Advances in Social Sciences Research Journal – Vol.7, No.2
Publication Date: Feb. 25, 2020
DoI:10.14738/assrj.72.7911.

Fu, Y. (2020). Assess the heart disease risk of the Chinese elderly using a predictive model. *Advances in Social Sciences Research Journal, 7(2)* 251-262.

Assess the heart disease risk of the Chinese elderly using a predictive model

Yu Fu
PhD student,
Central University of Finance and Economics, Beijing.

ABSTRACT

The accelerating aging process worldwide makes chronic diseases the predominant risk for public health, and heart disease is in the top causes of the mortality of the elderly. Studies have verified the interventions can prevent, reduce or delay the onset of chronic diseases. This paper aims to find the domain predictors of heart disease by applying a machine learning technique Extreme Gradient Boosting to 89 predictors extracting from genetic, lifestyle, economic condition, isolation, stressful life events, nutrition and availability of medical service indexes. The individual-level data used is Chinese Longitudinal Healthy Longevity Survey with the time range of 2000 to 2002, and 2011 to 2014. We apply the imputation and oversampling technique to improve the prediction performance and use a step by step parameter tuning process to get the best hyper-parameters needed in the modeling. The fitted predictive model reaches a prediction accuracy of above 90% in the independent test data set. Comparing the first investigated period of 2000 to 2002 with the second period of 2011 to 2014, the predictors associated with economic condition play an important role in the prediction. The nutrition factor, surprisingly, does not contribute significantly to the prediction capability.

Key Words: heart disease, Extreme Gradient Boosting, predictive model, elderly in China

INTRODUCTION

Ageing is the main causes for lots of chronic diseases and functional obstacles that limit healthy life span. The accelerating ageing process worldwide and the expanding life expectancy are bringing about the growth of the population with chronic diseases, in which heart disease occupy the predominant position in terms of causes of the elderlies' mortality. Both social security system and individuals encounter heavy burden concerning medical service supply and financial cost. Previous studies found the interventions on dietary, exercise that works in lifespan extending can often postpone or prevent the onset of many chronic diseases [6]. Concerning the necessity of targeting the medical and other public health resources to prevent, reduce or delay the onset of heart disease, it is important to illustrate how different individual characteristics influence the risk of heart disease, especially the early onset of heart disease. On the basis of the illustration of heart disease predictors, promoting longer healthy lives for the elderly rather than treating the disease becomes a feasible option, which will be financially beneficiary for both the society and the household and promote the happiness of elderlies' well-being.

Functional decline in organism is far more than a natural phenomenon, in other words, an inevitable result of ageing. It is the result of not only genetic factors, but also the accumulated effects of experiences across the life span [2-5, 7]. The paper aim to develop a multi-variable heart disease risk prediction models using Extreme Gradient Boosting (Xgb) machining learning methods which are of advantage in computing large number of variables and identify the important predictors of heart disease. The model with best performance will be chosen by tuning the hyper parameters step by step using cross-validation and grid search. Regularization technique is used to avoid over-fitting. To avoid the drawbacks of the data-driven models, careful candidate predictor selection procedure and model validation process are conducted. The candidate predictors are picked from 8 categories, including genetic, lifestyle, economic condition, isolation, stressful life events, nutrition and availability of medical service after a thorough review of previous literatures. Random-forest-based imputation is used to handle the missing data. As the number of events (onset of heart disease) is far less than the number of those remain heart disease free, which leads to an imbalanced data set, we use over sampling technique to increase the samples of the rare events.

We apply the Chinese Longitudinal Healthy Longevity Survey (CLHLS) data in the model development. CLHLS is chosen as it covers the widest time range and provide rich information on interviewees' health status as well. It now has released data of seven waves, which are 1998, 2000, 2002, 2005, 2008, 2011, 2014. But since heart disease index starts to be included in the second wave 2000, we use the 2000-2002 wave data and 2011-2014 wave data for the model development procedure to make a comparison of the changes in terms of the predictors and the health state transition.

The empirical results find the economic condition factors play an important role in the prediction of heart disease, and the impact is upgrading as time pass. Predictors categorized to isolation and the availability of medical resources factors also contribute to the prediction performance. But the nutrition predictors, surprisingly, do not contribute significantly in prediction.

DATA DESCRIPTION

Chinese Longitudinal Healthy Longevity Survey (CLHLS) data is used for the model development and validation. It has a relatively complete cover of the population in China since the survey is conducted in 23 out of 31 provinces, and the population in the 23 provinces comprises 85% of the population national wide [10]. The interviewees that were dead or loss to follow are replaced by new samples geographically nearby. It also is an ongoing survey with a longer survey period from 1998 to 2014 comparing with other individual level data set such as China Health and Retirement Longitudinal Study (CHARLS), which is of great importance since we aim to compare the changes of the heart risk as time pass. The survey is conducted by Center for Healthy Aging and Development Studies of Peking University, and currently has released data of 7 waves, which are 1998-1999, 2000, 2002, 2005, 2008-2009, 2011-2012, 2014. The questionnaire covers 180 indexes in individual characteristics, family relationship, activity of daily living, body functionality, cognitive function, lifestyle, and social and family care, and provide basic health and exercise test.

Table 1: The number of interviewees in each survey year in CLHLS (Categorized by the year joins the survey

	Interview year						
	1998	2000	2002	2005	2008	2011	2014
Joining year							
1998	9090	4831	2643	1052	358	128	47
2000		11199	6315	2629	950	363	143
2002			16064	8175	4191	2514	1681
2005				15638	7472	4192	2791
2008-09					16954	8425	5245
2011-12						9765	6066
2014							7192

In the pre-predictor selection procedure, we pick 89 candidate variables coming from the 7 categories including genetic, life style, economic condition, isolation, stressful life events, nutrition and availability of medical service according to previous literatures [3,7] As we focus on the incidence of heart disease in 2 investigated periods from 2000 to 2002 and from 2011 to 2014, only the interviewees initially without heart disease at the beginning of the investigated period and has non-missing heart disease report are the valid observations for our modelling. There are 19008 observations with valid two consecutive records of heart disease from 2000 to 2014. In terms of the numbers of samples exposure to heart disease risk at the beginning of each investigated period, there are 5367 valid samples in the 2000, and 7353 valid interviewees 2011. There are 337 interviewees who get onset of heart disease in 2002 among valid samples in 2000, and 5030 interviewees remained heart disease free in 2002. In the 2011 2̃014 period, there are 403 people who were initially heart disease free but got heart disease in 2014, and 6950 people kept healthy.

We further pre-treat the valid samples in two aspects. First, as missing values exist among lots of predictors which is a common phenomenon in Chinese data, we impute the missing value using an algorithm based on random forest (missForest) in R. Second, the same as other studies on disease incidence, the onset of heart disease is a rare event, which means the samples suffering from heart disease in the follow-up investigated year only comprise a small portion of the whole samples in the initial investigated year. If we use this original dataset for the model fit, the algorithm will give more weights to majority (those who do not occur heart disease) and the prediction accuracy on the event we interested (those who onset of heart disease) in final model will be reduced. To promote the entire accuracy of both the majority and the minority class, we oversample the rare events using SMOTE algorithm in R, the principle of which is to use bootstrapping and k-nearest neighbour to create extra samples. The advantage of handling the imbalanced data using bootstrapping based algorithm is supported by lots of studies, for example, it can improve the prediction accuracy and adjust for over-fitting [1, 8, 9]. We use ten times oversampling technique to get the additional samples onset of heart disease and twice for those who do not incur heart disease in the follow-up investigated year.

So the number of samples onset of heart disease in the 2000 to 2002 wave is 3707, and the number of samples heart disease free is 6740. Correspondingly, the number of samples experience the heart disease event in the 2011 to 2014 wave is 4433.

MODEL AND METHODOLOGY

We assume the transition from heart disease free to heart disease follow a two-state Markovian process, which does not take into account the past transitions or the time spend in the previous states.

The transition from heart disease free to heart disease is assumed to be a binary distribution, and the transition probability is modelled as

$$\ln\frac{p(i)}{1-p(i)} = \beta_{static} \cdot X_{static} + \gamma_{time-varying(i)} \cdot X_{time-varying(i)} \qquad (1)$$

where $p(i)$ is the transition probability at time i, X_{static} is the static variables in the predictors which does not change with time, such as gender, residence, $X_{time-varying(i)}$ is the time varying factors such as age. The explanatory variables and the dependence variable is connected by $f(x)$ which allows non-linear relationship.

We use the Extreme Gradient Boosting (Xgb) method to do the classification. All the calculations use R (version 3.6.1). We first split the entire data set into training and testing by half.

The key point in the machine learning technique is to set the hyper parameters appropriately. So before the modelling, we first apply the training dataset to the R package Classification and regression training (caret) to tune the hyper parameters needed in Xgb step by step. We cut the training dataset into 3 folds to do an inner cross-validation to avoid over-fit. A cut off of 0.5 is used to calculate the prediction accuracy which serves as the principle to pick out the parameter value with best performance. For every parameter, we train it twice. First, we tune the parameters in a sequence of possible values as the initial value with a relatively larger gap between the values. Second, we assign the best tuned value from former step as the initial value, and then tune it in the neighbourhood, reaching a more accurate parameter candidate. If the best tuned parameter falls on the edge of the initial set, the tune process will continue with an extensive range. The tuning process will stop until the best tuned parameters shows in the middle of the interval.

We use grid search to find the initial value of the hyper parameters. The initial range of the hyper parameters are set as following steps,

1. Number of iterations (nrounds)
   A sequence from 25 to 500 using 50 as the grid in the first searching round, and 25 as the grid in the second searching round.
2. Learning rate (eta)
   A sequence from 0.1 to 0.6 using 0.1 as the grid in the first searching round, and 0.02 as the grid in the second searching round.
3. Maximum tree depth (max_tree_depth)
   A sequence from 4 to 14 using 2 as the grid in the first searching round, and 1 as the grid in the second searching round.
4. Min_child_weight
   A sequence from 1 to 16 using 3 as the grid in the first searching round, and 1 as the grid in the second searching round.
5. colsample_by tree
   The same as the setting of Min_child_weight.
6. The ratio of sub-sample (subsample)
   A sequence from 0.5 to 1 using 0.1 as the grid in the first searching round, and 0.05 as the grid in the second searching round.
7. The minimum splitting loss (gamma)
   A sequence from 0 to 1 using 0.2 as the grid in the first searching round, and 0.05 as the grid in the second searching round.

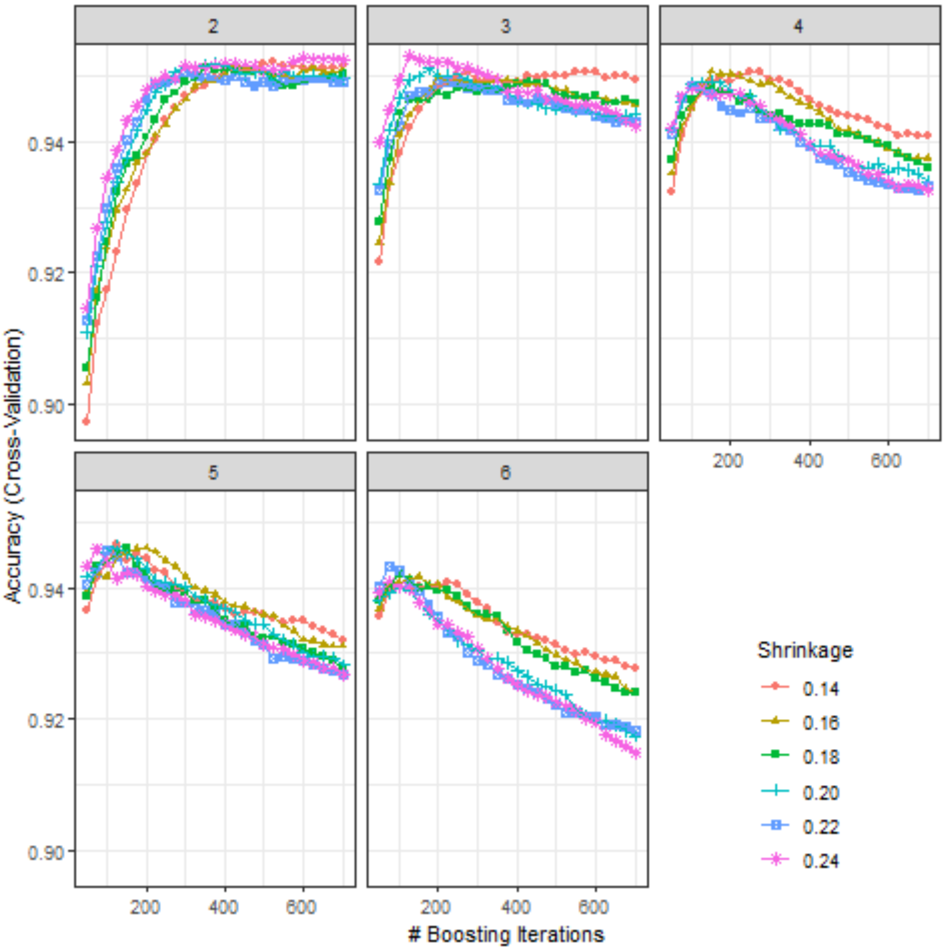
As mutual affect exists between some hyper parameters, we tune the parameters according to the steps below. If without further notation, we set the values in each step in the similar way. That is, for the parameters that wait to be tuned in the current step, set the initial values according to the former paragraph above, and for those parameters that have already been tuned in former steps, use the best tuned value. To make sure the iterations can cover the range the parameter tuning needed, the number of iterations is always from 50 to 1000 using 50 as the jump gap. Other parameters are set as default if have not been tuned. We use the data from 2011 to 2014 wave to show the parameter tuning process as follows, and the model using the data from 2000 to 2002 wave is of the same procedure.

Step 1. Tune the number of iterations, learning rate and maximum tree depth together. The prediction accuracy of each combinations of these 3 parameters using cross validation in the training set is shown in Figure 1. The best tuned values of the parameters are

Table 2: The best tuned parameter of nrounds, eta and max_tree_depth

	nrounds	max_depth	eta	gamma	colsample_ bytree	min_child_w eight	subsample
step1	1000	3	0.1	0	1	1	1

Figure 1: Paramers Tuning step 1

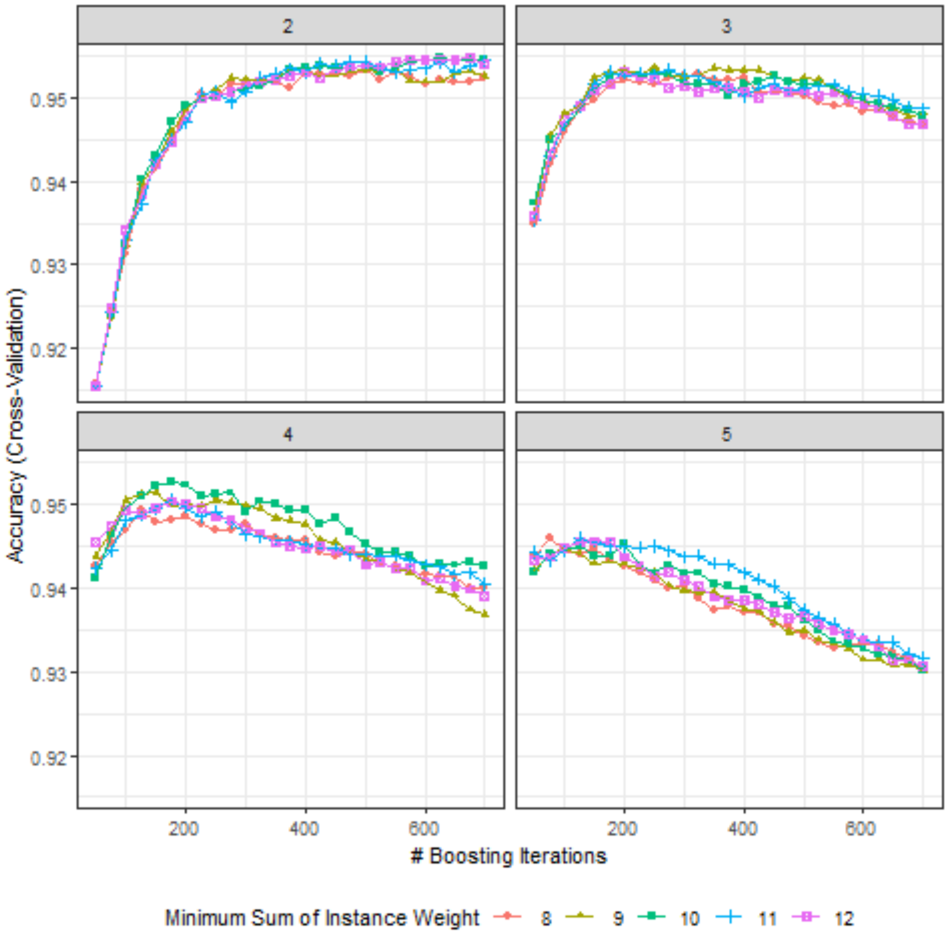


Step 2. Tune the Min_child_weight and adjust the maximum tree depth a little bit further. The changes of prediction accuracy as these two parameters change are presented in Figure 2. The parameter values get are shown in table 3.

Table 3: The best tuned parameter of Min_child_weight

	nrounds	max_d epth	eta	gamma	colsample_ bytree	min_child_ weight	subsample
step2	750	2	0.1	0	1	11	1

Figure 2: Paramers Tuning step 2



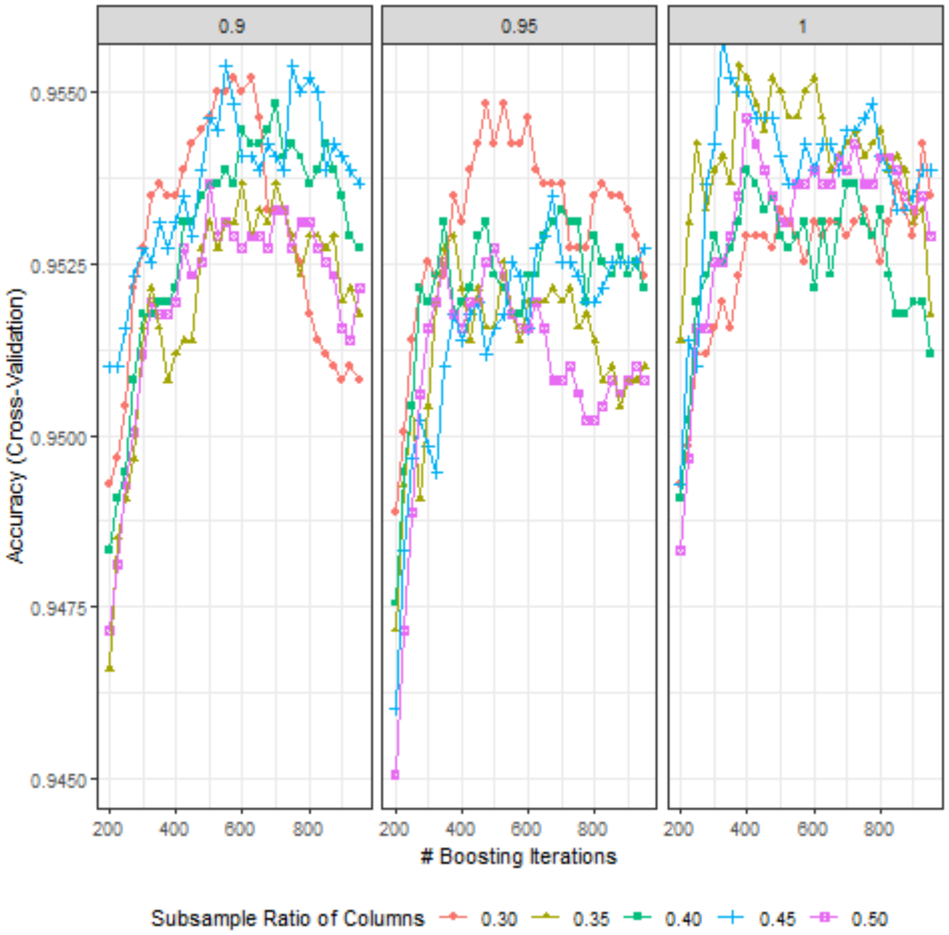
Step 3. Tune the number of variables used in each tree (colsample_bytree) and proportion of sub samples (subsample) in each training process.

The changes of prediction accuracy as these two parameters change are presented in Figure 3, and the best tuned values are

Table 4: The best tuned parameter of colsample_bytree and subsample

	nrounds	max_depth	eta	gamma	colsample_ bytree	min_child_ weight	subsample
step3	875	2	0.1	0	0.7	11	1

Figure 3: Paramers Tuning step 3

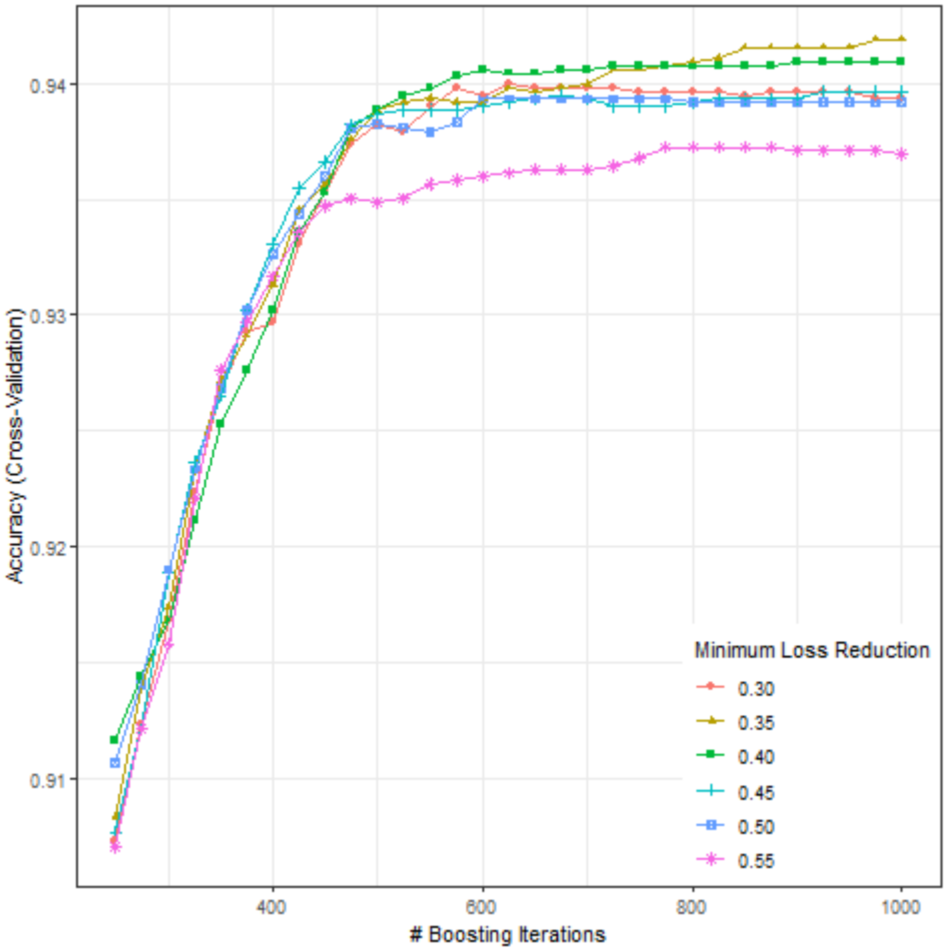


Step 4. Tune the minimum splitting loss parameter gamma The prediction accuracy for each candidate value of gamma is presented in Figure 4. The best tuned values are

Table 5: The best tuned parameter of gamma

	nrounds	max_ depth	eta	gamma	colsample_ bytree	min_child_ weight	subsample
step 4	975	2	0.1	0.35	0.7	11	1

Figure 4: Paramers Tuning step 4



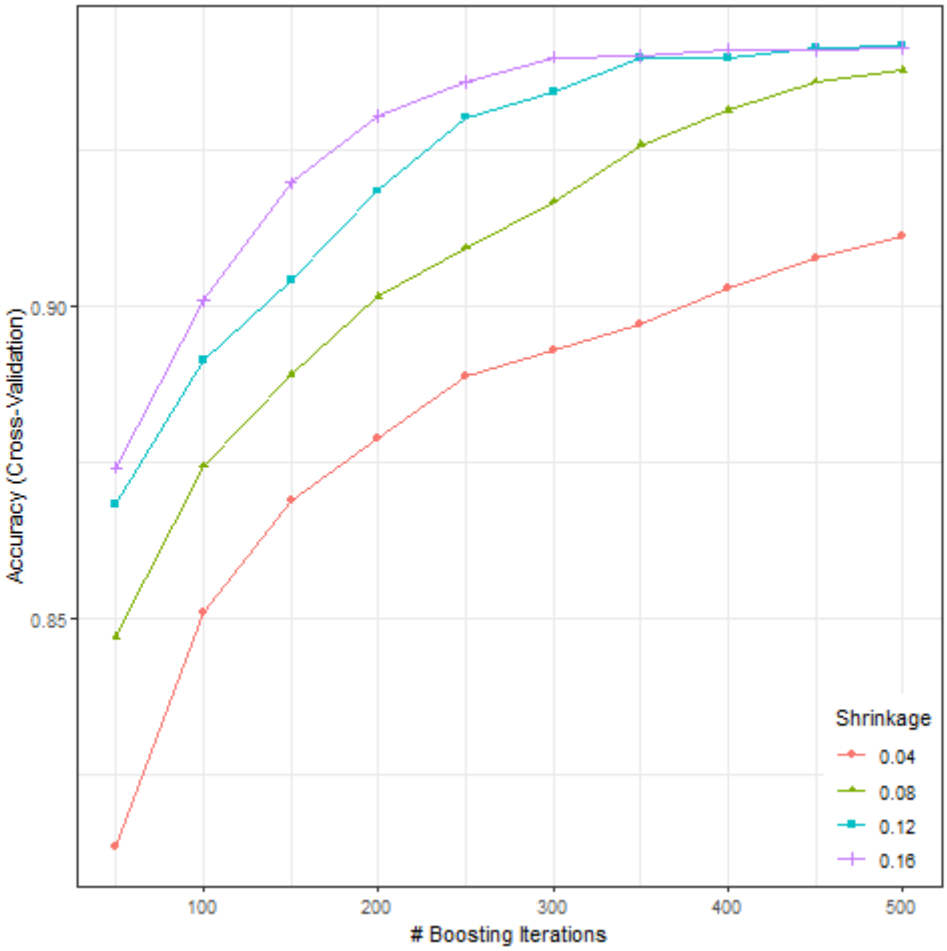
Step 5. The re-tuning of the learning rate

Tune the learning rate again to adjust with other parameters fixes in the above steps. The candidate values for searching are (0.5, 1, 1,5 ,2) times the best tuned values in the step 1. The prediction accuracy for each candidate eta is shown in Figure 5. The final best tuned hyper parameters are shown in Table 6.

Table 6: The best tuned hyper parameters

	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample
step5	500	2	0.1	0.35	0.7	11	1

Figure 5: Paramers Tuning step 5



Here we got the best tuned hyper parameters of the Xgb method. Applying these to the xgb algorithm, we get the fitted model.

RESULTS

Prediction accuracy

We first test the prediction accuracy using the test data set (Table 7).

For the investigated period from 2000 to 2002, the classification accuracy for those who transit from healthy to have heart disease is 90.8%, and for those who keep heart disease free reaches 98.4%. For the period from 2011 to 2014, the prediction accuracy for those with onset of heart disease is 90.3%, and for those who remain healthy is 96.4%. So, using an independent validation in the testing data set, the prediction accuracy of the model always keeps higher than 90%, with excessively higher accuracy in the disease-free class, which suggests the prediction power of the model is satisfying.

Table 7: The prediction accuracy of the heart disease predictive model

2000-2002			Actual		True positive	False negative
			0	1		
	Prediction	0	3315	170	90.8%	98.4%
		1	55	1683		
2011-2014	Prediction		0	1		
		0	3884	214	90.3%	96.4%
		1	146	2002		

Important predictors

The top 20 explanatory variables of heart disease risk for the period 2000 2̃002 are presented in Table 8, and for the investigated period are shown in Table 9.

Table 8: Top 20 explanatory variables of heart disease risk in the 2000 - 2002 wave

	Overall	Category	Qs
1	100	Isolation	Able to go outside to visit neighbors?
2	36.9	Lifestyle	Age when began doing physical labor
3	22.41	Isolation	Able to take public transportation?
4	11.16	Economic conditions	Number of biological siblings
5	6.18	Stressful life events	Respondent's age at father's death
6	5.51	Economic conditions	Birth order of respondent
7	3.72	Economic conditions	How many years did your mother attend school?
8	3.36	Economic conditions	Do you have the new rural cooperative medical insurance at present
9	2.82	Genetics	The 3rd sibling's age at present if alive, or age at death if died
10	2.82	Economic conditions	Father's main occupation before age 60
11	2.69	Economic conditions	Years of schooling
12	2.47	Economic conditions	During past 1 year, was your home damaged from broken pipes or heavy rain?
13	2.23	Genetics	The 1st sibling's age at present if alive, or age at death if died
14	2.19	Stressful life events	Respondent's age at mother's death
15	1.91	Lifestyle	Age when stopped doing physical labor
16	1.91	Economic conditions	Birth order of respondent
17	1.55	Economic conditions	Do you have collective medical insurance for urban residents at present
18	1.45	Isolation	Feel lonely and isolated
19	1.45	Availability of medical service	How far from your home to the nearest hospital (in kilometers)?
20	1.35	Genetics	Mother's age at death

The predictors having significant impact on the heart disease risk show some importanct features. First of all, the economic condition predictors give the most important predictive power. there are 9 predictors belong to the economic condition factor in 2000 - 2002, and up to 13 in 2011 - 2014 wave. Second, the variability of factors showing great impact shrinks as time pass. The top 20 predictors besides economic conditions also include isolation factors, life style factors, stressful life event factors and the availability of medical service factors in 2000 2̃002, while when coming to the 2011 2̃014 wave, in addition to the increasing number of predictors categorized to economic conditions, the rank of availability of medical services factor also climbs, pointing that the differences of social-economic status contributing more to the disparity of the heart disease risk. Third, the nutrition factor does not count a lot in the prediction, as there are no predictors in top 20 categorized to nutrition conditions. In fact, the first nutrition factor comes at No.33 in 2000 - 2002 wave and at No.41 in 2011 - 2014 wave.

Table 9: Top 20 explanatory variables of heart disease risk in the 2011 - 2014 wave

	Overall	Category	Qs
1	100	Economic conditions	Years of schooling
2	69.21	Availability of medical service	How far from your home to the nearest hospital (in kilometers)?
3	26.15	Economic conditions	Number of biological siblings
4	21.1	Economic conditions	Do you have a retirement pension?
5	14.85	Isolation	Able to take public transportation?
6	14.19	Economic conditions	Main source of financial support
7	13.54	Lifestyle	Age when stopped doing physical labor
8	11.78	Isolation	Time since isolation
9	11.65	Economic conditions	Do you have a retirement pension?
10	9.11	Economic conditions	Do you have medical insurance for urban workers at present
11	7.99	Economic conditions	Father's main occupation before age 60
12	7.42	Economic conditions	Do you have a retirement pension?
13	6.93	Genetics	Age
14	3.68	Economic conditions	Father's main occupation before age 60
15	3.61	Economic conditions	Was the place of birth an urban area or a rural area at time of birth?
16	3.33	Economic conditions	Do you have the new rural cooperative medical insurance at present
17	2.85	Economic conditions	How many years did your mother attend school?
18	2.69	Lifestyle	Age when began doing physical labor
19	1.68	Genetics	BMI
20	1.64	Economic conditions	Main occupation of the latest spouse before age 60

To be more specific, the top 5 predictors in 2000 - 2002 wave are isolation factor (Able to go outside to visit neighbours, Able to take public transportation), life style factor (Age when began doing physical labor), economic condition factor (Number of biological siblings), and stressful life events factor (Respondent's age at father's death). As isolation is also a kind of lifestyle, the heart disease risk in this period is closely connected to lifestyle. In the wave 2011 - 2014, the 4 in top 5 predictors are associated with social-economic conditions, including the economic condition (Years of schooling, Number of biological siblings, Do you have a retirement pension?), the availability of medical service (How far from your home to the nearest hospital (in kilometres)?) and isolation (Able to take public transportation?). The changes in the important predictors give a sign that the heart disease risk among groups of different economic conditions is expanding.

CONCLUSION

We apply the Xgboost algorithm to the incidence of heart disease modelling. Imputation technique based on random forest is used to generate values of the missing data. Oversampling method based on bootstrap and k-nearest neighbour (SMOTE from package DMwR) is used to handle the imbalanced data problem. We tune the hyper parameters in the Xgb step by step using a cross-validation based training process from 'caret' package. The fitted predictive model for heart disease risk using Xgboost performs well, reaching a prediction accuracy of above 90% in both the majority and the minority class in the independent test data set. Comparing the first investigated period of 2000 - 2002 with the second period of 2011 - 2014, the predictors associated with economic condition play a more important role in the prediction. The nutrition factor, surprisingly, does not contribute significantly in the prediction capability. We discuss these findings in light of improvements in treatments and changes in the environments of older adults.

References

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 158, no. 3, pp. 419-444.

De Leon, C. F. M., T. A. Glass, L. A. Beckett, T. E. Seeman, D. A. Evans, and L. F. Berkman (1999). Social networks and disability transitions across eight intervals of yearly data in the New Haven EPESE. The Journals of Gerontology: Series B, vol. 54B, no. 3, S162-S172.

European Innovation Partnership on Active and Healthy Ageing(EIP on AHA) (2016). Renovated Action Plan A3.

Freedman, V. A., E. M. Agree, L. G. Martin, and J. C. Cornman (2006). Trends in the use of assistive technology and personal care for late-life disability, 1992-2001. The Gerontologist, vol. 46, no. 1, pp. 124-127.

Fried, L. P. and J. M. Guralnik (1997). Disability in older adults: Evidence regarding significance, etiology, and risk. Journal of the American Geriatrics Society, vol. 45, no. 1, pp. 92-100.

Kennedy, B. K., S. L. Berger, A. Brunet, J. Campisi, A. M. Cuervo, E. S. Epel, C. Franceschi, G. J. Lithgow, R. I. Morimoto, J. E. Pessin, et al. (2014). Geroscience: linking aging to chronic disease. Cell, vol. 159, no. 4, pp. 709-713.

Kennedy, S., E. Goyder, A. Haywood, and S. Parker (2013). Ageing Populations and Age Related Health Inequalities: Evidence, issues and implications for policy and practice.

Sauerbrei, W., P. Royston, and H. Binder (2007). Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Statistics in medicine, vol. 26, no. 30, pp. 5512-5528.

Sauerbrei, W., A.-L. Boulesteix, and H. Binder (2011). Stability investigations of multivariable regression models derived from low-and high-dimensional data. Journal of biopharmaceutical statistics, vol. 21, no. 6, pp. 1206-1231.

Zeng, Y. (2004). Chinese longitudinal healthy longevity survey and some research findings. Geriatrics and Gerontology International, vol. 4, S49-S52.