Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 10, No. 1, 2023

A Study on Chinese Residents' Willingness to Delay Retirement Based on Machine Learning Algorithm

Dongmei He, Hang Yi, Jialin Li, Yunquan Song *

College of Science, China University of Petroleum (East China), Qingdao, Shandong, 266580, China
* **Corresponding author**: Yunquan Song (Email: syqfly1980@163.com)

Abstract: At present, the problem of population aging caused by the increase of average life expectancy continues to worsen, and thus leads to a series of problems. The implementation of a delayed retirement policy in China will affect everyone. In this paper, the data of China Health and Pension tracking survey in 2018 are used. Firstly, the optimal parameter combination of deep neural network, RF, GBDT and XGBoost are obtained by grid search algorithm to conduct single model modeling. On the basis of the above four single models, linear effects are introduced by logistic regression. By using the idea of model fusion, a new prediction model is obtained by weighted fusion learning of the above five models. The important factors affecting the public's retirement intention are found and the high precision prediction of retirement intention classification is carried out.

Keywords: Retirement intention; Data driven; Machine learning; Model fusion; Public will.

1. Introduction

With the aggravation of the aging degree, delaying retirement age has become a widely accepted point of view in our country and even in the world. Most foreign scholars hold two opinions: one is to explore the choice of retirement intention from the individual perspective. The research shows that the factors influencing individual retirement intention are very diversified, and the main factors are as follows.

Gary & Fields (1982) found that the group with higher salary is more inclined to leave the job as soon as possible, and once their expected income becomes higher, they tend to retire later. Handwerker (2011) found that once affected by family factors, such as raising the elderly and their children, parents would choose to extend the retirement age. Topa et al. (2011) found that if an individual has a higher social status, he tends to delay retirement. McGarry (2004) found that workers attach great importance to their health status, and they may be inclined to retire early once they think their health status has changed significantly. Vere (2011) found that the expected income of workers is closely related to social welfare. Once the level of social welfare is increased, workers will choose to delay the retirement age due to the increase of wage level. The factors mentioned above are discussed from the perspective of individuals, and quite a few scholars analyze from the perspective of policy makers. Heijdra & Romp (2009) believe that an important incentive for people to delay retirement is that they still have pension as a source of income after retirement. Further, Cremer & Pestieau (2003) suggests that increasing the retirement age can have a positive impact on pension fund balance.

Delayed retirement is not only closely related to individuals, but also to the formulation of national policies and the revival of the Chinese nation, so once implemented, it must be supported by reliable data. This paper will make recommendations based on the results of the empirical study to provide complete data support for the implementation of the delayed retirement policy in China.

2. Method

2.1. 2.1 Fully connected neural network (DNN)

The connection mode of neural network mainly includes intra-layer connection and extra-layer connection. Meanwhile, activation function is often used in node to calculate the probability of final output. Neural network model has been widely used in big data analysis and modeling because of its fast computation speed, high prediction accuracy and strong learning ability when processing nonlinear data. The fully connected neural network has more than two hidden layers, which has deeper abstraction and dimension reduction ability. The structure of DNN is similar to that of neural network, which consists of input layer, hidden layer and output layer. It usually uses back propagation to update and iterate the parameters of the initial model and uses optimization algorithms such as gradient descent to minimize the difference between the predicted results and the observed ones.

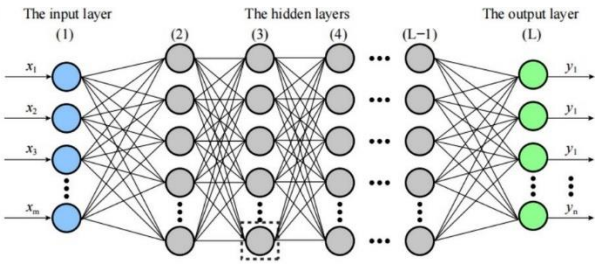

Fig. 1 DNN network structure

2.2. Random forest (RF)

Based on the idea of voting, the random forest integrates the results of multiple decision trees, and these classifiers make unified decisions. The output of the decision tree with the most votes is considered to be the final output. The classification process of random forest model is as follows.

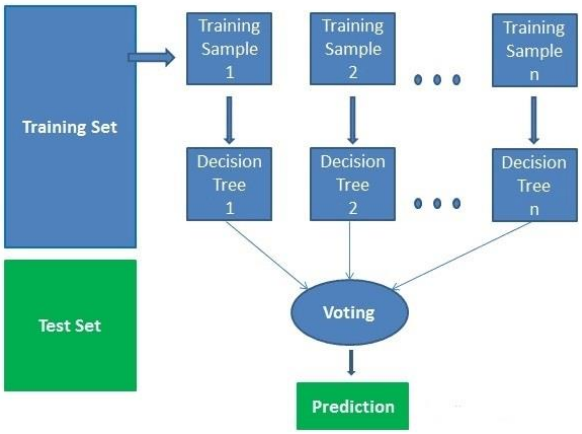

Fig. 2 RF decision process

Random forest can use errors to generate unbiased estimates of data, so as to prevent overfitting of the model. If the data distribution is unbalanced, the random forest model can still have good generalization ability.

2.3. Gradient Boosting Decision Tree (GBDT)

Through multiple iterations, each iteration generates a weak classifier, and each classifier is trained on the basis of the residual of the previous round. Each new model is built to make the residual of the previous model drop in the direction of gradient. The algorithm steps are as follows.

Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.

For $m = 1$ to M:

(a) For $i = 1,2, ...,N$ compute $r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$.

(b) Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm}, j = 1,2, ..., J_m$.

(c) For $j = 1,2, ..., J_m$ compute $\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$.

(d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

(3) Output $\hat{f}(x) = f_M(x)$.

Where y_i ($i = 1,\ldots,N$) is the dependent variable, x_i ($i = 1,\ldots,N$) is the covariate and r_{im} ($i = 1,\ldots,N, m = 1,\ldots,M$) the approximate residual.

2.4. Extreme Gradient Boosting (XGBoost)

XGBoost is an improved version of GBDT and one of Boosting algorithms. The second-order Taylor expansion of the loss function is carried out, and the specific loss function is as follows.

$$L(\phi) = \sum_{i=1}^{n} l\left(y_i', y_i\right) + \sum_k \Omega(f_k) \quad (1)$$

Where n is the number of training function samples, l is the loss of a single sample, assuming it is a convex function, and is the predicted value of the model for the training sample, and is the real label value of the training sample. The regularization term defines the complexity of the model is as follows.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| w\|^2 \quad (2)$$

Where γ and λ are manually set parameters, w is the vector formed by the values of all leaf nodes in the decision tree, and T is the number of leaf nodes.

2.5. Logistic Regression

Logistic regression is a generalized linear regression analysis model, mainly used to solve the binary classification problem, can also solve the multi-classification problem. The model is trained by a given n sets of data (training set), and the test set is classified after the training. The model form is as follows.

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \quad (3)$$

3. Research and Analysis

3.1. Data Sources

This article is based on data from the China Health and Retirement Tracking Survey, which aims to focus on more and more serious aging problem is an important data on retirement intention of Chinese middle and old people. This based on the newly released 2018 national tracking survey data, this paper investigates the respondents' willingness to delay retirement and a total of 4854 interviewees with employment experience were selected as the sample.

3.2. Model Study

3.2.1. Model Building and Optimization

In this paper, DNN, Random Forest, GBDT, XGBoost and logistic regression were used to conduct modeling analysis on the data of China's pension and health tracking survey, and each model method was optimized and analyzed. Finally, five single models were used to predict the test set, and the prediction accuracy of each model was obtained as shown in Table 1.

Table 1. Model prediction accuracy

Model	Prediction accuracy
DNN	0.7289
RF	0.7529
GBDT	0.7612
XGBoost	0.7543
Logistic Regression	0.7618

The prediction accuracy of each model is above 0.75. Based on the difference of classification accuracy of the above models, the comparison of the importance of features in different models, and the quantitative analysis of the importance of features based on the logistic regression model, this paper further carries out fusion learning of these machine learning models, and obtains a weighted fusion model with improved performance in all aspects (Weight of each model = prediction accuracy of each model/sum of prediction accuracy of five models). The prediction accuracy of the weighted fusion model is shown in the Fig. 3.

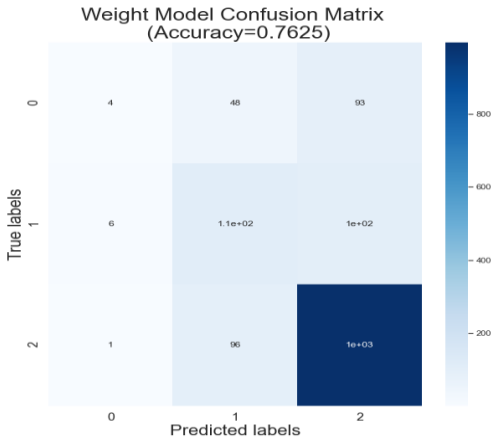

Fig. 3 Confusion matrix of weighted fusion model

The prediction accuracy of the weighted fusion model is improved compared with that of the other five single models after the fusion of nonlinear factors and linear factors, which indicates that the reprediction strategy of each single model adopted in this paper is effective after the optimal parameter adjustment of each single model. It can predict the public retirement intention well under the condition of known influencing factors.

In this paper, ROC and PR curves were further used to evaluate the predictive performance of the single model and the weighted fusion model. In both cases, the larger the area under the curve, the better the model effect. In ROC curve, the horizontal axis is FPR (false positive case rate) and the vertical axis is TPR (true case rate). PR curve is made with precision and recall as two variables. Due to the imbalance of sample data, introducing PR curve compared with ROC curve can better reflect the performance of the model.

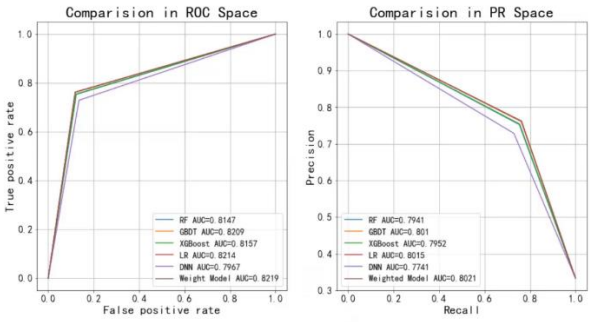

Fig. 4 ROC and PR curves of different models

In the figure, there are only three curves in each subgraph. This is because ROC curves and PR curves of some models overlap with other models, so there are only three curves in each graph. The AUC values of each model in ROC space are not different. Visually, the weighted fusion model has the best performance on the data set in this paper under ROC evaluation criteria. But the performance of DNN and weighted fusion model is not different. In PR space, the interval between three PR curves increases relative to that in ROC curve. At this time, the weighted fusion model is still the best among many models under PR evaluation index index, while the difference between the performance of DNN and that of the weighted fusion model is greater under PR index. It more intuitively reflects that the model based on the multi-model weighted fusion proposed in this paper has better performance in predicting the public's intention to delay retirement.

3.2.2. Feature Importance

In this section, the optimization results of random forest, GBDT and XGBoost integrated learning parameters obtained above are brought into the corresponding model and fitted on the training set to obtain the feature importance under different models. By averaging the importance of the same feature of different models, the final importance of the feature can be used to balance the training results of the three models, so as to facilitate the understanding of the feature importance of the subsequent fusion models. In order to compare the difference of feature importance in different models, this paper arranges the feature importance table in descending order according to the feature importance results in random forest, GBDT and XGBoost integrated learning to get three different data tables, which are show as follow.

Table 2. Feature importance that sorted by RF column (part)

Code	Feature	RF	GBDT	XGBoost	Rank
f6	age	0.085	0.101	0.018	1
f3	Monthly household consumption expenditure	0.082	0.108	0.012	2
f4	Food expenditure	0.080	0.110	0.013	3
f2	Educational background	0.058	0.057	0.011	4
f5	Other borrowings	0.046	0.073	0.018	5

Table 3. Feature importance that sorted by GBDT column (part)

Code	Feature	RF	GBDT	XGBoost	Rank
f4	Food expenditure	0.080	0.110	0.013	1
f3	Monthly household consumption expenditure	0.082	0.108	0.012	2
f6	age	0.085	0.101	0.018	3
f5	Other borrowings	0.046	0.073	0.018	4
f2	Educational background	0.058	0.057	0.011	5

In order to eliminate the influence of model selection on feature importance, the importance of the same feature of the above three models is averaged as the final feature importance, and then the mean column is sorted down, as shown in Table 5.

A comprehensive analysis of the feature importance of Table 2, 3, 4 and 5 shows that gender, age, income and expenditure have the greatest impact on residents' willingness to delay retirement.

Table 4. Feature importance that sorted by XGBoost column (part)

Code	Feature	RF	GBDT	XGBoost	Rank
f40	female	0.043	0.034	0.142	1
f66	Non-salary dividend	0.030	0.024	0.052	2
f43	Urban dweller	0.024	0.020	0.036	3
f65	Pay dividends	0.022	0.021	0.029	4
f24	Hope for the future takes up most of the day	0.018	0.017	0.025	5

Table 5. Final significance of feature (part)

Code	Feature	RF	GBDT	XGBoost	Mean	Rank
f40	Female	0.043	0.034	0.142	0.073	1
f6	age	0.085	0.101	0.018	0.068	2
f4	Food expenditure	0.080	0.110	0.013	0.068	3
f3	Monthly household consumption expenditure	0.082	0.108	0.012	0.067	4
f5	Other borrowings	0.046	0.073	0.018	0.046	5

4. Conclusions and Recommendations

The experimental results show that the weighted fusion model based on DNN, random Forest, GBDT, XGBoost and logistic regression models can improve the prediction accuracy of residents' intention to delay retirement. Through this model to intervene important factors, so as to formulate more in line with the national conditions and public demands of the policy. The main influencing factors and suggestions are as follows.

Female are more willing to delay retirement than male, and more and more female are eager to realize their social value in the workplace.

There are significant differences in retirement intention among people of different ages. The older people are, the more inclined they are to delay retirement. The reason may be the difference of people's life concept in different times. Therefore, when the state formulates the delayed retirement policy, it should consider implementing the delayed retirement policy step by step according to different age groups, and carry out publicity and education for people under the age of 50 to overcome the resistance.

The more educated people are, the more likely they are to delay retirement. Therefore, it is necessary to establish a flexible retirement mechanism. It takes a lot of time and energy to train talents for some jobs with high technical requirements. Therefore, welfare benefits should be improved to encourage them to choose to retire later. For some jobs that may harm an individual's health, such people can be encouraged to opt for early retirement and receive an early pension. It is necessary to establish a fair deferred retirement system. Fairness is the key to the retirement system and pension distribution system. We must achieve equality between civil servants and other professions, and fine-tune our policies according to the will of the people and the actual situation.

The willingness of urban residents to delay retirement is higher than that of rural residents, so the policy of postponing retirement can be implemented by different regions. The policy of postponing retirement should be implemented in big cities first and expanded from big cities to small cities. At the same time, high-quality education and employment resources are closer to small and medium-sized cities and rural areas, increasing the fairness of public resource allocation.

References

[1] Gary S. Fields. Book Review: Income Security, Insurance, and Benefits: Retirement Income Opportunities in an Aging America: Income Levels and Adequacy[J]. Industrial & Labor Relations Review, 1983, 36 (2).

[2] Elizabeth Weber Handwerker. Delaying Retirement to Pay for College[J]. Industrial & Labor Relations Review, 2011, 64 (5).

[3] Gabriela Topa, Juan A.Moriano, Marco Depolo, Carlos-María Alcover, Ana Moreno. Retirement and Wealth Relationships[J]. Research on Aging 2011, 33 (5).

[4] David Joulfaian, Kathleen McGarry. Estate and Gift Tax Incentives and Inter Vivos Giving[J]. National Tax Journal, 2004, 57 (2).

[5] James P. Vere. Social Security and elderly labor supply: Evidence from the Health and Retirement Study[J]. Labour Economics, 2011, 18 (5).

[6] Ben J. Heijdra, Ward E. Romp. Retirement, pensions, and ageing[J]. Journal of Public Economics, 2008, 93 (3).

[7] Helmuth Cremer, Pierre Pestieau.The Double Dividend of Postponing Retirement[J]. International Tax and Public Finance, 2003, 10 (4).