SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM

TMLAI TRANSACTIONS ON
MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Volume 8 No 5
ISSN 2054-7390

Development and Validation of the Elderlies' Diabetes Risk Predictive Model using the Chinese Data

Yu Fu[1]
[1] PhD student in Central University of Finance and Economics, Beijing
fuyu_bee@live.cn

ABSTRACT

Ageing is closely related to the functional decline and is the predominant causes of the chronic diseases such as cardiovascular disease, stroke and diabetes. Population ageing worldwide accelerates the prevalence of the chronic disease. Ageing China is suffering from the diabetes risk more than other countries according to WHO reports. We adapt a machine learning algorithm Extreme Gradient Boosting to model the incidence rate of diabetes in China using a large amount of individual-level characteristic indexes as predictors. The model performance is guaranteed with a prediction accuracy above 85%, arising from the use of minority class oversampling and a multi-variable grid search technique. We apply the 2000-2002 wave and 2011-2014 wave of the Chinese Longitudinal Healthy Longevity Survey (CLHLS) to investigate how the leading predictors of the diabetes risk change as time pass. The importance of social-economic status, life-style and the access to the medical service rise in the later wave, and the relative importance of isolation and stressful life events which are related to social-psychological health decline in the investigated period, indicating a disparity of the diabetes risk within subgroups of different economic conditions.

Key Words: diabetes risk, predictors, China, CLHLS

1 Introduction

Population ageing arising from longer life expectancy and declining fertility rate is a global phenomenon, especially in China, where the population of the baby boomer steps into ageing. The population of middle old and oldest elder in China is expected to grow rapidly over the next few decades [2]. Population ageing makes the non-communicable diseases the main type of which are cardiovascular diseases, cancers, and diabetes the leading cause of mortality, since chronic diseases is highly related to the ageing group. The deaths attribute to non-communicable diseases reaches 41 million people each year, comprise 71% of all deaths globally, in which 1.6 million is induced by diabetes 2. Studies finds the chronic diseases lead to a sharp decline of the active life expectancy, and the loss of the active years due to diabetes is high for both female and male [see e.g.][1].

[2] https://www.who.int/en/news-room/fact-sheets/detail/noncommunicable-diseases

DOI: 10.14738/tmlai.85.7915
Publication Date: 12th Aug, 2020
URL: http://dx.doi.org/10.14738/tmlai.85.7915

The pathway of developing diabetes is complex since many chronic diseases are correlated to each other [7]. As the growing incidence of diabetes is an universal thesis today, attention on the preventions of diabetes guided amounts of research in recent decades, and some factors are verified correlated to the diabetes risk and thus can work as predictive factors. For example, interventions on lifestyle have been approved effective in reducing the diabetes risk [4,8]. Finds the predictors of diabetes is important as it is the basis of risk prevention. The traditional findings focus more on the clinical data, which is a collective sample of those who have already been suffering diabetes. But some areas such as insurance product pricing need evidences from the normal population, which include both those who are with the disease and those who are not, to give a more general view of the incidence of diabetes risk and the key predictors. Besides, the policy design of the public health system also relies on evidences from larger sample with greater disparity. Due to the availability of the individual-level longitudinal data of the elderlies' health status, study can be done to deeply assess strong factors contributing to healthy-diabetes transition and consequently predict the diabetes risk and its changes.

For the purpose of facilitating the health insurance product pricing and effectively locating public health resources, estimating the diabetes risks and finding predictors using a more generalized population sample set rather than clinical data is in imperative need. This paper studies the transition probability from diabetes-free to diabetes using a tree-based machine learning technique Extreme Gradient Boosting (Xgb), which is both efficiency and compatible for incorporating amounts of variables into the modelling. All model development and validation is based on R(version 3.6.1). To develop the model with best performance, hyper parameters in the Xgb are tuned by using a customized training process using R package "caret", which incorporate grid search and cross validation technique. Missing data imputation and rare events oversampling are all used to tackle the drawback of the original data set.

We use the Chinese Longitudinal Healthy Longevity Survey (CLHLS) data of 2000-2000 wave and 2011-2014 wave for the model fit. Chinese Longitudinal Healthy Longevity Survey (CLHLS) starts from 1998, and has a 2 or 3 years gap between waves with the latest available data from 2014. It covers the widest time range and provide more waves of information comparing with other individual level data set such as China Health and Retirement Longitudinal Study (CHARLS) and China Health and Nutrition Surveys (CHNS). With the advantage of the longitudinal data, the trend of the diabetes risk can be obtained easily. It collects information on health status of the elderly aged 65 and above in 22 provinces. The candidate predictors used in the modelling can be categorized to 7 categories based on literatures [3,7], including genetic factors, lifestyle factors, social-economic status factors, communication or isolation factors, stressful events factors, the access to medical resources factors and nutrition factors.

The predictive model reaches a classification accuracy of above 85%, which suggests it is a good model that can be applied in insurance industry and the public as well to assess the diabetes risk. Our empirical results provide evidence of a disparity of the diabetes risk in populations with different social-economic conditions. The economic conditions, life-style, social isolation, stressful life events and the access to the medical service all attribute to the prediction of the diabetes risk, but relative importance of social-economic conditions such as economic conditions, life-style and the access to the medical service rises.

The reminder of the paper is organized as follows. In section 2 we give the data description. In section 3 we give a brief introduction about the model and the estimation methods. In section 3 we present the estimation results based on the data. Section 4 concludes.

2 Data preprocessing and description

We choose to use the Chinese Longitudinal Healthy Longevity Survey (CLHLS) data to fit the model among surveys containing the health status information for several reasons. First, it is an ongoing survey that tracks a large sample in frequent intervals which is every two or three years starting from 1998, making it available for more waves and easier to update the study in the future . Second, it concentrates on aged 65 and above, covering the wide age scope more likely to experience the non-diabetes to diabetes transition. Third, it is based on face to face interviews rather than questionnaires, providing better data reliability. Last, its interviewees come from 22 provinces, the population in which constitute 85% of China's national population [12]. It randomly selects half of the cities and towns in these provinces to conduct the survey. As there are samples quitted from the survey due to death or lost to follow, it substitutes samples with new samples selected from neighbourhood.

The transition probability modelling of the diabetes risk uses two consecutive survey year as one investigated period, using the diabetes-free interviewees in the first survey year as the base samples and tracking their status (keep diabetes-free or incidence of diabetes) in the second survey year. To compare the change of the diabetes risk, we use two investigated periods. One is 2000-2002 wave, the other is 2011-2014 wave. Figure 1 shows the age distribution of the interviewees in the 2 waves. The samples are evenly distributed among age 67 to around age 100 ensuring enough samples at each age to generate the transition probability.

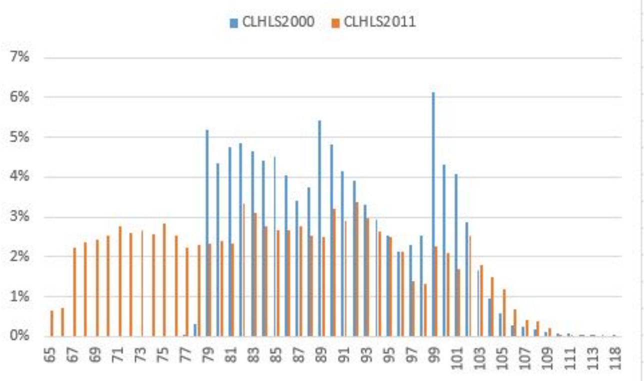

Figure 1: The age distribution of the interviewees in CLHLS (2000 and 2011 wave)

To define the valid samples, we introduce two indicator indexes. $R_k(i)$ is used to indicate if individual k at survey year i is exposed to the diabetes risk, and if $R_k(i) = 0$, the individual k is not exposed to the diabetes risk which means he/she is either already with diabetes or dead in a scenario considering only diabetes risk. If $R_k(1) = 1$, then individual k is exposed to the diabetes risk at the beginning of an investigated period, which means he/she is diabetes-free and can be counted as a valid sample. $Y_k(i)$ is to express if individual k experiences the transition from healthy (diabetes free) to diabetes between survey year i to survey year $i + 1$, with $Y_k(i) = 1$ means yes and $Y_k(i) = 1$ otherwise. So, two conditions are needed to make individual k a valid sample in the modelling. One is $R_k(1) = 1$, which means individual k is exposed to the diabetes risk, and the other is $Y_k(1)$ and $Y_k(2)$ all have valid record to acknowledge if individual k experiences the health status transition. There are 5695 interviewees exposed to the diabetes risk at the beginning of the first investigated period (2000-2002), and 8057 interviewees at the beginning of the first investigated period (2011-2014). In the first investigated period, 83 people of those valid samples got diabetes in the tracking year 2002. In the second investigated period, 144 people were onset of diabetes in the follow-up year 2014.

We select 89 variables as candidate predictors of the diabetes risk, in which 80 variables are categorical, and 19 variables are numeric. We give some technical preprocessing to the data set before the analysis. There is a large proportion of missing values in some predictors, which is the universal situation of most Chinese survey data set. Although the Xgb is tolerant to the missing values, the algorithm used to tune the hyper parameters is not missing-value adaptive. A imputation method based on random forest (missForest function in R) is conducted to replenish the missing data. In our study, the time variable is of importance concerning the accuracy of transition probability calculation, We find a majority of the dates and a fraction of the months missing when identifying the survey time, to keep as much information as possible, we time the survey conducing with the precision of month, and further use the mode of the survey months in a particular wave to replace an interviewees missing survey month if the survey year is available. We also remove the observations with contradictory variable values. For example, we include both ‘father’s death age’ and the interviewee’s ‘age at father’s death’, which will give the father’s age at the interviewee’s birth with a simple subtraction. But there exist some observations that this gap is smaller than 12, which is unlikely to happen in reality.

Another problem the original data has is the class distribution is not balanced that the number of the minority class, the transitions from diabetes-free to diabetes happens, is far less than the number of those keep diabetes-free. It makes the model give more credit to the majority class and will lower the prediction accuracy of the minority class. We use the SMOTE function from package (DMwR) to over-sample the minority class (the transition happens) by 10 times, obtaining a more balanced data set, which significantly improves the accuracy in the test validation [10,11].

3 Model estimation

The model of transitions from healthy (diabetes-free) to diabetes is developed in the Markov Chain scenario, indicating the conditional probability distribution of future states depends only upon the present state, and the dependence of the transition histories as well as the length of time spending in the previous states are not considered. It has been approved as a simple but effective assumption [5,9].

We assume the transition from healthy to diabetes follows the binary distribution and use logistic regression to model the transition probability. Then the transition probability for individual k at survey year i is given by

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

where $p(i)$ is the transition probability at investigated period i. X_{static} is the static predictors which does not vary with time, such as residence, father’s age at the interviewee’s death. $X_{time-varying(i)}$ is the time varying predictors such as age, divorce or not.

Extreme Gradient Boosting (Xgb) method is applied to train the classification. We use half of the data set as the training data. To get the best model performance, we incorporate the grid search and cross validation to the hyper parameter tuning process by writing a customized training function using the R package Classification and regression training (caret). To modify the prediction accuracy in the tuning process, we use a probability cut-off of 0.5 to classify the samples. We adapt a step by step training process which tune single or several related parameters each step and collect a set of the tuned parameters from all steps. The training data set is split into 3 folds to do a internal cross-validation to get the best tuned hyper parameters. Here we give the detailed hyper parameters tuning process as follows,

1. Tune the learning rate (eta), maximum tree depth (max_tree_depth) and the number of iterations (nrounds) together.

As these 3 parameters interact with each other, so we tune them together in the first step. To make sure the best parameters can be found in the tuning process, we first search the parameters in the domain area with a relatively large step between grids and find the possible best tuned values (round 1), then search them again in the neighborhood of the parameters got previously (round 2).

The learning rate is searched in the range from 0.1 to 0.6 in round 1 with a step jump of 0.1, and further searched with a step jump of 0.02. The maximum tree depth is searched in the range from 2 to 14 in round 1 with a step jump of 2, and further searched with a step jump of 1.

The number of iterations is searched in the range from 25 to 500 in round 1 with a step jump of 50, and further searched with a step jump of 25.

We set the multi-variables grid-search to test the prediction accuracy in the separated sub-testing dataset in the training data set. The search will go through each combination of these parameters.

The prediction accuracy based on 3-folds cross validation in the training data set against different combinations of the 3 parameters is presented in Figure 2. If without further notation, the tuning process in the steps below are of the same way as this step. The following steps share the same procedure if without specification.

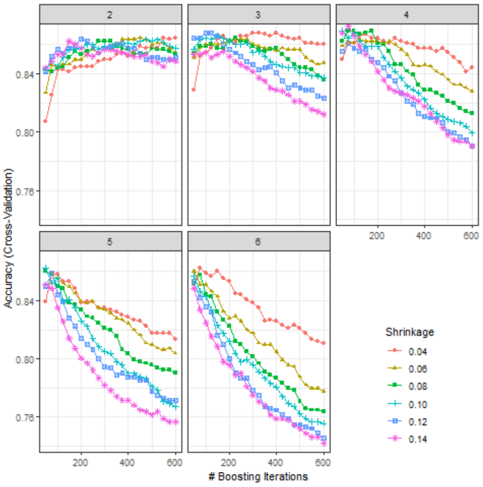

Figure 2: The prediction accuracy of different combination of eta, max_tree_depth and nrounds

2. Tune the minimum sum of instance weight (hessian) needed in a child (Min_child_weight) and explore the maximum tree depth again

As the Min_child_weight is also related to maximum tree depth, we tune the maximum tree depth in the neighbourhood again when we tune the Min_child_weight.

The Min_child_weight is searched in the range from 1 to 16 in round 1 with a step jump of 3, and further searched with a step jump of 1.

See Figure 3 for the prediction accuracy for each combination of the 2 parameters.

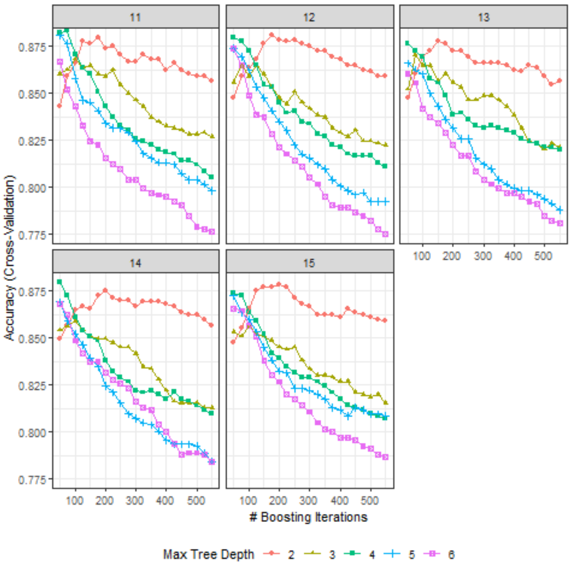

Figure 3: The prediction accuracy of different combination of Min_child_weight and max_tree_depth

3. Tune the subsample ratio of columns when constructing each tree (colsample_bytree) and sub-sample ratio of the training instances (subsample) in each training process.

The colsample_bytree is searched in the range of Min_child_weight. The "subsample" is searched in the range from 0.5 to 1 in round 1 with a step jump of 0.1, and further searched with a step jump of 0.05.

The changes of prediction accuracy as these two parameters change are presented in Figure 4.

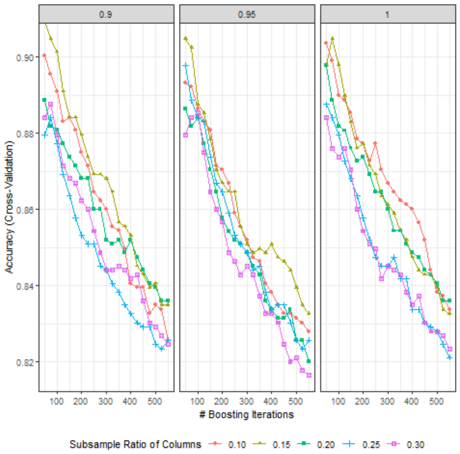

Figure 4: The prediction accuracy of different combination of colsample_bytree and subsample

4. Tune the minimum splitting loss (gamma)

To avoid over-fitting and miscalibration, we apply shrinkage or penalization procedures in the Xgb method variants of it, as it is particularly useful when a model is developed with rare events or from a very large number of predictorsambler2012evaluation. The minimum splitting loss parameter "gamma" is searched in the range from 0 to 1 in round 1 with a step jump of 0.2, and further searched with a step jump of 0.05.

The prediction accuracy for each candidate value of gamma is presented in Figure 5.

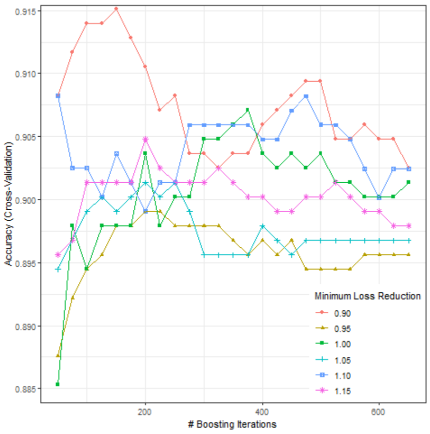

Figure 5: The prediction accuracy of different values of gamma

5. The re-tuning of the learning rate

The learning rate needs to be tuned again as other parameters have changed. Its values for the search are the current learning rate with the multiplier set (0.5, 1, 1,5 ,2), and the corresponding prediction accuracy for each candidate η is shown in Figure 6.

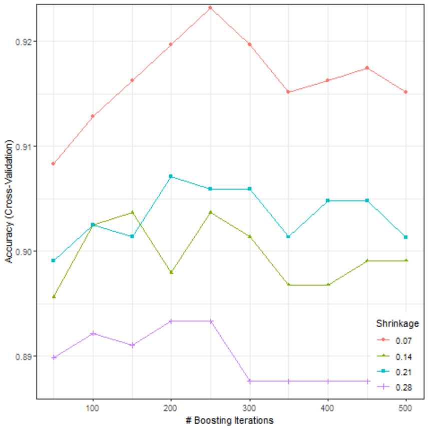

Figure 6: The prediction accuracy of different values of eta

Note need to be made that if the hyper parameter we finally get falls on the edge of the search area, the search process will restart to with an extensive range. The tuning process will stop until the best tuned parameters shows in the middle of the interval. The final best tuned hyper parameters in each step above are shown in Table 1.

Table 1: The best tuned hyper parameters in each tuning step

	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample
step1	75	4	0.14	0	1	1	1
step2	75	4	0.14	0	1	11	1
step3	50	4	0.14	0	0.15	11	0.9
step4	150	4	0.14	0.9	0.15	11	0.9
step5	250	4	0.07	0.9	0.15	11	0.9

We then use the best tuned hyper parameter to train the Xgboost algorithm to get the final fitted model.

4 Results

4.1 Model performance

By applying the fitted model into the testing data set, we get the cross-validated prediction accuracy (See Table 2). For the investigated period 2000 - 2002, the model classifies 94.1% of those who were diabetes-free in 2000 and incurred diabetes in 2002 correctly, and 89.9% of those who kept diabetes-free in 2002. For the second investigated period from 2011 to 2014, the prediction accuracy for those onsets of diabetes in 2014 is 92.3%, and for those who were still healthy is 86.0%

Table 2: The prediction accuracy of the diabetes predictive model

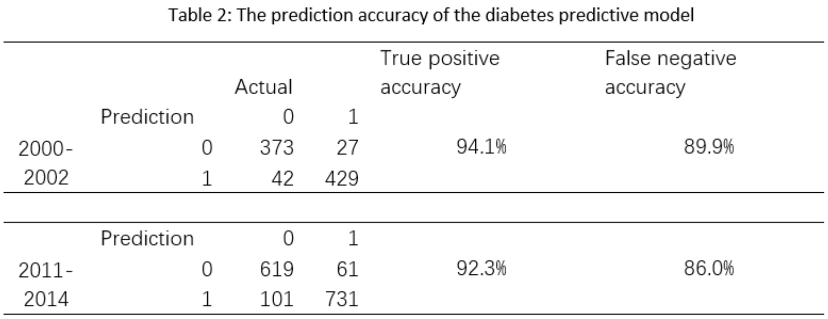

Table 2: The prediction accuracy of the diabetes predictive model

		Actual		True positive accuracy	False negative accuracy
	Prediction	0	1		
2000-2002	0	373	27	94.1%	89.9%
	1	42	429		
	Prediction	0	1		
2011-2014	0	619	61	92.3%	86.0%
	1	101	731		

The prediction accuracy keeps above 85%, indicating the model performs well in predicting the diabetes incidence.

4.2 Predictors

We rank the predictors according to their importance in predicting the diabetes incidence, and extract the top 20 most important predictors, see Table 3 for the first investigated wave, and Table 4 for the second investigated wave.

Table 3: Top 20 explanatory variables of the diabetes risk in the 2000 - 2002 wave

	Overall	Category	Qs
1	100	Isolation	Able to go outside to visit neighbors?
2	76.07	Isolation	Able to take public transportation?
3	58.34	Economic conditions	Number of biological siblings
4	28.14	Stressful life events	Respondent's age at mother's death
5	22.98	Economic conditions	Do you have the new rural cooperative medical insurance at present
6	18.36	Stressful life events	Respondent's age at father's death
7	14.22	Economic conditions	Self-reported quality of life
8	13.89	Economic conditions	Birth order of respondent
9	13.83	Genetics	The 3rd sibling's age at present if alive, or age at death if died
10	13.49	Lifestyle	Age when stopped doing physical labor

11	13.47	Isolation	Time since isolation
12	13.33	Genetics	The 2nd sibling's age at present if alive, or age at death if died
13	11.74	Economic conditions	How many years did your mother attend school?
14	10.68	Genetics	The 4th sibling's age at present if alive, or age at death if died
15	9.64	Genetics	The 1st sibling's age at present if alive, or age at death if died
16	9.53	Genetics	Mother's age at death
17	9.51	Genetics	Father's age at death
18	9.15	Economic conditions	Years of schooling
19	7.63	Economic conditions	Main occupation of the latest spouse before age 60
20	7.43	Lifestyle	Age when began doing physical labor

In the first investigated wave, economic condition, isolation, stressful life events, genetic factors and lifestyle as well all contribute to the diabetes risk, suggested by the top 20 predictors, indicating a more diversified distribution of the diabetes risk. When it comes to the 2011 - 2014 wave, the importance of economic condition and the availability of medical service rises. The number of economic condition predictors goes up to 10 in this wave, while it is 7 in the 2000 - 2002 wave.

Table 4: Top 20 explanatory variables of the diabetes risk in the 2000 - 2002 wave

	Overall	Category	Qs
1	100	Economic conditions	Years of schooling
2	54.48	Lifestyle	Age when stopped doing physical labor
3	49.41	Availability of medical service	How far from your home to the nearest hospital (in kilometers)?
4	40.85	Economic conditions	Do you have medical insurance for urban workers at present
5	40.62	Lifestyle	Age when began doing physical labor
6	37.91	Economic conditions	Number of biological siblings
7	36.07	Economic conditions	Was the place of birth an urban area or a rural area at time of birth?
8	22.95	Economic conditions	Do you have a retirement pension?
9	22.87	Economic conditions	Do you have a retirement pension?
10	22.61	Isolation	Time since isolation
11	13.92	Economic conditions	Father's main occupation before age 60
12	12.37	Isolation	Able to take public transportation?
13	12.17	Economic conditions	Do you have the new rural cooperative medical insurance at present
14	11.45	Genetics	The 3rd sibling's age at present if alive, or age at death if died
15	11.02	Economic conditions	How do you rate your economic status compared with other local people?
16	10.54	Availability of medical service	Are social and recreation services available in your community?
17	6.78	Stressful life events	Respondent's age at father's death
18	6.73	Isolation	Feel useless with age
19	6.62	Nutrition	How often eat milk products at present
20	6.28	Economic conditions	Birth order of respondent

More specifically, the top 5 predictors in the 1st wave are isolation factors (Able to go outside to visit neighbors, Able to take public transportation) ,the joint isolation and Economic conditions factors

(Number of biological siblings), economic conditions (Do you have the new rural cooperative medical insurance at present), and stressful life events (Respondent's age at mother's death). It is interesting the new rural cooperative medical insurance plays an important role in first wave, as the insurance launched in some pilot areas with better economic performance. So it can be a good proxy of the economic condition in that period. In the 2011 to 2014 wave, the importance of isolation, life style goes down, and the most important top 5 factors are economic conditions (Years of schooling ,Do you have medical insurance for urban workers at present), life style factors (Age when stopped doing physical labor, Age when began doing physical labor) and availability of medical service(How far from your home to the nearest hospital (in kilometer's)?). As both lifestyle and availability of medical service all connect to living conditions, so the social-economic conditions, to some extent, domain the diabetes risk.

5 Conclusion

We assess the diabetes risk by modelling the incidence rate of diabetes using a carefully developed predictive model. The big data analysis is introduced by incorporating 89 predictors categorized by economic condition, isolation, stressful life events, lifestyle, genetic, nutrition and availability of medical service into the modelling. A binary distribution is assumed for the transition from healthy to onset of diabetes, and the Logistic regression is adapted for the model form.

The model is developed using a machining learning technique, Extreme Gradient Boosting (Xgb). We incorporate several techniques to improve the model performance. First, we use the rare event oversampling from SMOTE method in R. Second, we use a multi-variable grid search to find the best tuned hyper parameters of the Xgb algorithm. These improve the prediction accuracy of the minority class by more than 30%, making the classification accuracy reach 90% above for those onsets of diabetes and 85% for those who kept diabetes free.

In terms of predictors, the importance of economic condition, lifestyle and the availability of medical service ascent during the investigated period from 2000 to 2014. Besides, the isolation, stressful life events and genetic factors that count in 2000 to 2002 wave are not so important in the late 2011 - 2014 wave. The change suggests a disparity of the diabetes within subgroups of different economic conditions.

REFERENCES

[1]. Belanger, A., L. Martel, J.-M. Berthelot, and R. Wilkins (2002). Gender difference in disability-free life expectancy for selected risk factors and chronic conditions in Canada. Journal of women & aging, vol. 14, no. 1-2, pp. 61-83.

[2]. China National Working Commission on Ageing (2016). The 4th Urban and Rural Elderly Life Quality Sampling Investigation in China report.

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

[4]. Group, D. P.P.D. R. et al. (2002). The Diabetes Prevention Program (DPP): Description of lifestyle intervention'. Diabetes care, vol. 25, no. 12, pp. 2165-2171.

[5]. Hoem, J. M. (1988). The versatility of the Markov chain as a tool in the mathematics of life insurance. In: Transactions of the 23rd International Congress of Actuaries. Vol. 3, pp. 171-202.

[6]. 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.

[7]. 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.

[8]. Lindström, J., Ilanne-Parikka, P., Peltonen, M., Aunola, S., Eriksson, J. G., Hemiö, K., ... & Louheranta, A. (2006). Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study. The Lancet, 368(9548), 1673-1679.

[9]. Sherris, M. and P. Wei (2019). A multi-state model of functional disability and health status in the presence of systematic trend and uncertainty. Available at SSRN 3445761.

[10]. Steyerberg, E.W., F. E. Harrell Jr, G. J. Borsboom, M. Eijkemans, Y. Vergouwe, and J. D. F.Habbema (2001). Internal validation of predictive models: effciency of some procedures for logistic regression analysis. Journal of clinical epidemiology, vol. 54, no. 8, pp. 774-781.

[11]. Steyerberg, E. W., S. E. Bleeker, H. A. Moll, D. E. Grobbee, and K. G. Moons (2003). Internal and external validation of predictive models: a simulation study of bias and precision in small samples. Journal of clinical epidemiology, vol. 56, no. 5, pp. 441-447.

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