Mortality Rates Forecasting with Data Driven LSTM, Bi-LSTM and GRU: the United States Case Study

Yuan Chen [a], Abdul Q. M. Khaliq [a]

[a] Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN, USA

Abstract

Mortality rate is always an important issue in insurance industry. There have been a large number of studies applied the deep learning approaches to the time series mortality rate prediction these years. In this study we proposed the comparison study on mortality forecasting between three different recurrent neural networks, which are Long Short-Term-Memory, Bidirectional Long Short-Term Memory, Gated Recurrent unit and the Lee-Carter model. The study will be based on the yearly-age mortality rate data from 1966-2015 of the United States. We found that all the deep learning models have the comparable results to the Lee-Cater model in terms of accuracy and Bidirectional Long short-term memory gave the best prediction results among these models.

Keywords: Mortality prediction, Lee-Carter model, Long short-term memory, Bidirectional LSTM, Gated recurrent unit

1. Introduction:

Modeling and forecasting the future mortality rate is one of the most challenging problems for life actuaries. Many countries experienced the rapidly increasing life expectancy in last few decades, which even extending the difficulties of modeling and predicting the future mortality evolution. During this time, several stochastic mortality models were proposed, Including the famous Lee-Carter model by Lee and Carter (1992), a Poisson log-bilinear regression to construct the life table by Brouhns et al. (2002) and a Two-Factor model for stochastic mortality with parameter uncertainty, as known as Cairns-Blake-Dowd model (CBD model) by Cairns et al. (2006). Several studies proposed some improvement to the Lee-Carter model. Renshaw and Haberman (2006) proposed a cohort-based extension to the Lee-Carter model. Deprez et al. (2017) and Levantesi, Pizzorusso (2019) proposed a random forest algorithm approached to the Lee-Carter model . Recently, with the developing of the machine learning and deep learning. The neural networks started to be used in the mortality forecasting field. As a strong and flexible model, the neural

networks is well fitting to the multiple features problems and could make the prediction with different time steps, e.g. weekly data and daily data. Some of the researches tried to combine the deep learning and Lee-Carter model. Nigri et al. (2019) applied the recurrent networks with Long Short-Term Memory (LSTM) architecture to predict the time index of Lee-Carter model and compared with the ARIMA models in several countries and both genders. Nigri et al. (2021) provided an estimation of parameter uncertainty of the LSTM, besides the point estimate in Lee-Carter model. Marino and Levantesi (2020) measuring longevity risk through a neural network Lee-Carter model. However, this paper more focus on the comparison of the prediction results between Lee-Carter model and deep learning models. Similar to the research by G´abor Petnehazi and Jozsef Gall (2019), they proposed a comparison study on mortality prediction with LSTM model and Lee-Cater model on countries all around world.

In this paper we considered the case study of the US mortality rate predictions between Lee-Carter model, Long short-term memory model, Bidirectional Long short-term memory model and Gated recurrent unite. We measure the forecasting result by Mean Average Error (MAE) and Root Mean Square Error (RSME) in an out of sample test.

This paper will be organized the following sections: Section 2 will introduce the data and the neural networks. The concepts of Lee-Carter model will be included in the section 3. Section 4 will introduce the metrics of the model evaluation. Section 5 will show the experiment results and section will discuss the results.

2. Data and Neural Networks

2.1 Data

The study focuses on the mortality rates. To avoid the excess mortality from the COVID-19, the data is implemented for the US mortality rate during the period 1966-2015 and considering the age 0-110+. The data are collected from the Human Mortality Database (HMD) and will be split into two sets: the training set and the test set, with the 80% train (in-sample) and 20% test (out-of-sample). We will do the numerical experiment on the log-mortality rates instead of the mortality rates for the recurrent neural networks part, inspired by Lee-Carter model. But the accuracy evaluation will be based on the mortality rate.

2.2 Long Short-Term Memory (LSTM)

The Long short-term memory is an improved version of recurrent neural networks (RNN), uses special units in addition to standard units. The recurrent neural networks (RNN） could store the information from the past by using the output of the previous unit as the input, which means it could cause some problems when facing the long-term data. As a result, Hochreiter and Schmidhuber (1997) introduced this special RNN which could store the long-term memory as well as discard some useless memory. The special structure makes Long short-term memory model well fitting in classification problems, regression problems and especially on time-series tasks. Now, let's see some mathematical function in LSTM.

The sigmoid (σ) and the tangent hyperbolic (tanh) are the most frequently nonlinear activation functions in neural networks which will be shown in equation (1-2). The sigmoid (σ) is used to decide if the information received should be uploaded and the activation functions tangent hyperbolic (tanh) is used to select the useful information and discard the less important information.

Sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$ (1)

Tangent hyperbolic: $\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$ (2)

A common LSTM unit has 3 gates mechanism which will be discussed in the following sections.

2.2.1 Forget gate

The forget gate controls the degree of information loss from the previous cell state, in another word, it decides what in formation will be dropped or kept. As we mentioned, the sigmoid function would be used in this part.

2.2.2 Input gate

The input gate controls the degree of new information to store in the current cell. It uses a sigmoid layer to decide what information would be updated in the current cell sate and then uses a tanh layer to create new vector for the cell state.

2.2.3 Cell state

We can see that the works are related to the cell state. So, after the previous works, the old cell state will be updated with the information collected from the gates. This step will be achieved by multiplying the old cell state with the weight matrix generated by forget gate and filter the original information to decide the kept part and the dropped part. Then multiply the results in the input gate to obtain the new information which would be added to the cell state.

2.2.4 Output gate

The output gate calculates the new output value of the current cell. The sigmiod layer would be used again to generate the weight matrix and we will use this matrix to decide what would be the output of the cell state. Then the weight matrix will be multiplied with the input results to output the part of the cell state.

2.2.5 Equations and Structure

The equations in the LSTM model will be shown as equation (3-8):

The relationship between the forget gate (f_t), input gate (i_t), output gate (o_t) with the input vector (x_t) and hidden units (h_t) at time t are shown as equations (3-5).

$$f_t = \sigma(W_f x_t + U_f h_{t-1}) \tag{3}$$

$$i_t = \sigma(U_i x_t + W_i h_{t-1}) \tag{4}$$

$$o_t = \sigma(U_o x_t + W_o h_{t-1}) \tag{5}$$

The Us and the Ws are the weight matrices. We use C_t to represent the current cell state, $\widetilde{C}_t$ to represent the candidate cell state. We got the following equations (6-8).

$$\widetilde{C}_t = \tanh(U_C x_t + W_C h_{t-1}) \tag{6}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \widetilde{C}_t \tag{7}$$

$$h_t = o_t \circ \tanh(C_t) \tag{8}$$

A single LSTM unit structure will be shown in figure 1.

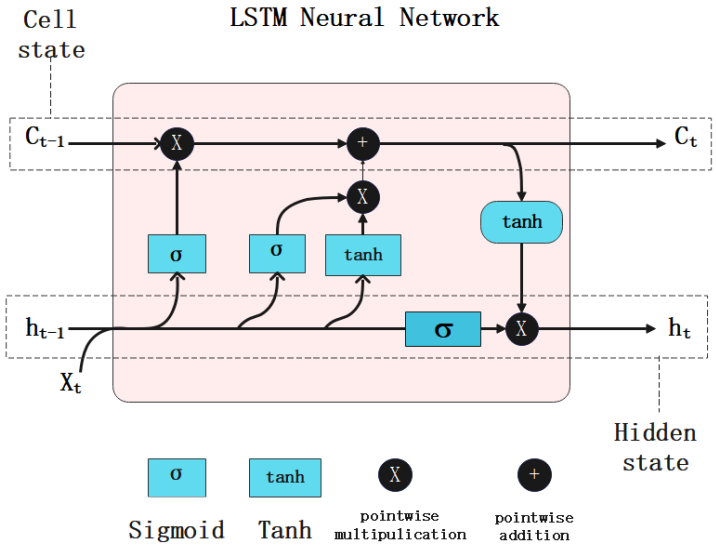

Figure1: LSTM unit structure

2.3 Bidirectional Long Short-Term Memory (Bi-LSTM)

And then, a new Long short-term memory model, named bidirectional Long short-term memory (Bi-LSTM) was invented by Schuster and Paliwa (1997). It is an extension of LSTM, the mainly different between Bi-LSTM and LSTM is, instead of one forward direction hidden layer, the bidirectional LSTM model uses two similar hidden layers with opposite directions. In the one forward direction, Bi-LSTM will learn in increasing order of sequence input and the backward direction, it would learn the information decreasing order of the sequence input. which means the both past and future information would be used. However, compared to LSTM, the Bi-LSTM model requires more to finishing the training, it will be a big challenge in practicing.

Bi-LSTM is doing well in natural language processing problems, such as sentence classification and translation. It could also be applied in handwritten recognition problem, sequence problem and some similar fields.

2.4 Gated Recurrent Unit (GRU)

The last type of recurrent neural networks is the gated recurrent unit (GRU), introduced by kyunghyun Cho et al (2014). It is quite similar to long short-term memory model (LSTM), but it has fewer parameters, gates and equations than LSTM. It merges the forget gate and input gate of

LSTM models into a single update gate. A gated recurrent unit has the following equations (9-12).

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \tag{9}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \tag{10}$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \circ h_{t-1})) \tag{11}$$

$$h_t = z_t \circ \tilde{h}_t + (1 - z_t) \circ h_{t-1} \tag{12}$$

The z_t denoted the update gate and the r_t denoted the reset gate, Ws are the weights, h_t is the output information to the next unit, $\tilde{h}_t$ is the current cell state and x_t denoted the input vector. The single gated recurrent unit structure will be shown in figure 2.

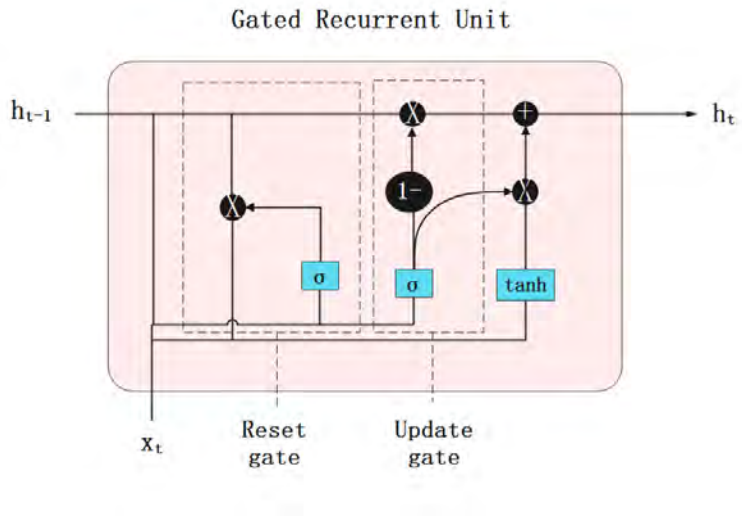

Figure2: Gated recurrent unit structure

2.5 Parameter Selection

We used the traditional three layers neural networks structure for each deep learning model (LSTM, Bi-LSTM and GRU), which contains a deep learning layer with 128 neurons, 2 dropout layers that would drop 20% of information out of the network and a dense layer. In the model compilation part, we proposed the optimizer adam and the loss function is mean square error (MSE). Each model will be trained for 300 epochs with 32 batch sizes each run.

Units in hidden layer	Batch size	Epochs	Dropout layer
128	32	300	20%

Table2: Parameters selection for LSTM/Bi-LSTM/GRU

3. Lee-Carter Model

In this part, we will discuss some important concept related to Lee-Carter model, the Lee-Carter model is a demographic model which is widely applied to the mortality prediction and life expectancy forecasting for different countries. Lee-Carter model implied a linear relationship between the mortality rate of a specific age $m_{x,t}$ and age interval x and year t. The equation usually describes it as equation 13:

$$m_{x,t} = \exp(\alpha_x + \beta_x \kappa_t) \tag{13}$$

Or we could rewrite it as equation 14:

$$\ln(m_{x,t}) = \alpha_x + \beta_x \kappa_t \tag{14}$$

Where the $m_{x,t}$ represents the central mortality rate for age group x in year t, α_x represents the specific average mortality rate, β_x means the deviation in mortality due to the age profile κ_t varies and κ_t is the time index to the year t. We use $\varepsilon_{x,t}$ to show a set of random disturbance. One more thing need to be imposed is that the Lee-Carter model is subject to the constraints on the parameters, so we have (15).

$$\sum_{x=x_1}^{x_p} \beta_x = 1 \text{ and } \sum_{t=t_1}^{t_n} \kappa_t = 0 \tag{15}$$

In practice, Singular Value Decomposition (SVD), Maximum Likelihood estimation (MLE) and Least Square (LS) are the three classical methods to estimate the parameters of Lee-Carter model. In this paper, we applied the Least Square (LS) approach to the Lee-Carter model and use the ARIMA process to estimate the time index κ_t .

4. Performance Evaluation

To measure the prediction performance, we selected 2 error criteria, mean absolute error (MAE)

and root mean square error (RMSE). The equations of the MAE and RMSE are shown by equation (16-17), which the number smaller shows the better prediction results. The total number of data is denoted by n, y_t represents the predicted mortality rate and h_t is the actual mortality rate.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - h_t| \tag{16}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - h_t)^2} \tag{17}$$

5. Result and Analysis

First, we process the prediction with Lee-Carter model. Firstly, calculate the parameters $\hat{\alpha}_x$, $\hat{\beta}_x$ and $\hat{\kappa}_t$ with the mortality data from 1966-2005, the estimations of parameters are shown in the figure3.

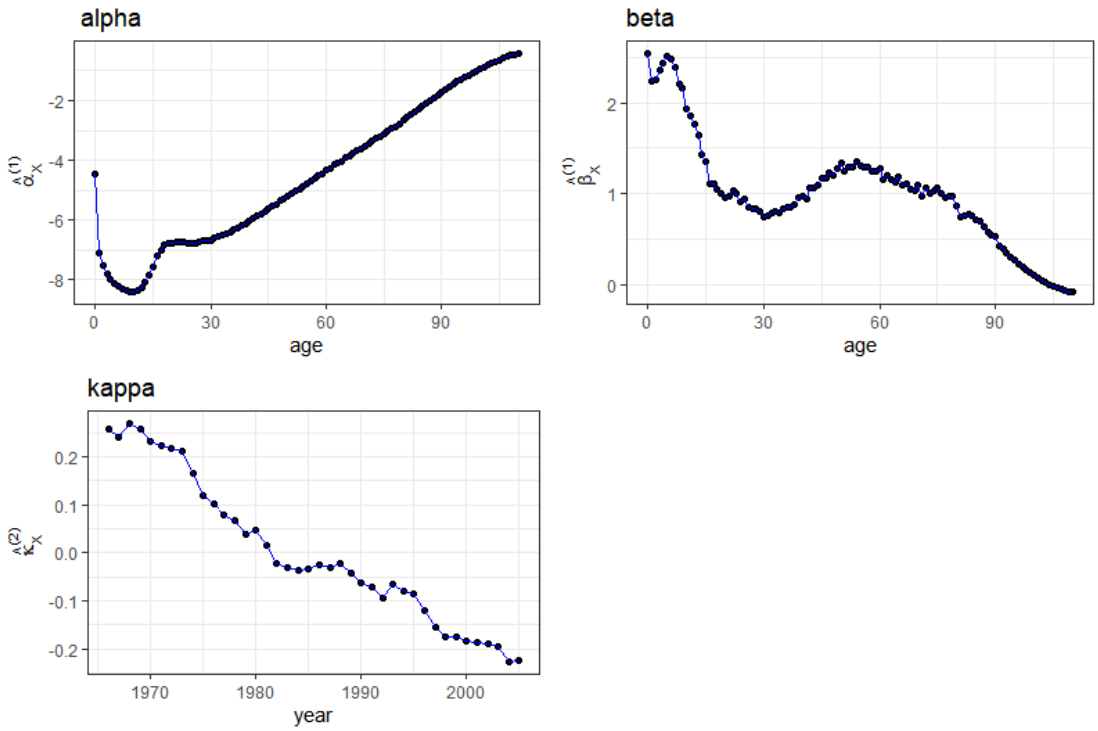

Figure 3: Parameters prediction Lee-Carter model $\hat{\alpha}_x$, $\hat{\beta}_x$ and $\hat{\kappa}_t$

Next, we used the Autoregressive Integrated Moving Average model (ARIMA) to predict the

future κ_t with the $\hat{\kappa}_t$, the prediction result will be shown in the following figure 4. In this study, the ARIMA (0,1,0) is used to achieve the prediction with AIC=164.88 and BIC=168.2.

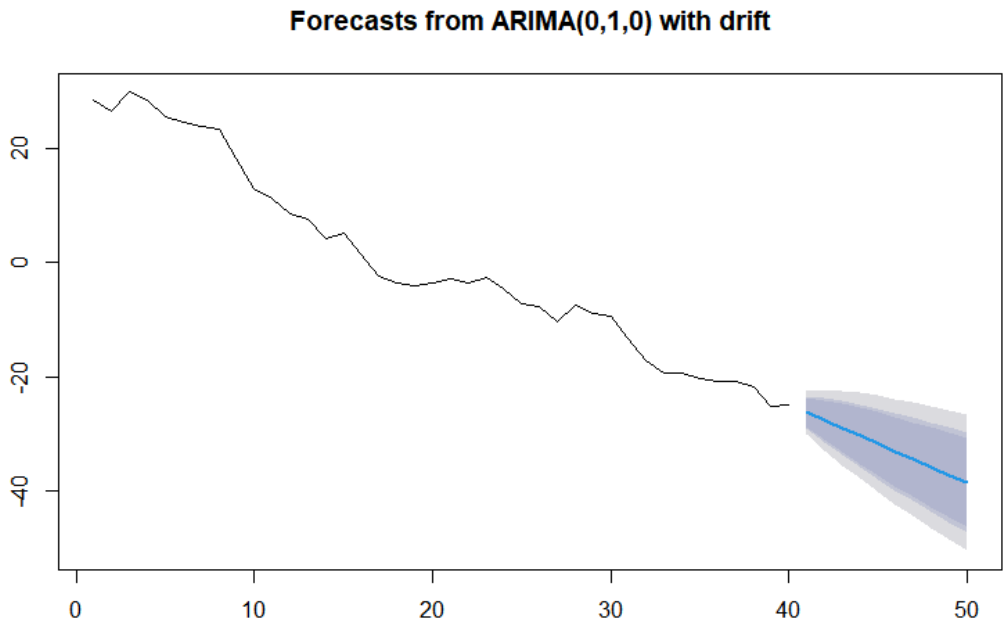

Figure 4: Forecast κ_t with ARIMA with drift

The next step is predicting the mortality rates in range 2006-2015 base on forecasting κ_t. The result will be shown later together with the deep learning models. To the deep learning model part, ten consecutive runs were conducted for each model and the mean error values were recorded. According to the average performance, the Bi- LSTM gave the best average prediction results with 0.00299874/0.0066071 (MAE/RMSE). Actually, every deep learning model gave a comparable performance to the Lee-Carter model.

Model	MAE	RMSE
Lee-Carter	0.00319467	0.0071826
LSTM	0.00326104	0.0072580
Bi-LSTM	**0.00299874**	**0.0066071**
GRU	0.00359175	0.0084659

Table2: MAE and RMSE for the models

Finally, the prediction results are shown in figure (3-5) for mortality rate predictions by LSTM, Bi-LSTM, GRU and Lee-Carter for the age 35, 55 and 100.

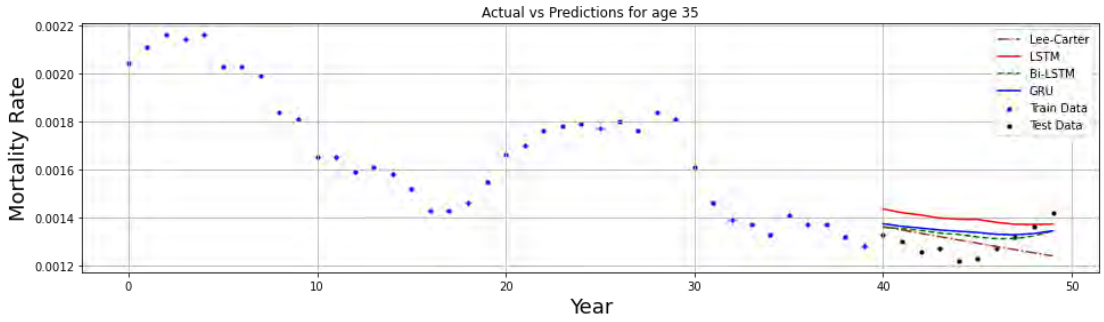

Figure3: Prediction with LC, LSTM, Bi-LSTM and GRU, x=35

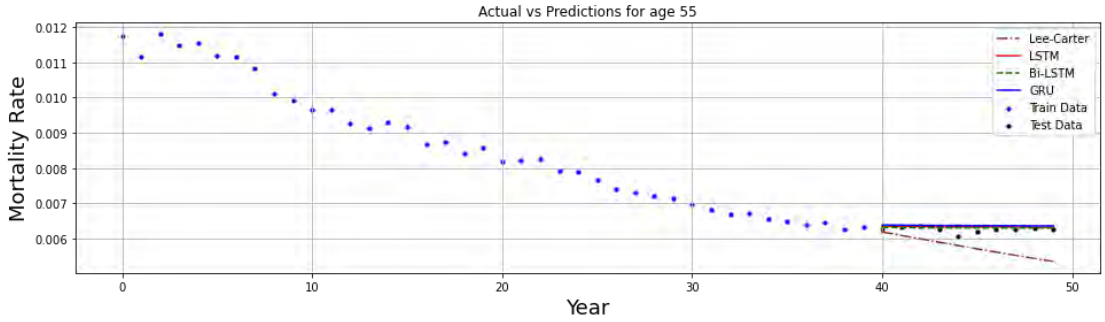

Figure4: Prediction with LC, LSTM, Bi-LSTM and GRU, x=55

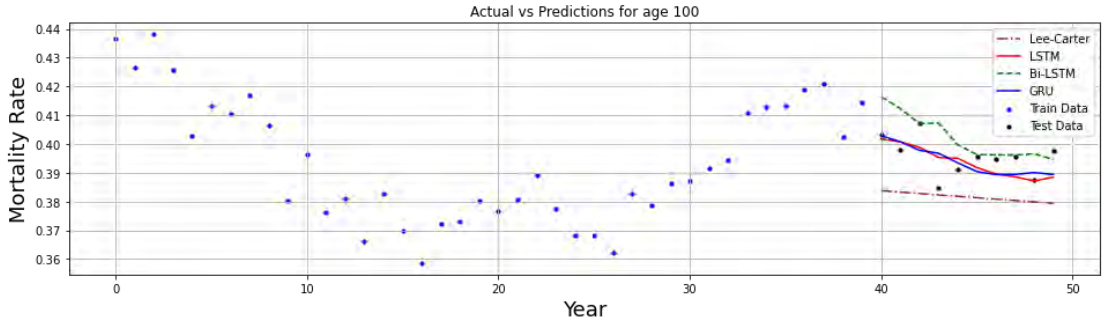

Figure5: Prediction with LC, LSTM, Bi-LSTM and GRU, x=100

From these three prediction results, can be observed that, for age 35, the Lee-Carter method and the deep learning approaches have the similar estimations, but deep learning approaches are capable to catch more future trend over the time. For some ages, such as 55, due to the pattern recognition ability, the deep learning approaches showed the better estimations than Lee-Carter model. However, for the group of age 100, it seems that all the models gave poor answers because of the data fluctuation.

6. Conclusion and Discussion

In this paper, we applied the three popular RNNs model, which are long short-term memory model, bidirectional long short-term memory model and gated recurrent units and the Lee-Carter model

on 1966-2015 USA mortality rate prediction study. A multivariate three layers deep learning networks architecture is used for each model.

The results of this study shows that the deep learning models gave comparable prediction performance to Lee-Cater model and Bi-LSTM model showed the most accurate performance among the models. The recurrent neural networks did better than Lee-Carter model on trend forecasting, more details could be caught. However, this simple prediction comparison has its limitation and shortcoming. As we know, the recurrent neural networks are data driven model, in this case, we only apply the models on the USA mortality rates, so the conclusion should more be considered as: the Bi-LSTM model gave the best forecasting results on 1966-2015 the USA mortality rates with these specific parameters. We do not know the performance of the deep learning models if we extend the forecasting period or apply on other data. Moreover, we have to show the respect to the simplicity of the Lee-Carter model, which shows the beauty of the mathematics.

The deep learning models could be improved by combining the traditional model or expanding the study to other data. To make the results more reliable, one possible method is replacing the ARIMA model with the deep learning models and explore the performance of the different RNN models under the Lee-Carter-RNN structure. That will be the next step of this study.

Reference

Broun's, N., Denuit, M. and Vermunt, J., 2002. A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3), pp.373-393.

Cairns, A., Blake, D. and Dowd, K., 2006. A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk & Insurance*, 73(4), pp.687-718.

Cho, K., van Merrienboer, B., Bahdanau, D. and Bengio, Y., 2022. *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1409.1259> [Accessed 5 September 2022].

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

Hocreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8):1735-1780.

Lee, R. and Carter, L., 1992. Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, 87(419), pp.659-671.

Levantesi, S. and Pizzorusso, V., 2019. Application of Machine Learning to Mortality Modeling and Forecasting. *Risks*, 7(1), p.26.

Marino, M. and Levantesi, S., 2020. Measuring Longevity Risk Through a Neural Network Lee-Carter Model. *SSRN Electronic Journal*,.

Marino, M., Levantesi, S. and Nigri, A., 2022. *Deepening Lee-Carter for longevity projections with*

uncertainty estimation. [online] arXiv.org. Available at: <https://arxiv.org/abs/2103.10535> [Accessed 5 September 2022].

Nigri, A., Levantesi, S. and Marino, M., 2020. Life expectancy and lifespan disparity forecasting: a long short-term memory approach. *Scandinavian Actuarial Journal*, 2021(2), pp.110-133.

Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S. and Perla, F., 2019. A Deep Learning Integrated Lee–Carter Model. *Risks*, 7(1), p.33.

Petneházi, G. and Gáll, J., 2022. *Mortality rate forecasting: can recurrent neural networks beat the Lee-Carter model?*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1909.05501> [Accessed 5 September 2022].

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