2019 (World Bank, 2019). The unanticipated increases in life expectancy not only lead to significant financial cost for pension sponsors, but also make it more challenging to predict future mortality in the long run (Gaille and Sherris, 2011). Second, mortality risk is also volatile and dispersed among populations, due to varying socioeconomic factors, e.g., income per capita, poverty rate, crime rates, etc., and environmental factors, e.g., pollution, natural disasters, pandemics, etc. (see, e.g., Garcia-Menendez et al., 2015; Bunker et al., 2016; Beven et al., 2003; Li and Hyndman, 2021; Boing et al., 2020). For example, according to the World Health Organization (WHO), the excess deaths worldwide during 2020–2021 are estimated

to be 14.91 million, and the world life expectancy in 2020 is reduced by one year (World Bank, 2019). Since mortality forecasts are typically based on historical data and statistical models, such volatility makes it more challenging to successfully predict into the future based on historical experiences. Third, mortalities are dependent among different populations and regions. Sophisticated mortality models that incorporate mortality dependence can improve the accuracy of mortality prediction, and also have important implications regarding mortality/longevity risk management and reduce basis risk (Li and Hardy, 2011; Chen et al., 2015; Chan et al., 2016; Li et al., 2017).

Since the seminal paper by Lee and Carter, in which they proposed a linear extrapolation model for stochastic mortality modeling and forecasting, various single mortality models have been developed (e.g., Cairns et al., 2006; Renshaw and Haberman, 2006; Cairns et al., 2009; Haberman and Renshaw, 2012; Chuang and Brockett, 2014; Currie, 2016). In addition, there is an emerging volume of literature discussing multipopulation mortality modeling, which includes research that propose multipopulation stochastic mortality model (e.g., Li and Lee, 2005; Li and Hardy, 2011; Dowd et al., 2011; Cairns et al., 2011; Jarner and Kryger, 2011; Zhou et al., 2013, 2014; Li et al., 2017; Li and Liu, 2020; Li et al., 2021a), and studies that develop models to capture mortality dependence structure (e.g., Chen et al., 2015; Wang et al., 2015; Zhu et al., 2017; Zhou, 2019; Zhou and Ji, 2021).[1] These classic mortality models can perform long-term forecasting conveniently. However, the linear extrapolation structure may restrict the forecasting performance.

In recent years, machine learning techniques have gained increasing attention in different fields including actuarial science (LeCun et al., 2015; Richman, 2021a,b). Many machine learning models, including deep learning, tree-based models, etc., can discover meaningful, nonlinear, and complicated patterns within data, and hence, are promising alternatives for mortality prediction. There is an increasing volume of pioneering work to apply machine learning techniques in both non-life sector (e.g., Wüthrich, 2017, 2018; Gao et al., 2019; Kuo, 2019; Noll et al., 2020; Gabrielli et al., 2020; Wuthrich and Buser, 2021; Gao et al., 2022; Chen et al., 2023), and life sector, especially mortality forecasting (e.g., Deprez et al.,

[1]For a comprehensive review of the recent development of mortality modelling and forecasting, see Blake and Cairns (2021) and the references therein.

2017; Richman and Wüthrich, 2019; Hainaut, 2018; Dong et al., 2020; Perla et al., 2021; Wang et al., 2021). We provide a comprehensive literature review on mortality modeling with machine learning in Section 2.

In this paper, we propose a novel framework for long-term mortality forecasting. First, we construct neighbourhood mortality data, which is able to capture the neighbourhood effect that is not identifiable by classic mortality models, and hence, improves the prediction accuracy (Wang et al., 2021). Next, we apply various tree boosting models to construct base prediction models for mortality, including XGBoost (Chen and Guestrin, 2016), CatBoost (Ke et al., 2017), and LightGBM (Prokhorenkova et al., 2018). Finally, we employ model ensembling idea to address the challenge of a shrinking pattern when performing longer-term mortality using information from neighbouring ages and cohorts. The ensembling process also reduces model variance and generalization errors and enjoys higher stability. We perform an extensive empirical analysis on various populations in the Human Mortality Database (HMD). Results show that the prediction error of the proposed framework is only half of the classic stochastic mortality models. More importantly, the proposed model also outperforms deep learning-based benchmarks.

Our empirical results also confirm that among all tested tree boosting algorithms, XGBoost has the best performance for long-term mortality forecasting. The XGBoost algorithm is based on adding a regularization term to the loss function to control the complexity of the model. It also expands a drift function through a Taylor second-order expansion to achieve better prediction, which makes the algorithm stand out from other machine learning algorithms. For example, in the professional Kaggle machine learning competition, out of 29 Kaggle winning teams in 2015, 17 teams used the XGboost algorithm compared to 11 teams using deep learning. Due to its excellent performance in prediction, the XGBoost algorithm has been widely used in finance-related topics such as stock price prediction (Basak et al., 2019), crude oil price prediction (Nobre and Neves, 2019), and corporate credit rating prediction (Son et al., 2019). However, to the best of our knowledge, there is no application of the XGBoost algorithm to mortality modelling in the literature.

The rest of the paper proceeds as follows. Section 3 proposes the mortality model ensembling framework to address some challenges in long-term mortality forecasting. Section 4

introduces the tree boosting algorithms that are used in the empirical analysis. Section 5 presents the empirical test results. Section 6 concludes the paper.

2 Literature review and contribution

There has been a substantial and growing body of literature exploring the utilization of various machine/deep learning algorithms across various fields of actuarial science, including asset allocation and investment decisions (e.g., Li and Forsyth, 2019; Carbonneau, 2021; Jin et al., 2021), catastrophe bond pricing (e.g., Makariou et al., 2021), nonlife insurance modeling and pricing (e.g., Wüthrich, 2017; Gabrielli and V. Wüthrich, 2018; Lally and Hartman, 2018; Lee and Lin, 2018; Diao and Weng, 2019; Gao et al., 2019; Noll et al., 2020; Devriendt et al., 2021; Henckaerts et al., 2021; Lee, 2021; Wuthrich and Buser, 2021; Gao et al., 2022; Hu et al., 2022; Meng et al., 2022; Xin and Huang, 2023; Fissler et al., 2023), insurance reserving (e.g., Wüthrich, 2018; Kuo, 2019; Lopez et al., 2019; Gabrielli et al., 2020; Al-Mudafer et al., 2022), climate and weather risk management (e.g., Brock Porth et al., 2020; Ghahari et al., 2019; Chen et al., 2023), lapse risk management (e.g., Loisel et al., 2021), fraud detection (e.g., Gomes et al., 2021; Debener et al., 2023), cyber insurance (e.g., Farkas et al., 2021), regulation (e.g., Castellani et al., 2018), as well as mortality modeling. For a in-depth review of the recent advancements of machine learning in actuarial science, refer to Richman (2021a) and Richman (2021b), as well as the references therein. In this section, we will focus on mortality modeling and forecasting, and provide a comprehensive overview of the burgeoning literature in machine learning with its application in this specific area.

Pioneering the field, Deprez et al. (2017) discuss the application of machine learning techniques in the analysis and evaluation of stochastic mortality models. In particular, they found that *boosting regression trees* can detect weaknesses in traditional mortality models, such as the Lee-Carter model (Lee and Carter, 1992) and Renshaw-Haberman (Renshaw and Haberman, 2006). This study underscored the invaluable role of machine learning in augmenting our comprehension and assessment of mortality models. Expanding upon this pioneering research, Levantesi and Pizzorusso (2019) and Levantesi and Nigri (2020) enhance

the prediction accuracy of standard stochastic mortality models by incorporating different tree-based models. Meanwhile, Bjerre (2022) propose to use tree-based models to predict differenced log-mortality rates, and investigate the interpretability of these constructed machine learning models.

In a distinct avenue of exploration, Hainaut (Hainaut, 2018) presented the first paper that proposes to model mortality with an autoencoder. The proposed framework extends traditional Lee-Carter model to capture the nonlinearies in the mortality data, enhancing the estimation of latent mortality factors. In a related study, Miyata and Matsuyama (2022) introduce a Bayesian extension of Lee-Carter model, incorporating a variational autoencoder.

Richman and Wüthrich (2019) propose a neural-network extension of the Lee-Carter model for multiple populations. The innovative framework allows neural networks to automatically select an optimal structure to improve mortality forecasting. Since then, numerous studies have investigated to improve the mortality prediction performance through the utilization of various neural network architectures, including recurrent neural network (RNN) (Nigri et al., 2019; Petneházi and Gáll, 2019; Richman and Wuthrich, 2019; Bravo, 2021a,b; Nigri et al., 2021; Perla et al., 2021; Chen and Khaliq, 2022; Lindholm and Palmborg, 2022; Zhang et al., 2022; Marino et al., 2023), and Convolution neural network (Perla et al., 2021; Wang et al., 2021; Schnürch and Korn, 2022; Zhang et al., 2022). For example, Nigri et al. (2019) and Lindholm and Palmborg (2022) improve the Lee-Carter model by fitting a Long Short-Term Memory (LSTM) model to the latent factor, κ_t. Nigri et al. (2021) proposes a LSTM network-based model to improve the forecast accuracy of life expectancy and lifespan disparity. Perla et al. (2021) discuss both RNN and CNN architectures in mortality modeling. Wang et al. (2021) adopt a unique approach by interpreting mortality data as images and constructing CNN models for mortality prediction.

Some studies employ tensor-based approach for mortality modeling. For example, Dong et al. (2020) use tensor decomposition to construct multi-population mortality models. Cardillo et al. (2022) apply a tensor-based approach to study the cause-of-death mortality patterns. Most recently, transformer model has been proposed to predict mortality data in major countries (Wang et al., 2023).

Furthermore, researchers have also investigated the uncertainty of the constructed model.

Schnürch and Korn (2022) propose a CNN architecture for the Lee-Carter model and study the prediction uncertainty using bootstrapping. Marino et al. (2023) formalize the uncertainty estimation of LSTM-based Lee-Carter model by proposing prediction intervals of the mortality index. Richman (2022) provides a comprehensive review of deep learning models for mortality forecasting, and in particular, discusses model uncertainty estimations using quantile regression and deep ensembles.

Contribution of this paper This paper makes important contributions to the literature, in three ways. First, we enhance the long-term mortality forecasting accuracy through a novel framework. By leveraging ideas from neighboring prediction, model ensembling, and tree boosting, the framework proposed in this paper significantly improves both the accuracy and stability for long-term mortality prediction. This is crucial for various financial applications including life insurance and pension fund risk management, where accurate long-term mortality forecasts are essential for risk management and investment decisions. Second, for the first time, this paper introduces the application of XGBoost and other boosting algorithms to enhance mortality forecasting, with a particular emphasis on long-term predictions. The existing literature on machine learning for mortality forecasting has predominantly focused on deep learning models, except the work by Deprez et al. (2017), Levantesi and Pizzorusso (2019) and Levantesi and Nigri (2020). Our research demonstrates that tree-based methods can achieve satisfying performance in mortality forecasting. The prediction accuracy is higher than popular deep learning methods such as CNN. Finally, this paper extends and complements the work of Wang et al. (2021) and other related studies that employ deep learning techniques for mortality prediction. It addresses some unique challenges associated with long-term forecasting using existing models. These challenges are pervasive whenever machine learning techniques are used for mortality prediction from neighboring ages and cohorts. Importantly, while this paper primarily focuses on tree boosting algorithms, the proposed framework is flexible and adaptable to alternative machine learning models. The proposed framework's ability to improve prediction accuracy will directly benefit the financial and insurance industry by enabling better mortality risk modeling, ultimately leading to more stable and efficient operations.

3 Mortality model ensembling

3.1 The neighbourhood mortality data

In this paper, we will construct a mortality prediction framework with neighbourhood data, that is, we train machine learning models, $\mathcal{M}\big(\mathcal{E}_{m_{x,t-1}}(k_1, k_2, s)\big)$, to predict future mortality $m_{x,t}$, where $\mathcal{M} : \mathbb{R}_+^{(k_1+k_2+1)\times s} \to \mathbb{R}_+$, is a function to map neighbourhood mortality data, which will be introduced below, to mortality. Such model is able to capture the intricate and highly nonlinear structure in the mortality data, pushing the boundary beyond the linear decomposition. In the past decades, a number of stochastic mortality models have been proposed. These include Lee-Carter model (Lee and Carter, 1992) and its various extensions (see, e.g., Renshaw and Haberman, 2003; Cairns et al., 2006; Renshaw and Haberman, 2006; Cairns et al., 2009). Most of the existing stochastic mortality models assume that the mortality rates can be decomposed into a few additive latent factors, and each factor can be further decomposed into age, period, and cohort effects. More specifically,

$$g(m_{x,t}) \quad = \quad \sum_{i}^{I} \beta_x^{(i)} \cdot \kappa_t^{(i)} \cdot \gamma_{t-x}^{(i)}, \tag{1}$$

where $m_{x,t}$ is the central death rate at age x and year t, $x_0 \le x \le X$, $0 \le t \le T$; $g(\cdot)$ is a transformation function such as logarithmic function or logit transformation; $\beta_x^{(i)}$ represents age-related effect; $\kappa_t^{(i)}$ represents period-related effect; and $\gamma_{t-x}^{(i)}$ represents cohort-related effect, $i = 1, 2, \ldots, I$.

While the linear decomposition framework is parsimonious and achieves great success in mortality prediction, this framework is still restrictive as it only captures linear relationship of different latent factors and only focuses on effects along three dimensions: age, period, and cohort. Wang et al. (2021) propose a neighbouring prediction model which pushes the boundary further by predicting $m_{x,t}$ with a neighbourhood around $m_{x,t}$. The neighbouring prediction framework is shown to capture the neighbourhood effect in mortality data and achieves superior prediction performance. The key is to construct the neighbourhood data, $\mathcal{E}_{m_{x,t}}(k_1, k_2, s)$. For each mortality rate at age x and year t, $\mathcal{E}_{m_{x,t}}(k_1, k_2, s)$ is a two-dimensional neighbourhood around $m_{x,t}$, including mortality information for ages

$x-k_1,\ldots,x,\ldots,x+k_2$, lagging s-years. More specifically,

$$\mathcal{E}_{m_{x,t}}(k_1,k_2,s)=\begin{bmatrix} m_{x-k_1,t-s}, & \ldots & m_{x-k_1,t-2}, & m_{x-k_1,t-1} \\ \vdots & \ldots & \vdots & \vdots \\ m_{x-1,t-s}, & \ldots & m_{x-1,t-2}, & m_{x-1,t-1} \\ m_{x,t-s}, & \ldots & m_{x,t-2}, & m_{x,t-1} \\ m_{x+1,t-s}, & \ldots & m_{x+1,t-2}, & m_{x+1,t-1} \\ \vdots & \ldots & \vdots & \vdots \\ m_{x+k_2,t-s}, & \ldots & m_{x+k_2,t-2}, & m_{x+k_2,t-1} \end{bmatrix}. \tag{2}$$

The neighbourhood data $\mathcal{E}_{m_{x,t}}(k_1,k_2,s)$ can be viewed as a 2D image, where not only mortality rate of age x is included in the model, the population with very close ages in a neighbourhood of $[x-k_1,x+k_2]$ are also used to construct the prediction model.

3.2 Long-term forecast using neighbourhood data: Challenges

It is very natural to perform one-year-ahead forecasts using the neighbourhood data constructed in Subsection 3.1, since $\mathcal{E}_{m_{x,t}}(k_1,k_2,s)$ is $\mathcal{F}_{t-1}$ measurable. However, performing long-term forecasts (i.e., multi-year forecasts) are not simple extensions of a single-year forecast, which has a shrinking pattern of the prediction results.

To illustrate, let us first consider a naïve recursive procedure, that is, we use historical and predicted values as inputs to forecast future years. Assume that the historical sample period is from 0 to T, and the forecast horizon is τ years. Also assume that $x_0\le x\le X$, i.e., the age range is from x_0 to X. Then, the forecast of mortality rate of age x in year $T+1$, $\widehat{m}_{x,T+1}$, uses $\mathcal{E}_{m_{x,T}}(k_1,k_2,s)$ as inputs, $\widehat{m}_{x,T+2}$ uses $\mathcal{E}_{\widehat{m}_{x,T+1}}(k_1,k_2,s)$ as inputs, and so on. More specifically, the recursive procedure is as follows:

$$\begin{aligned} \widehat{m}_{x,T+1} &= \mathcal{M}\big(\mathcal{E}_{m_{x,T}}(k_1,k_2,s)\big), \\ \widehat{m}_{x,T+2} &= \mathcal{M}\big(\mathcal{E}_{\widehat{m}_{x,T+1}}(k_1,k_2,s)\big), \\ &\ldots \\ \widehat{m}_{x,T+\tau} &= \mathcal{M}\big(\mathcal{E}_{\widehat{m}_{x,T+\tau-1}}(k_1,k_2,s)\big). \end{aligned} \tag{3}$$

The above recursive procedure works for time series inputs, but is problematic for neighbourhood data (i.e., $k_1 > 0$ or $k_2 > 0$). For example, in year $T + 1$, the predictions will be missing if $x < x_0 + k_1$ or $x > X - k_2$. Similarly, in year $T + 2$, when $x < x_0 + 2k_1$ or $x > X - 2k_2$, predictions are unavailable due to missing inputs. As such, naïve recursive prediction procedure will produce a shrinking pattern, which is illustrated in Figure 1.

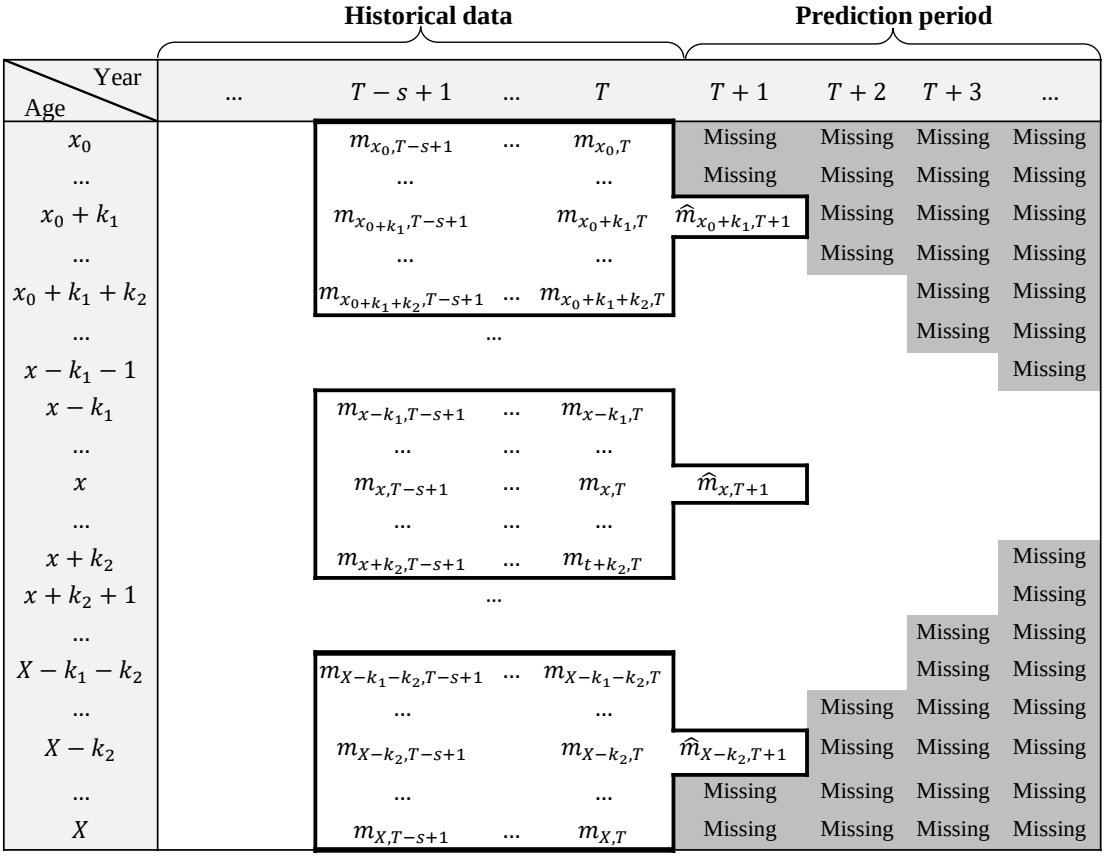


Figure 1. **Shrinking pattern of long-term prediction using neighbourhood data.**

3.3 Model ensembling

In this paper, we propose an ensemble modelling framework to address the aforementioned challenge. In machine learning, ensemble modelling is a commonly used technique to reduce model variance or bias (Dietterich, 2000). The idea is to train multiple different base models, and aggregate the prediction of each base model as the final model. While the motivation for using ensemble models in this paper is to overcome the challenge of a shrink-

ing prediction horizon in long-term mortality forecasting, our resulting long-term prediction models also enjoy all existing benefits of model ensembling, including reduced generalization errors, higher stability, lower model variance, etc.

The ensemble model proposed in this paper utilizes different neighbourhood mortality data constructions. Its general representation can be expressed as follows:

$$\widehat{\mathcal{M}}_{x,t} = \begin{cases} \dfrac{1}{x-x_0+1}\displaystyle\sum_{i=0}^{x-x_0} \widehat{\mathcal{M}}_{a,i+1}\big(\mathcal{E}_{m_{x,t}}(i,k_1+k_2-i,s)\big) & \text{if } x < x_0+k_1 \\ \dfrac{1}{k_1+k_2+1}\Bigg[\displaystyle\sum_{i=0}^{k_1-1} \widehat{\mathcal{M}}_{a,i+1}\big(\mathcal{E}_{m_{x,t}}(i,k_1+k_2-i,s)\big) & \\ \quad +\widehat{\mathcal{M}}_0\big(\mathcal{E}_{m_{x,t}}(k_1,k_2,s)\big) & \text{if } x_0+k_1 \le x \le X-k_2 \\ \quad +\displaystyle\sum_{j=0}^{k_2-1} \widehat{\mathcal{M}}_{b,j+1}\big(\mathcal{E}_{m_{x,t}}(k_1+k_2-j,j,s)\big)\Bigg] & \\ \dfrac{1}{X-x+1}\displaystyle\sum_{j=0}^{X-x} \widehat{\mathcal{M}}_{b,j+1}\big(\mathcal{E}_{m_{x,t}}(k_1+k_2-j,j,s)\big) & \text{if } x > X-k_2. \end{cases} \quad (4)$$

where $\widehat{\mathcal{M}}_{x,t}$ is the ensemble model for mortality rate of age x in year t. For example, if $x = x_0$, we construct model $\widehat{\mathcal{M}}_{a,1}\big(\mathcal{E}_{m_{x,t}}(0,k_1+k_2,s)\big)$, which is the model based on neighbourhood data $\mathcal{E}_{m_{x,t}}(0,k_1+k_2,s)$ to predict $m_{x,t}$; if $x = x_0+1$, we construct two models, model $\widehat{\mathcal{M}}_{a,1}\big(\mathcal{E}_{m_{x,t}}(0,k_1+k_2,s)\big)$ and model $\widehat{\mathcal{M}}_{a,2}\big(\mathcal{E}_{m_{x,t}}(1,k_1+k_2-1,s)\big)$, which are models based on neighbourhood data $\mathcal{E}_{m_{x,t}}(0,k_1+k_2,s)$ and $\mathcal{E}_{m_{x,t}}(1,k_1+k_2-1,s)$, respectively, and use the average of these two models to predict $m_{x,t}$; and so on. Similarly, if $x = X$, we will only use model $\widehat{\mathcal{M}}_{b,X}\big(\mathcal{E}_{m_{x,t}}(k_1+k_2,0,s)\big)$ to predict $m_{x,t}$; and if $x = X-1$, we will use the average of model $\widehat{\mathcal{M}}_{b,1}\big(\mathcal{E}_{m_{x,t}}(k_1+k_2,0,s)\big)$ and model $\widehat{\mathcal{M}}_{b,2}\big(\mathcal{E}_{m_{x,t}}(k_1+k_2-1,1,s)\big)$ to predict $m_{x,t}$. When $x_0+k_1 \le x \le X-k_2$, all k_1+k_2+1 models will be included in the ensembling.

To illustrate, let us consider a concrete example. Suppose $x_0 = 65$ and $X = 89$, and consider a hypothetical mortality table (e.g., Figure 3). We are interested to perform a long-term forecast from 2000 - 2020, with neighbourhood data with $k_1 = k_2 = 6$. Using the model ensembling idea discussed above, we will construct and ensemble in total 13 models. The model construction and ensembling procedures following Equation (4) are shown in Figure 2. In particular, $\widehat{\mathcal{M}}_{a,1}$ uses $\mathcal{E}_{m_{x,t}}(0,12,s)$ as inputs, that is, neighbourhood data of ages from

x to $x+12$ of the past s years. Then we predict mortality of age 65 with only $\widehat{\mathcal{M}}_{a,1}$, i.e., $\widehat{\mathcal{M}}_{65} = \widehat{\mathcal{M}}_{a,1} = 0.12$. $\widehat{\mathcal{M}}_{a,2}$ uses $\mathcal{E}_{m_{x,t}}(1, 11, s)$, i.e., neighbourhood data of ages from $x-1$ to $x+11$ of the past s years, as inputs. We predict mortality of age 66 as the average of $\widehat{\mathcal{M}}_{a,1}$ and $\widehat{\mathcal{M}}_{a,2}$, i.e., $\widehat{\mathcal{M}}_{66} = \frac{1}{2}\left(\widehat{\mathcal{M}}_{a,1} + \widehat{\mathcal{M}}_{a,2}\right) = 0.14$. This asymmetric neighbourhood model construction goes on until the seventh model, $\widehat{\mathcal{M}}_0$, which uses a symmetric neighbourhood data of ages from $x-6$ to $x+6$. Age 77 mortality will be predicted by averaging predictions of all 13 models, i.e., $\widehat{\mathcal{M}}_{77} = \frac{1}{13}\left(\widehat{\mathcal{M}}_{a,1} + \widehat{\mathcal{M}}_{a,2} \ldots + \widehat{\mathcal{M}}_0 + \ldots \widehat{\mathcal{M}}_{b,2} + \widehat{\mathcal{M}}_{b,1}\right) = 0.34$. The next six models are again based on asymmetric neighbourhood data but utilize more younger ages. For example, for $\widehat{\mathcal{M}}_{b,6}$ uses $\mathcal{E}_{m_{x,t}}(12, 0, s)$, i.e., neighbourhood data of ages from $x-12$ to x. Prediction of age 89 will be based on $\widehat{\mathcal{M}}_{b,6}$ only. Figure 3 shows historical sample period and prediction period in the example. We can see that after model ensembling, the shrinking prediction pattern issue is addressed.

Age \ Model	$\widehat{\mathcal{M}}_{a,1}$	$\widehat{\mathcal{M}}_{a,2}$	$\widehat{\mathcal{M}}_{a,3}$	$\widehat{\mathcal{M}}_{a,4}$	$\widehat{\mathcal{M}}_{a,5}$	$\widehat{\mathcal{M}}_{a,6}$	$\widehat{\mathcal{M}}_{0}$	$\widehat{\mathcal{M}}_{b,1}$	$\widehat{\mathcal{M}}_{b,2}$	$\widehat{\mathcal{M}}_{b,3}$	$\widehat{\mathcal{M}}_{b,4}$	$\widehat{\mathcal{M}}_{b,5}$	$\widehat{\mathcal{M}}_{b,6}$	$\widehat{\boldsymbol{\mathcal{M}}}$
65	0.12													**0.12**
66	0.13	0.15												**0.14**
67	0.12	0.13	0.14											**0.13**
68	0.16	0.18	0.15	0.14										**0.16**
69	0.21	0.19	0.18	0.17	0.19									**0.19**
70	0.21	0.18	0.19	0.21	0.21	0.19								**0.20**
71	0.21	0.22	0.23	0.18	0.19	0.22	0.21							**0.21**
72	0.23	0.21	0.25	0.24	0.23	0.22	0.21	0.23						**0.23**
73	0.21	0.23	0.25	0.24	0.26	0.21	0.19	0.27	0.26					**0.24**
74	0.31	0.27	0.28	0.26	0.28	0.19	0.18	0.22	0.24	0.25				**0.25**
75	0.32	0.28	0.29	0.26	0.24	0.25	0.33	0.26	0.23	0.22	0.28			**0.27**
76	0.31	0.28	0.29	0.33	0.27	0.33	0.32	0.31	0.29	0.31	0.29	0.36		**0.31**
77	0.36	0.32	0.34	0.35	0.36	0.38	0.29	0.32	0.31	0.33	0.35	0.34	0.36	**0.34**
78		0.41	0.38	0.39	0.36	0.37	0.38	0.36	0.35	0.32	0.33	0.38	0.31	**0.36**
79			0.44	0.38	0.42	0.39	0.36	0.38	0.39	0.42	0.44	0.35	0.41	**0.40**
80				0.44	0.46	0.44	0.41	0.43	0.46	0.42	0.48	0.49	0.45	**0.45**
81					0.49	0.52	0.56	0.48	0.55	0.49	0.56	0.52	0.51	**0.52**
82						0.58	0.56	0.62	0.56	0.66	0.52	0.54	0.61	**0.58**
83							0.66	0.59	0.62	0.58	0.63	0.65	0.64	**0.62**
84								0.74	0.75	0.71	0.69	0.73	0.71	**0.72**
85									0.75	0.78	0.82	0.77	0.79	**0.78**
86										0.89	0.88	0.92	0.85	**0.89**
87											0.94	0.92	0.88	**0.91**
88												0.99	0.92	**0.96**
89													0.99	**0.99**

Figure 2. **Model ensembling procedure in the concrete example.**

4 Model boosting

Gradient Boosting Decision Tree (GBDT) developed by Friedman (2001) is a widely-used machine learning technique that shines in applications of many areas. XGBoost is a famous

	Historical data						Prediction period			
Years / Ages	1994	1995	1996	1997	1998	1999	**2000**	**2001**	...	**2020**
63	...	...	...	...	...	...			...	
64	...	...	...	...	...	...			...	
65	0.14	0.14	0.13	0.12	0.12	0.11	0.12	0.12	...	0.11
66	0.16	0.15	0.14	0.14	0.13	0.13	0.13	0.14	...	0.13
67	0.18	0.17	0.16	0.16	0.15	0.14	0.14	0.13	...	0.14
68	0.2	0.19	0.19	0.18	0.16	0.16	0.15	0.16	...	0.15
69	0.22	0.21	0.19	0.19	0.18	0.17	0.17	0.19	...	0.16
70	0.23	0.23	0.23	0.21	0.21	0.2	0.18	0.20	...	0.18
71	0.25	0.24	0.24	0.23	0.22	0.22	0.22	0.21	...	0.18
72	0.28	0.27	0.27	0.26	0.25	0.24	0.23	0.23	...	0.21
73	0.32	0.3	0.29	0.29	0.28	0.27	0.25	0.24	...	0.23
74	0.36	0.32	0.31	0.32	0.31	0.3	0.28	0.25	...	0.26
75	0.37	0.37	0.36	0.34	0.33	0.32	0.31	0.27	...	0.29
76	0.43	0.39	0.4	0.38	0.36	0.36	0.32	0.31	...	0.31
77	0.47	0.44	0.44	0.43	0.41	0.41	0.36	0.34	...	0.35
78	0.46	0.49	0.49	0.46	0.46	0.44	0.38	0.36	...	0.38
79	0.48	0.46	0.44	0.48	0.42	0.49	0.42	0.40	...	0.39
80	0.53	0.51	0.48	0.46	0.45	0.48	0.48	0.45	...	0.46
81	0.51	0.58	0.58	0.56	0.57	0.55	0.56	0.52	...	0.56
82	0.68	0.67	0.66	0.66	0.65	0.67	0.61	0.58	...	0.58
83	0.69	0.68	0.67	0.66	0.63	0.62	0.63	0.62	...	0.66
84	0.85	0.84	0.83	0.84	0.82	0.81	0.78	0.72	...	0.71
85	0.89	0.88	0.87	0.85	0.85	0.83	0.82	0.78	...	0.78
86	0.94	0.94	0.93	0.92	0.91	0.88	0.88	0.89	...	0.88
87	0.98	0.96	0.95	0.95	0.94	0.93	0.92	0.91	...	0.91
88	0.99	0.99	0.99	0.98	0.98	0.96	0.95	0.96	...	0.92
89	0.99	0.98	0.97	0.99	0.98	0.98	0.98	0.99	...	0.99
90	...	...	...	...	...	...			...	
91	...	...	...	...	...	...			...	

Figure 3. **Historical period and prediction period in the example.**

effective implementation of GBDT. In this section, using XGBoost as an example, we will review the general idea of tree boosting, and then we will briefly introduce three commonly used algorithms, including XGBoost, and two of its extensions, LightGBM and CatBoost. In the empirical analysis in Section 5, we will use these three tree boosting models to construct individual base prediction models for mortality.

4.1 Gradient tree boosting and XGBoost

XGboost proposed by Chen and Guestrin (2016) extends the original GBDT algorithm by adding a second-order Taylor expression term and a penalty term to the objective loss function, leading to more efficient and precise predictions. The XGBoost algorithm is a supervised machine learning decision tree model that uses an ensemble learning perspective

to combine many weak models to form a strong model, thereby improving the robustness of the model.

Given a set of N observations, $D = \{(X_i, y_i)\}, i = 1, \ldots, N$, where $X_i \in R^m$ is an input feature of m dimensions; $y_i \in R$ is a one-dimensional output or prediction target and is predicted using addictive functions, which is expressed as follows:

$$\hat{y}_i = \phi(X_i) = \sum_{k=1}^{K} f_k(X_i), \quad f_k \in \Psi, \tag{5}$$

where K is the number of trees and Ψ is the function space of the regression tree. Assuming that the current regression tree node is split by the dth feature value, the samples with feature values less than s are divided into the tree on the left, and those greater than s are divided into the tree on the right, i.e.,

$$R_1(d, s) = \{X | X^{(j)} \leq s\} \text{ and } R_2(d, s) = \{X | X^{(j)} \geq s\}. \tag{6}$$

To avoid the overfitting problem, the objective function of the XGBoost algorithm is the following regularized function:

$$L(\phi) \quad = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k). \tag{7}$$

There are two components in the objective function: l represents the differentiable convex loss function. For example, the most commonly used Square Loss in regression trees is $l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$; $\Omega = \Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||w||^2$ is the regularization function that penalizes the complexity of the model. It also contains two parts. The first part is related to the number of leaf nodes, where T is the number of leaves in the tree and γ is the corresponding regularization parameter; the second part is related to the value of control leaf nodes, where w is leaf score and λ is the regularization parameter. The XGBoost algorithm prevents overfitting problems through the regularization function. When the regularization parameter is zero, the objective function, $L(\phi)$, degenerates to the original GBDT.

The gradient tree boosting procedure in general is to combine multiple tree models, and generate new trees by using the results of previous iterations, with each new tree added to

minimize the objective function. In other words, each time a tree model is generated, the residuals of the previous prediction are fitted. Let $\hat{y}^{(0)} = 0$, then the prediction process can be expressed as follows

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(X_i) = \hat{y}_i^{(t-1)} + f_t(X_i). \tag{8}$$

Therefore, without considering the regularization function, the objective function of the tth iteration $L^{(t)}$ can be written as

$$L^{(t)} = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(X_i)\right). \tag{9}$$

According to the second-order Taylor expansion:

$$L^{(t)} = \sum_{i=1}^{N} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_i(X_i) + \frac{1}{2} h_i f_t^2(X_i) \right] + \text{Constant}, \tag{10}$$

where g_i and h_i are the first and second-order derivatives of the loss function, respectively. Since the constant term does not affect the differentiation result, the target function for the tth iteration can be simplified as

$$\begin{aligned} \bar{L}^{(t)} &= \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)}\right) - \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2 \right] + \gamma T, \\ G_j &= \sum_{i \in I_j} \quad \text{and} \quad H_j = \sum_{i \in I_j} h_i, \end{aligned} \tag{11}$$

where $I_j \in \{i | q(X_i) = j\}$ is the instance set of leaf j, and $q : \mathbb{R}^m \to T$ is the representation of the tree structure. The optimal value of objective function for a fixed structure q can be obtained by talking first-order condition:

$$\bar{L}^{(t)*}(q) = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)}\right) - \frac{1}{2} \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2 \right] + \gamma T. \tag{12}$$

However, it is impossible to enumerate all possible structures. As a result, similar to Fried-

man (2001), XGBoost algorithm also uses a greedy algorithm to perform node splitting. Similar to Equation (6), the gain before and after node splitting can be defined as

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} + \frac{(G_R+G_L)^2}{H_L+H_R+\lambda}\right] - \gamma. \tag{13}$$

Equation (13) consists of the fraction of new nodes on the left (subscripted as L), the fraction of new nodes on the right (subscripted as R), the fraction of original nodes, and the regularization parameter of the number of control nodes γ. Therefore, Equation (13) can be used to evaluate the split candidates.

4.2 LightGBM

LightGBM was developed by Microsoft in 2016 (Ke et al., 2017). It accelerates the training process by reducing the sample size and the number of features. Similar to XGBoost, it is an improvement and engineering implementation of the GBDT algorithm. In order to solve the problem of low efficiency of the GBDT algorithm when there is a large amount of data, LightGBM uses a novel histogram-based algorithm, which divides continuous features into discrete buckets, allowing the model to process the high-dimensional data more efficiently. In particular, LightGBM contains two techniques: Gradient-based One-Side Sampling and Exclusive Feature Bundling to address the issues of a large sample size and a large number of features. Compared with the pre-sorted algorithm, the histogram-based optimization algorithm reduces the calculation amount and memory usage of the partition feature, and facilitates parallel calculation. When generating a decision tree, LightGBM uses a leaf-wise method with a depth limit, which potentially could outperform the depth-wise method in XGBoost. In addition to its efficiency, LightGBM can directly process categorical features without converting to one-hot encoding. It can also support GPU training and handle large-scale data, and hence, has been widely used in industry.

4.3 CatBoost

CatBoost, developed by Prokhorenkova et al. (2018), is another popular gradient boosting toolkit developed by the search giant Yandex in 2017. CatBoost, for Categorical and

Boosting, proposes ordering principle to solve the problem of prediction shift in existing implementations of gradient boosting and is more efficient to process categorical features. Compared with XGBoost and Lightgbm, CatBoost embeds an innovative ordered boosting algorithm that automatically processes categorical features into numerical features, enriches the feature dimension, and handles the noisy points in the training set. With a fully symmetric tree as the base model, CatBoost potentially has a prediction accuracy, especially on data sets containing categorical features.

5 Empirical analysis

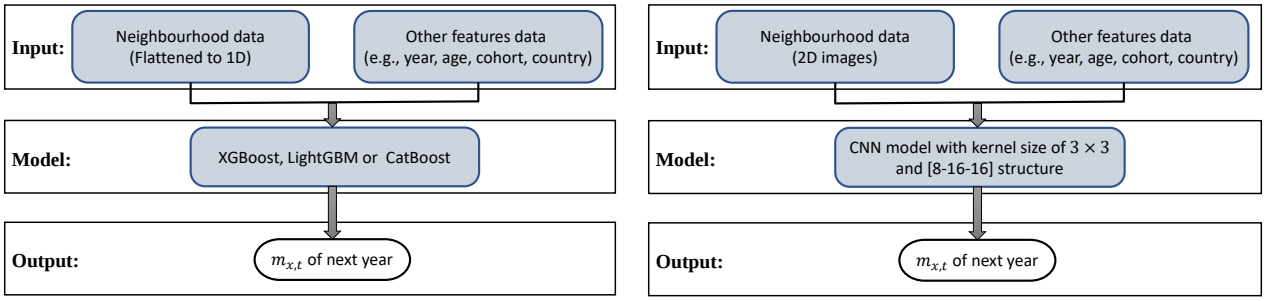
5.1 Data and benchmark models

In this section, we perform extensive empirical analyses to test the performance of the proposed long-term prediction framework. We use various populations in the Human Mortality Database (HMD; http://www.mortality.org). We include all 32 populations that have data available since 1960 in our analysis. Table 1 summarizes the populations that are used in the empirical analysis, together with their abbreviations. We use total population mortality data between age 65 and 89. Online Appendix A provides more detailed descriptions of the data.

Table 1. **Summary of Countries and Regions used in the empirical analysis.**

Country	Abbrev	Country	Abbrev	Country	Abbrev	Country	Abbrev
Australia	AUS	Denmark	DNK	Iceland	ISL	New Zealand	NZL
Austria	AUT	Spain	ESP	Italy	ITA	Poland	POL
Belgium	BEL	Estonia	EST	Japan	JPN	Portugal	PRT
Bulgaria	BGR	Finland	FIN	Lithuania	LTU	Russia	RUS
Belarus	BLR	France	FRA	Luxembourg	LUX	Slovakia	SVK
Canada	CAN	U.K.	GBR	Latvia	LVA	Sweden	SWE
Switzerland	CHE	Hungary	HUN	Netherlands	NLD	Ukraine	UKR
Czechia	CZE	Ireland	IRL	Norway	NOR	U.S.A.	USA

Following the machine learning convention, we split the sample into two subsamples: training set and test set, preserving the temporal order. In particular, all data earlier than 1999 are used as the training set while data from 2000 - 2020 are used as the test set. Note



(a) Tree model (b) CNN model

Figure 4. **Structure of individual base models for ensembling.**

that the year 2020, the first year of pandemic, is included in the test set, and it is interesting to compare the prediction performance of various models in this year.

We evaluate and compare the prediction power of different models based on mean absolute percentage error (MAPE) defined as follows:

$$MAPE = \frac{100\%}{n_t} \sum_t \sum_x \left| \frac{m_{x,t} - \hat{m}_{x,t}}{m_{x,t}} \right|, \tag{14}$$

where $\hat{m}_{x,t}$ is the predicted value of $m_{x,t}$ based on different predictive models; n_t is the number of observations in the test set. MAPE is a commonly used measure to evaluate forecasting for mortality.

Following Wang et al. (2021), in the empirical analysis of this paper, we construct neighbourhood mortality data with window width $k_1 = k_2 = s = 6$. For each individual base model, $\widehat{\mathcal{M}}_i$, we construct a boosted tree model with XGBoost, LightGBM, or CatBoost algorithms, where the number of iteration for boosting is set to be 3000, the maximum depth of a tree is 6, and the step size for shrinkage is set as 0.01. We have tested different choices of these parameters and found the results are robust. Figure 4(a) displays the structure of each tree-based model.

Benchmark mortality models A natural benchmark model for our proposed framework in this paper is the CNN model in Wang et al. (2021). Therefore, we consider a CNN-Hybrid model with the same structure of the best model in Wang et al. (2021). Note that

for this CNN model, we also follow the ensembling procedure described in Subsection 3.3 to perform long-term forecasting. Besides the CNN model, we also consider three representative mortality models that have been widely used in the literature, including Lee-Carter Model (LC; Lee and Carter, 1992), Cairns-Blake-Dowd Model (CBD; Cairns et al., 2006), and APC model (APC; Currie, 2006).

5.2 Prediction performance

This subsection presents prediction performance in the test set for different models. Table 2 summarizes the results for all benchmark models. Columns 2-4 show results of classic mortality models, including LC, CBD, and APC. Column 5 shows results of CNN model that is trained for each individual population. The last column (CNN(All)) shows results of the CNN model with all population data included in the model training. Table 2 presents prediction MAPE for each forecast year, as well as the average prediction MAPE of the entire testing period. Comparing these benchmarks, we can see that using single population data, CNN model has the best performance, with the MAPE of 12.88%, while LC, CBD, and APC have MAPE of 16.87%, 13.72%, and 14.67%, respectively.

It's noteworthy that CNN model performs satisfyingly well in early prediction years (e.g., first five years from 2000 - 2005), as implied by much lower MAPEs than classic benchmarks. However, its prediction errors become significantly larger when the prediction horizon extends beyond 5 years. After 15 years, we can see that the MAPEs of individual CNN model are close to those of classic models such as CBD and APC.

Including more mortality data from multiple populations will significantly improve the performance, as we can see CNN(All) reduces the average MAPE from 12.88% to 10.84%. This is because using mortality data from multiple populations increases the sample size, and more importantly, takes into consideration the mortality dependence of different regions, and hence, improves model performance. This is also consistent with results reported in Wang et al. (2021).

Table 3 reports the performance of different tree-based ensemble models, using individual population data. We can see that on average, CatBoost has the best performance with an average MAPE of 10.15%. XGBoost and LightGBM both have roughly 1% higher MAPEs

Table 2. **Summary of Prediction (%) Results for Benchmark Models.**

Year	LC	CBD	APC	CNN	CNN (All)
2000	7.12	5.56	6.29	4.36	4.22
2001	8.41	6.19	7.45	5.20	4.99
2002	8.55	6.08	7.51	5.67	5.73
2003	8.88	6.60	8.09	6.54	6.97
2004	12.21	8.91	10.68	8.09	7.39
2005	12.56	9.08	10.80	8.53	8.55
2006	14.46	10.92	12.42	9.89	9.19
2007	14.89	11.42	12.92	10.55	9.56
2008	16.31	12.61	14.43	11.66	10.64
2009	17.51	13.85	15.39	12.77	10.96
2010	18.32	14.65	16.05	13.51	12.05
2011	20.36	16.75	17.57	15.32	12.62
2012	20.32	16.91	17.89	15.68	13.24
2013	21.38	17.96	18.91	16.53	13.74
2014	23.06	19.65	19.38	18.17	13.60
2015	21.43	18.13	18.54	17.14	13.50
2016	22.56	19.32	19.53	18.18	14.11
2017	22.39	19.04	19.10	18.60	14.01
2018	22.46	18.94	18.89	18.51	14.41
2019	23.14	20.03	18.74	18.56	14.41
2020	18.04	15.47	17.55	17.05	13.68
Avg	**16.87**	**13.72**	**14.67**	**12.88**	**10.84**

than that of CatBoost. Comparing the results in Table 3 with Columns 2-5 in Table 2, we can see that tree-based boosting models perform much better than classic mortality models as well as its CNN competitor. The improvements of the best tree-based model are 6.72%, 3.57%, 4.52%, and 2.73% over LC, CBD, APC, and individual CNN, respectively. It's also important to note that even compared to CNN(All) model, the average prediction error of the best tree boosting model is still lower (10.15% v.s. 10.84%). This demonstrates the prediction power of model boosting in the application of mortality forecasting.

Table 4 reports the performance of tree-based ensemble models, using all mortality data. As expected, the prediction performance is further improved by including more mortality data into the model training procedure. The average MAPEs of XGBoost, CatBoost and LightGBM are 8.43%, 8.64%, and 8.82%, respectively. The best model with all mortality data is XGBoost, with MAPE improvement of 8.44%, 5.28%, 6.23%, 4.44%, and 2.40% over LC, CBD, APC, individual CNN, and CNN (All), respectively. CatBoost model, which performs the best using individual mortality data, performs slightly worse than XGBoost,

majorly due to its worse performance in recent years of prediction. For example, in the year 2020, the prediction error of CatBoost model is as high as 18.50%. In contrast, the MAPEs of XGBoost and LightGBM are only 11.69% and 12.17%, respectively. Figure 5 displays the prediction results of all test years for various mortality models. We can see that the tree boosting models perform consistently better than the benchmarks, especially XGBoost. In Online Appendix B, we show detailed prediction results for each individual country/region in representative years.

Table 3. **Summary of Prediction (%) Results for Different Ensemble Models**

Year	XGboost	Catboost	Lightgbm	Year	XGboost	Catboost	Lightgbm
2000	3.77	3.47	3.82	2011	13.81	11.15	13.78
2001	4.70	4.24	4.66	2012	13.94	11.63	14.05
2002	4.58	4.39	4.69	2013	14.73	12.41	14.90
2003	5.34	4.96	5.39	2014	16.07	13.45	16.11
2004	7.44	6.60	7.46	2015	14.67	13.44	14.75
2005	7.37	6.47	7.43	2016	15.40	14.01	15.51
2006	9.14	8.00	9.17	2017	15.35	14.74	15.45
2007	9.50	8.39	9.54	2018	15.02	14.75	15.31
2008	10.43	8.97	10.56	2019	15.26	14.96	15.34
2009	11.52	9.67	11.59	2020	13.99	17.27	14.11
2010	12.13	10.20	12.25	**Avg**	**11.15**	**10.15**	**11.23**

Table 4. **Summary of Prediction (%) Results for Ensemble Models with All Data Included**

Year	XGboost	Catboost	Lightgbm	Year	XGboost	Catboost	Lightgbm
2000	3.56	3.60	3.66	2011	10.35	9.68	10.66
2001	4.16	4.25	4.27	2012	10.26	9.70	10.39
2002	4.15	4.26	4.27	2013	10.58	10.14	11.00
2003	4.71	4.93	4.72	2014	11.38	10.50	12.18
2004	6.53	6.10	6.54	2015	9.86	10.31	10.71
2005	6.55	6.24	6.64	2016	10.26	10.50	11.37
2006	8.20	7.19	8.21	2017	10.66	11.12	11.32
2007	8.49	7.48	8.43	2018	9.80	11.38	10.84
2008	8.61	7.92	8.65	2019	8.73	10.72	10.17
2009	9.06	8.35	9.18	2020	11.69	18.50	12.17
2010	9.50	8.65	9.74	**Avg**	**8.43**	**8.64**	**8.82**

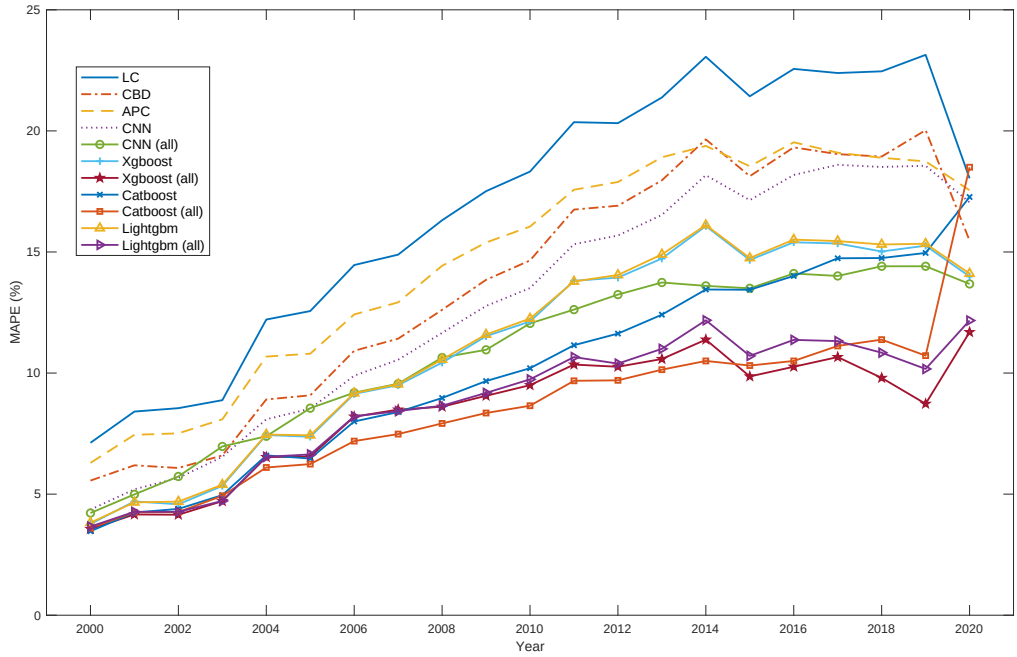


Figure 5. **Long-term prediction results of various mortality models.**

6 Conclusion

Long-term mortality forecasting is an important, yet, challenging task for the risk management of the life sector. In this paper, we propose a new long-term mortality forecasting framework. We first construct neighbouring mortality data to capture the neighbourhood effect. Then, we apply various tree boosting models to construct base prediction models for mortality, including XGBoost, CatBoost, and LightGBM. Tree boosting constructs and combines simple trees sequentially and improves model performance by learning slowly. Finally, we employ the model ensembling idea to aggregate predictions from multiple models, which not only addresses the challenge of a shrinking pattern when performing longer-term mortality forecasting, but also improves the model stability and reduces model variance and generalization errors. Our empirical analysis indicates that the prediction error of the proposed framework is only half of the classic stochastic mortality models. Moreover, the proposed model also outperforms CNN-based benchmarks.

Although this paper touches multi-population modeling by including all mortality data

and shows that the performance will be further enhanced, the focus of this paper is to perform long-term forecasting for individual population. An interesting future research direction is to formally introduce dependence models such as copulas in the current machine learning-based framework. Such multi-population extension will further facilitate research on pricing and hedging mortality/longevity risks and reduce basis risk.

References

Al-Mudafer, M. T., Avanzi, B., Taylor, G., and Wong, B. (2022). Stochastic loss reserving with mixture density neural networks. *Insurance: Mathematics and Economics*, 105:144–174.

Basak, S., Kar, S., Saha, S., Khaidem, L., and Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance*, 47:552–567.

Beven, J. L., Stewart, S. R., Lawrence, M. B., Avila, L. A., Franklin, J. L., and Pasch, R. J. (2003). Annual summary: Atlantic hurricane season of 2001. *Monthly Weather Review*, 131(7):1454–1484.

Bjerre, D. S. (2022). Tree-based machine learning methods for modeling and forecasting mortality. *ASTIN Bulletin: The Journal of the IAA*, 52(3):765–787.

Blake, D. (1999). Annuity markets: Problems and solutions. *Geneva Papers on Risk and Insurance. Issues and Practice*, 24(3):358–375.

Blake, D. and Burrows, W. (2001). Survivor bonds: Helping to hedge mortality risk. *Journal of Risk and Insurance*, 68(2):339–348.

Blake, D., Cairns, A., Coughlan, G., Dowd, K., and MacMinn, R. (2013). The new life market. *Journal of Risk and Insurance*, 80(3):501–558.

Blake, D. and Cairns, A. J. (2020). Longevity risk and capital markets: The 2018–19 update. *Annals of Actuarial Science*, 14(2):219–261.

Blake, D. and Cairns, A. J. (2021). Longevity risk and capital markets: The 2019-20 update. *Insurance: Mathematics and Economics*, 99:395–439.

Boing, A. F., Boing, A. C., Cordes, J., Kim, R., and Subramanian, S. (2020). Quantifying

and explaining variation in life expectancy at census tract, county, and state levels in the United States. *Proceedings of the National Academy of Sciences*, 117(30):17688–17694.

Bravo, J. M. (2021a). Forecasting longevity for financial applications: A first experiment with deep learning methods. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 232–249. Springer.

Bravo, J. M. (2021b). Forecasting mortality rates with recurrent neural networks. *CAPSI 2021 Proceedings*, pages 1–19.

Brock Porth, C., Porth, L., Zhu, W., Boyd, M., Tan, K. S., and Liu, K. (2020). Remote sensing applications for insurance: A predictive model for pasture yield in the presence of systemic weather. *North American Actuarial Journal*, 24(2):333–354.

Bunker, A., Wildenhain, J., Vandenbergh, A., Henschke, N., Rocklöv, J., Hajat, S., and Sauerborn, R. (2016). Effects of air temperature on climate-sensitive mortality and morbidity outcomes in the elderly; a systematic review and meta-analysis of epidemiological evidence. *EBioMedicine*, 6:258–268.

Cairns, A. J., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1–35.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., and Khalaf-Allah, M. (2011). Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin*, 41(1):29–59.

Carbonneau, A. (2021). Deep hedging of long-term financial derivatives. *Insurance: Mathematics and Economics*, 99:327–340.

Cardillo, G., Giordani, P., Levantesi, S., and Nigri, A. (2022). A tensor-based approach to cause-of-death mortality modeling. *Annals of Operations Research*, pages 1–20.

Castellani, G., Fiore, U., Marino, Z., Passalacqua, L., Perla, F., Scognamiglio, S., and Zanetti, P. (2018). An investigation of machine learning approaches in the Solvency II valuation framework. *Available at SSRN 3303296*.

Chan, W.-S., Li, J. S., Zhou, K. Q., and Zhou, R. (2016). Towards a large and liquid longevity market: A graphical population basis risk metric. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 41(1):118–127.

Chen, H., MacMinn, R., and Sun, T. (2015). Multi-population mortality model: A factor copula approach. *Insurance: Mathematics and Economics*, 63:135–146.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chen, Y. and Khaliq, A. Q. (2022). Comparative study of mortality rate prediction using data-driven recurrent neural networks and the Lee–Carter model. *Big Data and Cognitive Computing*, 6(4):134.

Chen, Z., Lu, Y., Zhang, J., and Zhu, W. (2023). Managing weather risk with a neural network-based index insurance. *Management Science, forthcoming*.

Chuang, S.-L. and Brockett, P. L. (2014). Modeling and pricing longevity derivatives using stochastic mortality rates and the esscher transform. *North American Actuarial Journal*, 18(1):22–37.

Cox, S. H., Lin, Y., and Wang, S. (2006). Multivariate exponential tilting and pricing implications for mortality securitization. *Journal of Risk and Insurance*, 73(4):719–736.

Currie, I. D. (2006). Smoothing and forecasting mortality rates with P-splines. https://www.actuaries.org.uk/documents/smoothing-and-forecasting-mortality-rates-p-splines-handouts. Accessed Oct 31, 2022.

Currie, I. D. (2016). On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal*, 2016(4):356–383.

Debener, J., Heinke, V., and Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. *Journal of Risk and Insurance*.

Deprez, P., Shevchenko, P. V., and Wüthrich, M. V. (2017). Machine learning techniques for mortality modeling. *European Actuarial Journal*, 7(2):337–352.

Devriendt, S., Antonio, K., Reynkens, T., and Verbelen, R. (2021). Sparse regression with multi-type regularized feature modeling. *Insurance: Mathematics and Economics*, 96:248–261.

Diao, L. and Weng, C. (2019). Regression tree credibility model. *North American Actuarial Journal*, 23(2):169–196.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Dong, Y., Huang, F., Yu, H., and Haberman, S. (2020). Multi-population mortality forecasting using tensor decomposition. *Scandinavian Actuarial Journal, forthcoming*.

Dowd, K., Cairns, A. J. G., Blake, D., Coughlan, G. D., and Khalaf-Allah, M. (2011). A gravity model of mortality rates for two related populations. *North American Actuarial Journal*, 15(2):334–356.

Farkas, S., Lopez, O., and Thomas, M. (2021). Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, 98:92–105.

Fissler, T., Merz, M., and Wüthrich, M. V. (2023). Deep quantile and deep composite triplet regression. *Insurance: Mathematics and Economics*, 109:94–112.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gabrielli, A., Richman, R., and Wüthrich, M. V. (2020). Neural network embedding of the over-dispersed poisson reserving model. *Scandinavian Actuarial Journal*, 2020(1):1–29.

Gabrielli, A. and V. Wüthrich, M. (2018). An individual claims history simulation machine. *Risks*, 6(2):29.

Gaille, S. and Sherris, M. (2011). Modelling mortality with common stochastic long-run trends. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 36(4):595–621.

Gao, G., Meng, S., and Wüthrich, M. V. (2019). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, 2019(2):143–162.

Gao, G., Wang, H., and Wüthrich, M. V. (2022). Boosting Poisson regression models with telematics car driving data. *Machine Learning*, 111(1):243–272.

Garcia-Menendez, F., Saari, R. K., Monier, E., and Selin, N. E. (2015). US air quality and health benefits from avoided climate change under greenhouse gas mitigation. *Environmental Science & Technology*, 49(13):7580–7588.

Ghahari, A., Newlands, N. K., Lyubchich, V., and Gel, Y. R. (2019). Deep learning at

the interface of agricultural insurance risk and spatio-temporal uncertainty in weather extremes. *North American Actuarial Journal*, 23(4):535–550.

Gomes, C., Jin, Z., and Yang, H. (2021). Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance*, 88(3):591–624.

Haberman, S. and Renshaw, A. (2012). Parametric mortality improvement rate modelling and projecting. *Insurance: Mathematics and economics*, 50(3):309–333.

Hainaut, D. (2018). A neural-network analyzer for mortality forecast. *ASTIN Bulletin: The Journal of the IAA*, 48(2):481–508.

Henckaerts, R., Côté, M.-P., Antonio, K., and Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 25(2):255–285.

Hu, C., Quan, Z., and Chong, W. F. (2022). Imbalanced learning for insurance using modified loss functions in tree-based models. *Insurance: Mathematics and Economics*, 106:13–32.

Jarner, S. F. and Kryger, E. M. (2011). Modelling adult mortality in small populations: The SAINT model. *ASTIN Bulletin*, 41(2):377–418.

Jin, Z., Yang, H., and Yin, G. (2021). A hybrid deep learning method for optimal insurance strategies: Algorithms and convergence analysis. *Insurance: Mathematics and Economics*, 96:262–275.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Kuo, K. (2019). Deeptriangle: A deep learning approach to loss reserving. *Risks*, 7(3):97.

Lally, N. and Hartman, B. (2018). Estimating loss reserves using hierarchical bayesian gaussian process regression with input warping. *Insurance: Mathematics and Economics*, 82:124–140.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American statistical association*, 87(419):659–671.

Lee, S. C. (2021). Addressing imbalanced insurance data through zero-inflated poisson regression with boosting. *ASTIN Bulletin: The Journal of the IAA*, 51(1):27–55.

Lee, S. C. and Lin, S. (2018). Delta boosting machine with application to general insurance. *North American Actuarial Journal*, 22(3):405–425.

Levantesi, S. and Nigri, A. (2020). A random forest algorithm to improve the Lee–Carter mortality forecasting: Impact on q-forward. *Soft Computing*, 24(12):8553–8567.

Levantesi, S. and Pizzorusso, V. (2019). Application of machine learning to mortality modeling and forecasting. *Risks*, 7(1):26.

Li, H. and Hyndman, R. J. (2021). Assessing mortality inequality in the us: What can be said about the future? *Insurance: Mathematics and Economics*, 99:152–162.

Li, H., Tan, K. S., Tuljapurkar, S., and Zhu, W. (2021a). Gompertz law revisited: Forecasting mortality with a multi-factor exponential model. *Insurance: Mathematics and Economics*, 99:268–281.

Li, J. S.-H., Chan, W.-S., and Zhou, R. (2017). Semicoherent multipopulation mortality modeling: The impact on longevity risk securitization. *Journal of Risk and Insurance*, 84(3):1025–1065.

Li, J. S.-H. and Hardy, M. R. (2011). Measuring basis risk in longevity hedges. *North American Actuarial Journal*, 15(2):177–200.

Li, J. S.-H., Li, J., Balasooriya, U., and Zhou, K. Q. (2021b). Constructing out-of-the-money longevity hedges using parametric mortality indexes. *North American Actuarial Journal*, 25(sup1):S341–S372.

Li, J. S.-H. and Liu, Y. (2020). The heat wave model for constructing two-dimensional mortality improvement scales with measures of uncertainty. *Insurance: Mathematics and Economics*, 93:1–26.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of population: An extension to the classical Lee-Carter approach. *Demography*, 42(3):575–594.

Li, Y. and Forsyth, P. A. (2019). A data-driven neural network approach to optimal asset allocation for target based defined contribution pension plans. *Insurance: Mathematics and Economics*, 86:189–204.

Lin, Y. and Cox, S. H. (2005). Securitization of mortality risks in life annuities. *Journal of Risk and Insurance*, 72(2):227–252.

Lindholm, M. and Palmborg, L. (2022). Efficient use of data for lstm mortality forecasting.

European Actuarial Journal, 12(2):749–778.

Loisel, S., Piette, P., and Tsai, C.-H. J. (2021). Applying economic measures to lapse risk management with machine learning approaches. *ASTIN Bulletin: The Journal of the IAA*, 51(3):839–871.

Lopez, O., Milhaud, X., and Thérond, P.-E. (2019). A tree-based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin: The Journal of the IAA*, 49(3):741–762.

Makariou, D., Barrieu, P., and Chen, Y. (2021). A random forest based approach for predicting spreads in the primary catastrophe bond market. *Insurance: Mathematics and Economics*, 101:140–162.

Marino, M., Levantesi, S., and Nigri, A. (2023). A neural approach to improve the Lee-Carter mortality density forecasts. *North American Actuarial Journal*, 27(1):148–165.

Meng, S., Gao, Y., and Huang, Y. (2022). Actuarial intelligence in auto insurance: Claim frequency modeling with driving behavior features and improved boosted trees. *Insurance: Mathematics and Economics*, 106:115–127.

Miyata, A. and Matsuyama, N. (2022). Extending the Lee–Carter model with variational autoencoder: A fusion of neural network and Bayesian approach. *ASTIN Bulletin: The Journal of the IAA*, 52(3):789–812.

Nigri, A., Levantesi, S., and Marino, M. (2021). Life expectancy and lifespan disparity forecasting: A long short-term memory approach. *Scandinavian Actuarial Journal*, 2021(2):110–133.

Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., and Perla, F. (2019). A deep learning integrated Lee–Carter model. *Risks*, 7(1):33.

Nobre, J. and Neves, R. F. (2019). Combining principal component analysis, discrete wavelet transform and xgboost to trade in the financial markets. *Expert Systems with Applications*, 125:181–194.

Noll, A., Salzmann, R., and Wuthrich, M. V. (2020). Case study: French motor third-party liability claims. *Available at SSRN 3164764*.

Oeppen, J. and Vaupel, J. W. (2002). Broken limits to life expectancy.

Perla, F., Richman, R., Scognamiglio, S., and Wüthrich, M. V. (2021). Time-series forecast-

ing of mortality rates using deep learning. *Scandinavian Actuarial Journal*, 2021(7):572–598.

Petneházi, G. and Gáll, J. (2019). Mortality rate forecasting: Can recurrent neural networks beat the lee-carter model? *arXiv preprint arXiv:1909.05501*.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Renshaw, A. E. and Haberman, S. (2003). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33(2):255–272.

Renshaw, A. E. and Haberman, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and economics*, 38(3):556–570.

Richman, R. (2021a). Ai in actuarial science–a review of recent advances–part 1. *Annals of Actuarial Science*, 15(2):207–229.

Richman, R. (2021b). Ai in actuarial science–a review of recent advances–part 2. *Annals of Actuarial Science*, 15(2):230–258.

Richman, R. (2022). Mind the gap–safely incorporating deep learning models into the actuarial toolkit. *British Actuarial Journal*, 27:e21.

Richman, R. and Wuthrich, M. V. (2019). Lee and Carter go machine learning: recurrent neural networks. *Available at SSRN 3441030*.

Richman, R. and Wüthrich, M. V. (2019). A neural network extension of the Lee–Carter model to multiple populations. *Annals of Actuarial Science*, page forthcoming.

Riley, J. C. (2001). *Rising life expectancy: A global history*. Cambridge University Press.

Schnürch, S. and Korn, R. (2022). Point and interval forecasts of death rates using neural networks. *ASTIN Bulletin: The Journal of the IAA*, 52(1):333–360.

Son, H., Hyun, C., Phan, D., and Hwang, H. J. (2019). Data analytic approach for bankruptcy prediction. *Expert Systems with Applications*, 138:112816.

Wang, C.-W., Yang, S. S., and Huang, H.-C. (2015). Modeling multi-country mortality dependence and its application in pricing survivor index swaps—a dynamic copula approach. *Insurance: Mathematics and Economics*, 63:30–39.

Wang, C.-W., Zhang, J., and Zhu, W. (2021). Neighbouring prediction for mortality. *ASTIN*

Bulletin: The Journal of the IAA, 51(3):689–718.

Wang, J., Wen, L., Xiao, L., and Wang, C. (2023). Time-series forecasting of mortality rates using transformer. *Scandinavian Actuarial Journal*, pages 1–15.

World Bank (2019). Life expectancy at birth, total (years). https://data.worldbank.org/indicator/SP.DYN.LE00.IN. Accessed: 2022-12-15.

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal*, 7(1):89–108.

Wüthrich, M. V. (2018). Neural networks applied to chain–ladder reserving. *European Actuarial Journal*, 8(2):407–436.

Wuthrich, M. V. and Buser, C. (2021). Data analytics for non-life insurance pricing. *Swiss Finance Institute Research Paper*, (16-68).

Xin, X. and Huang, F. (2023). Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *North American Actuarial Journal*, pages 1–35.

Zhang, N., Chen, H., and Liu, J. (2022). Mortality forecasting using lstm-cnn model. *Available at SSRN 4261735*.

Zhou, K. Q. and Li, J. S.-H. (2020). Asymmetry in mortality volatility and its implications on index-based longevity hedging. *Annals of Actuarial Science*, 14(2):278–301.

Zhou, K. Q. and Li, J. S.-H. (2021). Longevity greeks: What do insurers and capital market investors need to know? *North American Actuarial Journal*, 25(sup1):S66–S96.

Zhou, R. (2019). Modelling mortality dependence with regime-switching copulas. *ASTIN Bulletin: The Journal of the IAA*, 49(2):373–407.

Zhou, R. and Ji, M. (2021). Modelling mortality dependence: An application of dynamic vine copula. *Insurance: Mathematics and Economics*, 99:241–255.

Zhou, R., Li, J. S.-H., and Tan, K. S. (2013). Pricing standardized mortality securitizations: A two-population model with transitory jump effects. *Journal of Risk and Insurance*, 80(3):733–774.

Zhou, R., Wang, Y., Kaufhold, K., Li, J. S.-H., and Tan, K. S. (2014). Modeling mortality of multiple populations with vector error correction models: Applications to solvency II. *North American Actuarial Journal*, 18(1):150–167.

Zhu, W., Tan, K. S., and Wang, C.-W. (2017). Modeling multicountry longevity risk with

mortality dependence: A Lévy subordinated hierarchical Archimedean copulas approach. *Journal of Risk and Insurance*, 84(S1):477–493.

Online Appendices

A Summary of mortality data of various populations

This appendix provides a summary of mortality data for various countries and regions used in the empirical analysis of this paper.

Table A1. **Summary of Mortality Data for Various Countries and Regions**

Abbrev	Sample Period	Sample Size	Abbrev	Sample Period	Sample Size
AUS	1921-2019	2325	ISL	1901-2018	2800
AUT	1947-2019	1675	ITA	1872-2019	3550
BEL	1841-2020	4075	JPN	1947-2020	1700
BGR	1947-2017	1625	LTU	1959-2020	1400
BLR	1959-2018	1350	LUX	1960-2020	1375
CAN	1921-2019	2325	LVA	1959-2019	1375
CHE	1876-2020	3475	NLD	1850-2019	4100
CZE	1950-2019	1600	NOR	1846-2020	4225
DNK	1835-2021	4525	NZL	1948-2013	1500
ESP	1908-2020	2675	POL	1958-2019	1400
EST	1959-2019	1375	PRT	1940-2020	1875
FIN	1878-2020	3425	RUS	1959-2014	1250
FRA	1816-2019	4950	SVK	1950-2019	1600
GBR	1922-2018	2275	SWE	1751-2020	6600
HUN	1950-2020	1625	UKR	1959-2013	1225
IRL	1950-2017	1550	USA	1933-2019	2025

B Prediction results of individual countries in representative years

In this appendix, we show the prediction results of individual countries in three representative years: 2000, 2009, and 2019.

Table B1. **Prediction results for individual countries in 2000 for benchmark models.**

Region	LC	CBD	APC	CNN	CNN (All)	Region	LC	CBD	APC	CNN	CNN (All)
AUS	5.88	3.54	5.42	2.57	2.46	ISL	13.92	16.23	14.14	12.93	12.8
AUT	5.35	6.06	6.18	4.44	2.68	ITA	5.33	3.35	5.74	5.40	1.88
BEL	4.21	3.53	3.30	2.36	3.45	JPN	10.28	6.47	6.12	2.03	4.00
BGR	3.65	5.84	2.23	3.37	3.87	LTU	8.01	6.99	6.24	5.57	4.26
BLR	9.39	7.08	8.97	4.10	2.68	LUX	10.21	9.90	9.80	8.53	8.14
CAN	6.32	3.68	4.77	3.77	1.49	LVA	11.14	7.93	6.28	6.02	4.97
CHE	3.92	5.25	3.85	4.50	3.27	NLD	4.04	2.88	3.91	2.76	2.25
CZE	6.00	4.04	4.99	2.80	2.46	NOR	5.98	4.06	6.07	6.49	4.02
DNK	5.80	3.60	6.44	3.95	2.55	NZD	9.53	6.42	9.23	5.37	4.22
ESP	10.02	5.41	9.26	3.34	1.53	POL	7.91	4.61	7.30	3.25	3.24
EST	7.04	6.49	5.12	4.39	4.58	PRT	6.91	3.06	6.68	2.70	2.99
FIN	5.89	4.19	5.15	4.55	3.22	RUS	5.61	6.21	4.65	1.53	7.65
FRA	5.30	6.43	6.27	2.88	2.94	SVK	5.20	4.62	3.60	3.53	6.45
GBR	9.17	4.14	7.61	5.53	2.66	SWE	5.87	4.01	4.49	5.57	5.15
HUN	10.71	6.73	10.43	4.42	4.69	UKR	5.32	5.91	4.04	1.94	9.10
IRL	9.82	5.18	9.86	4.25	3.74	USA	4.33	4.03	3.17	4.63	5.76

	LC	CBD	APC	CNN	CNN (All)
Avg	7.13	5.56	6.29	4.36	4.22

Table B2. **Prediction results for individual countries in 2000 for ensemble models.**

Region	Xgboost	Catboost	Lightgbm	Region	Xgboost	Catboost	Lightgbm
AUS	1.97	2.49	2.14	ISL	13.47	12.57	13.65
AUT	3.54	2.97	3.51	ITA	2.28	2.49	2.37
BEL	2.33	1.90	2.98	JPN	3.73	3.47	4.26
BGR	2.75	2.59	2.82	LTU	5.28	3.88	4.80
BLR	5.94	5.12	5.73	LUX	9.05	9.06	8.68
CAN	2.56	2.80	3.15	LVA	4.62	3.38	4.21
CHE	2.71	2.73	3.12	NLD	2.07	2.01	2.14
CZE	2.38	2.25	2.85	NOR	3.39	2.95	3.94
DNK	3.06	2.16	2.45	NZD	5.63	5.47	5.95
ESP	3.91	3.80	4.21	POL	3.92	2.72	3.29
EST	4.66	4.73	4.79	PRT	2.89	2.56	3.37
FIN	3.24	2.70	3.24	RUS	2.05	1.76	1.86
FRA	1.69	1.81	2.32	SVK	2.73	2.55	2.52
GBR	3.51	2.89	3.65	SWE	2.10	2.00	2.20
HUN	5.09	4.87	4.57	UKR	2.62	2.94	2.02
IRL	4.27	4.15	4.05	USA	1.22	1.21	1.50

	Xgboost	Catboost	Lightgbm
Avg	3.77	3.47	3.82

Table B3. **Prediction results for individual countries in 2000 for ensemble models with all mortality data.**

Region	Xgboost	Catboost	Lightgbm	Region	Xgboost	Catboost	Lightgbm
AUS	3.16	4.42	3.53	ISL	14.41	13.58	14.06
AUT	3.18	2.71	2.96	ITA	2.44	2.74	2.58
BEL	1.96	2.04	1.83	JPN	3.20	3.80	4.06
BGR	4.31	3.27	4.49	LTU	4.45	4.02	4.15
BLR	5.51	4.29	5.39	LUX	8.79	8.94	8.80
CAN	2.53	2.46	2.98	LVA	3.77	3.59	3.75
CHE	2.32	3.54	3.10	NLD	1.58	1.96	1.57
CZE	1.73	1.91	1.64	NOR	3.17	3.28	3.27
DNK	2.79	2.67	2.82	NZD	6.05	4.49	5.68
ESP	3.32	1.62	2.85	POL	3.07	3.35	3.06
EST	5.17	4.60	4.99	PRT	2.37	2.09	2.48
FIN	2.62	3.04	2.65	RUS	1.02	3.07	1.64
FRA	1.21	1.88	2.05	SVK	2.31	3.02	2.33
GBR	3.52	3.91	3.49	SWE	1.32	1.91	1.35
HUN	5.47	4.67	5.35	UKR	2.15	2.46	2.41
IRL	4.26	3.99	4.51	USA	0.76	1.99	1.39

	Xgboost	Catboost	Lightgbm
Avg	3.56	3.60	3.66

Table B4. **Prediction results for individual countries in 2009 for benchmark models.**

Region	LC	CBD	APC	CNN	CNN (All)	Region	LC	CBD	APC	CNN	CNN (All)
AUS	17.48	13.85	9.47	4.52	9.14	ISL	17.54	13.76	12.93	13.80	12.11
AUT	10.38	9.38	12.50	9.02	5.66	ITA	16.13	13.26	12.46	17.77	8.10
BEL	16.48	13.79	7.56	8.70	5.48	JPN	9.49	5.17	10.37	6.00	5.48
BGR	15.04	12.14	14.34	16.97	9.30	LTU	11.18	10.12	10.18	4.19	9.18
BLR	20.81	16.10	20.66	16.10	9.43	LUX	16.99	13.57	18.73	9.56	13.15
CAN	17.17	12.89	8.77	12.62	5.41	LVA	14.89	10.98	24.15	10.36	8.34
CHE	12.57	10.19	5.69	11.56	8.01	NLD	21.11	18.10	16.01	20.07	12.77
CZE	16.61	11.80	15.91	6.55	10.08	NOR	19.91	16.83	9.82	19.77	12.47
DNK	16.66	14.26	20.81	15.02	9.60	NZD	17.80	13.60	16.65	8.23	8.72
ESP	17.62	12.71	10.82	15.77	8.00	POL	18.97	14.55	19.32	11.03	10.35
EST	21.46	15.02	25.45	15.10	13.81	PRT	19.27	16.08	13.90	12.42	7.23
FIN	18.87	16.07	10.51	16.62	6.45	RUS	22.02	16.55	23.12	16.47	15.03
FRA	18.09	15.52	16.97	7.66	6.17	SVK	16.43	11.68	19.57	10.12	20.83
GBR	22.74	18.67	19.74	22.03	12.64	SWE	15.51	13.39	6.14	18.93	10.06
HUN	19.12	13.94	26.44	11.63	10.56	UKR	16.03	11.02	16.69	7.50	40.48
IRL	31.88	28.40	28.73	21.06	19.83	USA	14.19	9.80	7.94	11.62	6.84

	LC	CBD	APC	CNN	CNN (All)
Avg	17.51	13.85	15.39	12.77	10.96

Table B5. **Prediction results for individual countries in 2009 for ensemble models.**

Region	Xgboost	Catboost	Lightgbm	Region	Xgboost	Catboost	Lightgbm
AUS	5.66	9.63	8.65	ISL	12.34	11.45	12.13
AUT	7.07	7.08	7.50	ITA	7.82	5.93	8.98
BEL	5.52	4.36	5.78	JPN	10.43	6.87	11.35
BGR	13.41	7.89	12.29	LTU	3.95	7.98	4.11
BLR	20.30	19.03	19.68	LUX	12.40	12.48	13.48
CAN	10.91	8.35	12.10	LVA	5.07	8.31	3.26
CHE	4.58	4.68	5.48	NLD	18.62	13.49	17.76
CZE	6.53	7.00	6.86	NOR	13.51	8.40	15.94
DNK	10.99	8.71	8.64	NZD	10.91	8.92	11.61
ESP	17.85	16.54	17.09	POL	11.35	7.67	9.78
EST	11.66	7.43	9.80	PRT	15.72	15.04	17.21
FIN	6.27	3.50	7.07	RUS	22.28	14.29	21.17
FRA	9.16	6.98	10.18	SVK	9.86	8.83	9.29
GBR	13.81	11.67	16.71	SWE	9.38	5.18	8.23
HUN	12.39	9.68	12.06	UKR	9.15	2.30	7.77
IRL	24.27	25.75	23.47	USA	15.44	14.18	15.49

	Xgboost	Catboost	Lightgbm
Avg	11.52	9.68	11.59

Table B6. **Prediction results for individual countries in 2009 for ensemble models with all mortality data.**

Region	Xgboost	Catboost	Lightgbm	Region	Xgboost	Catboost	Lightgbm
AUS	9.47	10.97	11.43	ISL	13.20	12.47	13.20
AUT	3.81	4.51	4.46	ITA	6.71	8.44	7.95
BEL	3.96	4.70	4.14	JPN	7.45	8.85	9.34
BGR	4.48	11.49	4.96	LTU	8.13	6.58	9.14
BLR	21.79	16.98	21.90	LUX	10.32	11.48	11.23
CAN	6.91	6.72	9.71	LVA	8.47	5.33	6.43
CHE	3.33	7.93	5.49	NLD	10.44	9.49	10.93
CZE	4.24	4.17	4.07	NOR	7.30	8.17	8.04
DNK	7.43	7.93	7.00	NZD	9.32	7.72	8.23
ESP	12.28	6.35	10.07	POL	8.83	7.71	7.22
EST	5.09	5.37	4.44	PRT	10.36	10.18	10.78
FIN	6.47	6.32	6.52	RUS	21.73	10.77	17.60
FRA	4.90	6.21	9.05	SVK	7.19	5.86	6.93
GBR	12.57	12.04	12.80	SWE	2.83	5.75	4.43
HUN	11.85	6.03	10.68	UKR	5.60	4.53	3.48
IRL	23.08	21.73	23.06	USA	10.32	4.42	9.01

	Xgboost	Catboost	Lightgbm
Avg	9.06	8.35	9.18

Table B7. **Prediction results for individual countries in 2019 for benchmark models.**

Region	LC	CBD	APC	CNN	CNN (All)	Region	LC	CBD	APC	CNN	CNN (All)
AUS	23.97	20.64	9.60	5.54	14.56	ITA	21.63	19.07	13.29	27.62	13.46
AUT	10.51	12.59	13.19	10.54	8.58	JPN	15.39	13.24	11.73	9.95	8.85
BEL	24.60	21.72	7.34	17.06	8.48	LTU	19.68	16.33	18.62	10.88	15.27
CAN	22.72	18.62	7.01	19.61	8.74	LUX	18.13	13.94	22.10	8.19	17.75
CHE	16.20	13.79	13.08	17.57	12.26	LVA	25.41	23.51	44.34	22.15	15.99
CZE	23.81	18.84	22.72	5.21	15.42	NLD	27.09	24.44	18.92	29.01	19.02
DNK	29.27	27.63	35.39	28.55	20.34	NOR	31.14	28.85	9.85	31.84	21.99
ESP	23.80	19.65	13.47	25.81	14.49	POL	26.20	21.76	25.96	17.57	15.30
EST	33.55	27.51	40.14	26.42	24.69	PRT	24.06	20.87	14.32	19.62	13.31
FIN	24.14	20.45	9.17	24.93	8.17	SVK	28.59	23.77	32.34	19.14	26.64
FRA	23.35	21.35	20.58	7.99	8.78	SWE	25.54	23.75	6.87	30.95	9.78
HUN	23.68	18.26	37.52	16.96	13.78	USA	12.84	10.13	2.20	12.41	10.28

	LC	CBD	APC	CNN	CNN (All)
Avg	23.14	20.03	18.74	18.56	14.41

Table B8. **Prediction results for individual countries in 2019 for ensemble models.**

Region	Xgboost	Catboost	Lightgbm	Region	Xgboost	Catboost	Lightgbm
AUS	13.76	38.82	14.20	ITA	11.42	20.01	15.90
AUT	8.72	13.12	12.10	JPN	15.85	9.08	18.63
BEL	6.13	11.04	7.68	LTU	13.78	10.53	10.73
CAN	16.16	6.95	17.63	LUX	13.81	11.17	13.40
CHE	6.59	22.41	8.02	LVA	13.37	14.09	8.20
CZE	9.46	13.76	11.23	NLD	24.81	12.25	24.56
DNK	20.83	15.47	17.40	NOR	23.45	13.08	27.90
ESP	29.88	24.36	27.39	POL	15.57	8.92	11.36
EST	20.39	12.28	18.21	PRT	20.00	19.55	23.64
FIN	6.63	21.22	7.53	SVK	19.27	16.79	18.03
FRA	6.94	9.71	10.22	SWE	15.70	10.56	11.71
HUN	16.17	9.43	14.80	USA	17.53	14.44	17.67

	Xgboost	Catboost	Lightgbm
Avg	15.26	14.96	15.34

Table B9. **Prediction results for individual countries in 2019 for ensemble models with all mortality data.**

Region	Xgboost	Catboost	Lightgbm	Region	Xgboost	Catboost	Lightgbm
AUS	8.51	13.25	14.52	ITA	4.77	13.38	8.11
AUT	6.53	10.29	5.38	JPN	5.54	24.12	8.94
BEL	4.13	5.05	5.55	LTU	11.14	12.11	16.08
CAN	4.04	6.18	11.05	LUX	10.60	11.34	10.24
CHE	7.93	21.66	11.97	LVA	17.30	14.28	14.18
CZE	6.95	11.52	8.06	NLD	7.60	8.94	10.79
DNK	10.98	10.20	10.62	NOR	8.68	10.81	12.01
ESP	15.35	10.86	14.05	POL	10.48	7.98	12.15
EST	15.15	6.23	7.69	PRT	8.37	9.40	10.96
FIN	4.14	7.27	5.12	SVK	11.73	6.68	10.91
FRA	3.40	11.23	8.56	SWE	4.52	11.45	6.72
HUN	14.01	6.34	12.27	USA	7.65	6.82	8.18

	Xgboost	Catboost	Lightgbm
Avg	8.73	10.72	10.17