Research on Diabetes Prediction Model Based on XGBoost Algorithm

Wu Hao

School of Economics and Management, Beijing Jiaotong University, Beijing China

Keywords: Diabetes; XGBoost; Forecast

Abstract: To explore the role of XGBoost algorithm in predicting the risk of diabetes mellitus. Pima Indian Diabetes data set in UCI machine learning database was selected and 70% of the samples were randomly selected. The plasma glucose concentration, diastolic blood pressure (mm Hg), triceps skinfold thickness (mm), 2-hour serum insulin (mu U/ml), body mass index (kg/m2), diabetic family function value and age were taken as eight factors after pregnancy, oral glucose tolerance test for 2 hours. Independent variable, with diabetes as dependent variable, based on Logistic regression and XGBoost, diabetes prediction models were established respectively. The prediction model is applied to the remaining 30% samples to evaluate the prediction effect of the model with the correct rate. The correct rates of Logistic regression model and XGBoost model were 77% and 83%, respectively. XGboost has better prediction accuracy than traditional Logistic regression.

1. Introduction

Diabetes mellitus (DM) is a metabolic disease. Its main characteristic is that the blood sugar of patients is higher than the standard value for a long time[1]. Hyperglycemia can cause symptoms commonly known as "more than three and less": eating more, drinking more, frequent urination and weight loss. With the rapid development of social economy and the improvement of people's living standards, the prevalence of diabetes is increasing year by year, and has become an important disease threatening human health. And diabetes is a chronic, refractory, lifelong disease. Only continuous medication, treatment, and will produce a series of complications. It brings tremendous life and economic pressure to patients'lives. So it is very important to predict diabetes. According to the data of the eighth edition of Diabetes World Map[2] published by IDF (International Diabetes Federation) in 2017, the number of diabetic patients in China is 114.4 million, ranking first in the world.

In recent years, many experts and scholars in this field have put forward some prediction models. Chen Yu et al. discussed the risk factors of type 2 diabetes mellitus and the application of logistic regression and BP neural network model in its risk prediction[3]. Su Ping et al. selected baseline non-type 2 diabetic patients aged 20-75 years from Shandong multi-center health management database to build a cohort, and used Cox proportional risk regression method to build a prediction model of type 2 diabetes[4]. However, these methods also have some limitations, such as over-fitting, falling into local minimum, insensitivity to random and volatile data, and unsatisfactory prediction results for unbalanced data. At the same time, the algorithms mentioned above are shallow learning algorithms, which are difficult to learn complex non-linear relationships from high-dimensional data samples, while deep learning is a stack network composed of interconnected neurons. It starts from the low-level data directly and learns to the high-level learning network of specific properties layer by layer, which effectively avoids the problem of poor training effect of traditional algorithms. This study explores the risk factors of diabetes mellitus based on diabetes related data, establishes a diabetes prediction model based on XGboost, and compares the predicted results of the model with those of the traditional logistic regression model. The aim is to find a better model for predicting the risk of type 2 diabetes mellitus, and to provide a reference for early detection and prevention of the disease.

2. Basic Theory

2.1. Basic Theory of XGBoost

XGBoost is a gradient lifting algorithm, residual decision tree, and an ensemble learning algorithm based on gradient Boosting[5]. Its basic idea is: one tree and one tree are added to the model gradually. When adding a CRAT decision tree, the overall effect should be improved. Multiple decision trees (multiple weak classifiers) are used to construct a combined classifier, and each leaf node is assigned a certain weight. It was put forward by Dr. Chen Tianqi. The decision tree mentioned here is the classification and regression tree (CART). The CART decision tree is a binary tree. The internal node features are "yes" and "no". The left branch is a branch with "yes" and the right branch is a branch with "no". CART assigns input to each leaf node according to its attributes, and each leaf node has a score. The following are some important deductive formulas of the algorithm: The objective function of XGBoost:

$$Obj(\theta)=\sum_i^n l(y_i,\hat{y}_i)+\sum_{k=1}^K \Omega(f_k) \tag{1}$$

The first part is training loss. The second part is the sum of the complexity of each tree. Take a way called additive training. That is, each iteration generates a new regression tree, so that the predicted value keeps approaching the real value (i.e. further minimizing the objective function). Each time the original model is retained unchanged, a new function f is added to the model:

$$\hat{y}_i^{(0)}=0 \tag{2}$$

$$\hat{y}_i^{(1)}=f_1(x_i)=\hat{y}_i^{(0)}+f_1(x_i) \tag{3}$$

$$\hat{y}_i^{(2)}=f_1(x_i)+f_2(x_i)=\hat{y}_i^{(1)}+f_2(x_i) \tag{4}$$

$$\hat{y}_i^{(t)}=\sum_{k=1}^t f_k(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i) \tag{5}$$

The selection must make our objective function as low as possible. Firstly, the objective function is rewritten as follows

$$obj^{(t)}=\sum_{i=1}^n\left(\hat{y}_i-\left(\hat{y}_i^{(t-1)}+f_t(x_i)\right)\right)^2+\sum_{i=1}^t\Omega(f_i) \tag{6}$$

$$=\sum_{i=1}^n\left[2\left(\hat{y}_i^{(t-1)}-y_i\right)f_t(x_i)+f_t(x_i)^2\right]+\Omega(f_t)+\text{constant} \tag{7}$$

Traditional GBDT only uses the first derivative information in optimization, while XGBoost expands the cost function with the second Taylor expansion, and uses the first derivative and the second derivative at the same time.

$$obj^{(t)}=\sum_{i=1}^n l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)+\Omega(f_t)+\text{constant} \tag{8}$$

Using Taylor expansion to approximate our original goal

$$f(x+\triangle x)\approx f(x)+f'(x)\triangle x+\frac{1}{2}f''(x)\triangle x^2 \tag{9}$$

$$g_i=\partial_{\hat{y}^{(t-1)}}l\left(y_i,\hat{y}_i^{(t-1)}\right),h_i=\partial^2_{\hat{y}^{(t-1)}}l\left(y_i,\hat{y}^{(t-1)}\right) \tag{10}$$

$$obj^{(t)}\approx\sum_{i=1}^n\left[l\left(y_i,\hat{y}_i^{(t-1)}\right)+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t)+\text{constant} \tag{11}$$

Because the minimum value of this objective function is required, the constant term C after the above formula is useless and can be removed directly. The Ω term_represents a regularization term,

which can be expressed as:

$$\Omega(f) = \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{12}$$

Then, the constant term is removed and the upper formula is introduced.

$$obj^{(t)} \approx \sum_{i=1}^{n}\left[g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2 \right] + \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{13}$$

$$= \sum_{j=1}^{T}\left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \Upsilon T \tag{14}$$

The optimal coefficients can be obtained by deriving w. Therefore, the optimal W and objective function formulas are

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{15}$$

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \Upsilon T \tag{16}$$

2.2. Basic Principles of Logistic Regression

Logistic regression, also known as logistic regression analysis, is a generalized linear regression analysis model, which is often used in data mining, automatic disease diagnosis, economic prediction and other fields[6]. In this paper, the risk factors of diabetes mellitus were discussed by logistic regression, and the probability of diabetes mellitus was predicted according to the risk factors. Independent variables can be either continuous or classified. Then through logistic regression analysis, we can get the weight of independent variables, so that we can roughly understand which factors are the risk factors of disease. At the same time, according to the weight, the possibility of a person suffering from diabetes can be predicted according to the risk factors.

Logical regression is a binary classification problem. If we neglect that the value of Y in binary classification problem is a discrete value (0 or 1), we continue to use linear regression to predict the value of Y. Doing so will result in y not being 0 or 1. Logistic regression uses a function to normalize the y value so that the value of Y is in the interval (0,1). This function is called Logistic function, also known as Sigmoid function. The function formula is as follows:

$$g(z) = \frac{1}{1 + e^{-z}} \tag{17}$$

Logistic function when z approaches infinity,g(z) approaches 1; when Z approaches infinity, g (z) approaches 0. Logical regression is essentially a linear regression. It only adds a layer of function mapping to the mapping from feature to result, that is, first the feature is linear summation, and then the most hypothetical function is predicted by using function g (z). g(z) can map continuous values between 0 and 1. When the expression of linear regression model is introduced into g (z), the expression of logical regression is obtained.

$$h_\theta(x) = g\left(\theta^T x\right) = \frac{1}{1 + e^{-\theta^T x}} \tag{18}$$

Now we normalize the value of y to (0,1) by logistic function. The value of Y has a special meaning. It represents the probability of taking the result of 1. Therefore, the probability of classifying the result of input x into category 1 and category 0 is respectively:

$$P\left(y = 1 \middle| x; \theta\right) = h_\theta(x) \tag{19}$$

$$P\left(y = 0 \middle| x; \theta\right) = 1 - h_\theta(x) \tag{20}$$

Combining the above expressions is:

$$P\left(y|x;\theta\right)=\left(h_\theta\left(x\right)\right)^y\left(1-h_\theta\left(x\right)\right)^{1-y} \tag{21}$$

The expression of logistic regression is obtained. Next, the likelihood function is constructed, and then the maximum likelihood estimation is obtained. Finally, the iteration renewal expression of theta is deduced.

$$\theta_j:=\theta_j+a\left(y^{(i)}-h_\theta\left(x^{(i)}\right)\right)x_j^{(i)} \tag{22}$$

3. Experiments

3.1. Data and data sources

The data in this paper comes from the IMA Indian Diabetes data set in the University of California, Irvine (UCI) machine learning database, which contains 768 data items. The 768 subjects were ordinary residents from Arizona. Because of the high incidence of diabetes among the local population, the National Institute of Diabetes, Digestive and Kidney Diseases will conduct a continuous survey of the population in this area. Each data item contains eight basic attributes, all of which are numerical. Six of them are quantitative and continuous.

The specific attributes are as follows: (1) Number of pregnancies, including abortion before the test; (2) The plasma glucose concentration 2 hours after oral glucose tolerance test; (3) diastolic blood pressure (mm Hg); (4) Skin fold thickness of triceps brachii (mm); (5) 2-hour serum insulin (mu U/ml); (6) Body mass index (kg/m2); (7) Family function value of diabetes mellitus; (8) Age; (9) Category value, 1 for diabetes, 0 for non-diabetes.

3.2. Data preprocessing

Because the data contains missing values, duplicate data, wrong data and other issues, it can not be used directly. In order to effectively reduce the impact of data noise on the accuracy of prediction, data preprocessing and data cleaning should be carried out. The experimental data contains a certain number of missing data, such as diastolic blood pressure data. As shown in the figure 1, the missing data is 0, and the diastolic blood pressure data contains 35 missing data. The median can avoid the influence of extreme values, so this data is cleaned by using median to fill missing values.

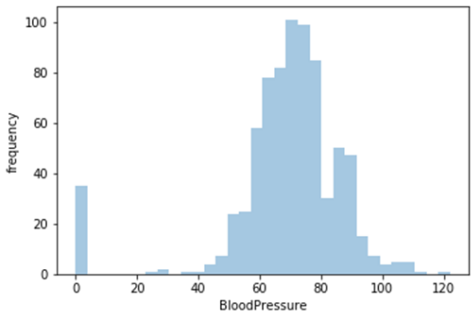

Figure 1. Blood pressure data

3.3. Experimental process

The pre-processed data are divided into training set and test set, 70% of which are training set and 30% are test set. The binary logistic regression model and XGboost gradient lifting model were established by Python 3.6. This experiment included 768 subjects, of which 268 were diabetic, accounting for 34.89%.

The accuracy rate is used to evaluate the prediction effect of the model. The definition formula is as follows. The accuracy rate is equal to the correct number of samples / the total number of samples * 100%. The XGBoost model predicts the decimal number between 0 and 1, binarizes the decimal number above 0.5 and below 1,0.5, then compares the predicted value with the actual value, and calculates the correct rate.

Parametric adjustment process: Logistic regression, regularization parameter C = 1 (default value) model in the training set correctness rate of 78%, in the test set correctness rate of 77%. When the regularization parameter C is set to 100, the accuracy of the model in the training set is slightly improved by 78.5%, but it slightly decreased by 76.6% in the test set. It shows that less regularization and more complex models are not necessarily better than default parameter models. Therefore, we choose the default value C = 1. Under the condition of default parameters, the accuracy of XGBoost is 91.7% in the training set and 82.2% in the test set, which may be over-fitting. In order to reduce over-fitting, pruning is carried out by limiting the maximum depth and reducing the learning rate. The maximum depth is 5, the learning rate is 0.05, the training set accuracy is 85% and the test set accuracy is 83%. The accuracy of training set is reduced.

3.4. Characteristic Importance Assessment Methods

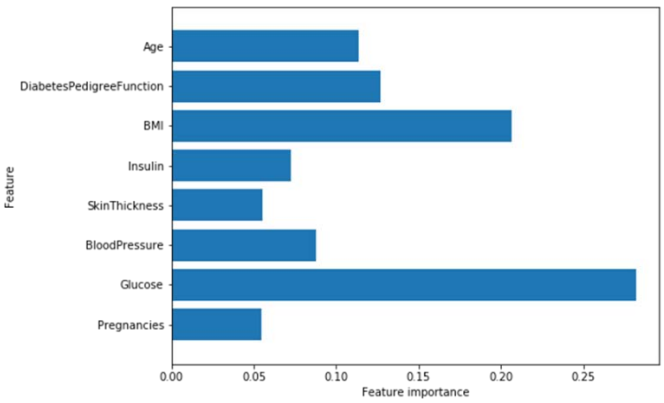

Figure 2. Feature importance

The contribution of each characteristic variable to the model can be judged by XGBoost model, and which characteristic variables have more significant impact on the risk of diabetes mellitus can be judged. As shown in the figure 2, plasma glucose concentration ranks first two hours after oral glucose tolerance test, which is also consistent with clinical practice. The clinical use of glucose tolerance test (OGTT), and combined with the results of blood sugar level test to make the final diagnosis. BMI ranks second, and body mass index (BMI) can reflect the overall body fat. Obese people are prone to diabetes, especially type 2 diabetes. Obese people should pay more attention to the prevention of diabetes and eat less food containing more sugar. The function value of diabetes family ranks third. Diabetes has a certain relationship with heredity. People with diabetes history in family should pay more attention to prevention.

4. Conclusion

In terms of model accuracy, XGBoost has better advantages. From the model, we can see the importance of fasting blood sugar, family history of diabetes mellitus, past history of cardiovascular and cerebrovascular diseases to the onset of type 2 diabetes mellitus. Through this study, we can predict diabetes mellitus by glucose tolerance test, body mass index, family function value of diabetes mellitus and avoid many complicated hospitals. The inspection procedure saves medical resources and provides guidance for diagnosis and prediction in some remote areas. However, due to the influence of data and region, follow-up work still needs a lot of data and investigation data, and reasonable and objective evaluation model.

References

[1] Wang Yao. Diabetes Health Data Analysis Method and Application [D]. Master's Degree Thesis, Harbin University of Technology, 2017: 36-45.

[2] International Diabetes Federation (IDF) DIABETES ATLAS (Eighth Edition), 2017. I.S. Jacobs

and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[3] Chen Yu, Zong Huijuan and Li Wei.Study on risk factors and risk prediction model of type 2 diabetes [J].Journal of Kunming University of Technology (Natural Science Edition), 2018,43(02): 60-64+70.

[4] Su Ping, Yang Yachao, Yang Yang Yang, et al. Prediction model of risk of type 2 diabetes in health management population [J]. Journal of Shandong University (Medical Edition), 2017, 55 (6): 82-86..

[5] Ogunleye Adeola Azeez,Qing-Guo Wang. XGBoost Model for Chronic Kidney Disease Diagnosis.[J]. IEEE/ACM transactions on computational biology and bioinformatics,2019..

[6] Luo Hao, Han Ruizhu. Logistic credit scoring model based on adaptive LASSO variable selection [J]. Quotient, 2016 (04): 161-162.