DOI: 10.53469/jtpes.2024.04(01).08

Diabetes Risk Analysis based on Machine Learning LASSO Regression Model

Sihao Wang[1,*], Yizhi Chen[2], Zhengrong Cui[3], Luqi Lin[4], Yanqi Zong[5]

[1]Mathematics, Southern Methodist University, Dallas, TX
[2]Information Studies, Trine University, Allen Park, MI, USA
[3]Software Engineering, Northeastern University, Shanghai, China
[4]Software Engineering, SunYat-sen University, Shanghai, China
[5]Information Studies, Trine University, PhoenixAZ, USA
Correspondence Author, sihaow@smu.edu

Abstract: ***With the continuous application of artificial intelligence in the field of medical research, machine learning has been widely used to solve many complex problems in the medical field. However, there are many risk factors affecting the development of diabetes, which are far more complex than the traditional disease prediction model. LASSO (Least Absolute Shrinkage and Selection Operator) regression model is an intelligent machine learning algorithm, which has the advantages of strong anti-overfitting ability and is not susceptible to collinearity between variables. The prediction model based on LASSO regression algorithm is conducive to finding and identifying different models and nonlinear relationships among multi-dimensional factors, so as to accurately predict the incidence of diabetes. In disease detection, the role of LASSO regression models is to help identify key characteristic variables associated with disease, thereby improving the predictive accuracy and interpretability of the models. By reducing uncorrelated variables, LASSO is able to process high-dimensional data more efficiently, reducing the risk of overfitting, and improving the model's ability to generalize. Based on these advantages of LASSO algorithm, this paper analyzes the risk detection of diabetes on the basis of machine learning.***

Keywords: Diabetes mellitus; LASSO regression model; Risk detection; Machine learning.

1. INTRODUCTION

Diabetes is a common chronic metabolic disease, the causes of which include genetic and environmental factors and many other complex factors, diabetes patients have high blood sugar levels, and may suffer from serious complications, such as cardiovascular disease, kidney disease, eye disease and so on. Worldwide, millions of people are diagnosed with diabetes each year, and the number continues to climb[1]. Chronic diseases such as diabetes not only affect people's quality of life, but also bring a heavy burden to the national health expenditure. Ai technology can now predict blood sugar levels in people with diabetes, a very common disease that more than 100 million Americans currently have diabetes or prediabetes, and about 1.4 million new cases are diagnosed each year, according to a recent report from the Centers for Disease Control and Prevention (CDC). Although medical advances have made it easier to treat diabetes, patients who remain stuck in high levels of blood sugar are still at risk of coma or even death. Earlier and more accurate identification and intervention of diabetes can effectively control its incidence, improve people's happiness of life, and reduce the medical burden of the country. In the context of the continuous improvement of information technology and data processing capabilities, the application of machine learning technology to the medical field has become a hot trend at present. Machine learning technology can establish predictive models based on a large number of diabetes patients' data, so as to accurately help improve the prevention and treatment of diabetes[2].

Diabetes is also a chronic metabolic disease characterized by high blood sugar levels due to insufficient insulin production or resistance to insulin action. Diabetes is associated with a variety of complications, including cardiovascular disease, kidney disease, neuropathy, and retinopathy. Based on the diabetes data set released by Alibaba Cloud Tianchi Competition, this paper first reviews the current situation of diabetes prediction at home and abroad and the research status of LOSSA regression model algorithm, expounds the basic theories of feature selection, LOSSA regression model and model parameter optimization, and then preprocesses the data: The abnormal points and data with more missing points are eliminated, the features with fewer missing points are filled with the mean value, the discrete variables are encoded, and the data with different dimensions are normalized. To analyze the risk factors of diabetes patients and explain the innovation and future prospects of machine learning in the field of disease detection.

2. RELATED WORK

The continuous progress of machine learning technology makes its application in the medical field a hot trend at present. In this context, the construction of accurate health prediction models based on machine learning algorithms has attracted much attention. With this

In terms of improving prediction accuracy, optimizing machine learning model parameters has become an important direction of current research. The purpose of this paper is to use LOSSA regression model algorithm to construct diabetes prediction model and risk.

2.1 LASSO Regression Model

LASSO regression is a variable screening method with high model stability. The coefficient is continuously compressed by introducing a penalty term into the model estimation. To achieve the purpose of simplifying the model, while effectively dealing with overfitting and multicollinearity problems[3]. In the regression coefficient trajectory diagram, each curve represents the change trajectory of each independent variable coefficient. Using the sum of the absolute value of the regression coefficients as a penalty function, the regression coefficients of the characteristic variables in the model with little correlation with diabetes outcome were compressed to 0 to achieve variable screening. In the cross-validation curve, λ with the smallest deviation is selected. At this value, the LASSO regression model has the best fitting effect.

$$w = arg\min_{w} \sum_{i=1}^{N} (y_i - w^T x_i) = (X^T X)^{-1} X^T y \tag{1}$$

Lasso regression model is a linear model used to estimate sparse parameters, especially for parameter reduction. For this reason, Lasso regression model is widely used in compressed sensing.

To understand Lasso regression, first the general linear model is Y = β0 + β1X1 +... + βnXn + e, the best fit attempts to minimize the residual sum of squares (RSS). RSS is the sum of squares of the difference between the actual number and the estimated number and can be expressed as e12 + e22 +... + en2. We can use regularization to add a new parameter to the RSS minimization process, called contraction penalty term. This penalty term contains the normalized results of the λ and β coefficients and weights.

2.2 LASSO Model Disease Prediction Principle

When the generalized linear model is established by Lasso regression, the response variables can include: one-dimensional continuous dependent variable; Multidimensional continuous dependent variable; Non-negative frequency dependent variable; Binary discrete dependent variable; Multivariate discrete variable.

In addition, regardless of whether the dependent variable is continuous or discrete, lasso can handle it. In general, lasso has extremely low requirements for data, so it is widely used. The complexity of lasso is controlled by λ, and the greater the λ, the greater the penalty on the linear model with more variables, and the final result is a model with fewer variables.

LASSO regression is characterized by Variable Selection and Regularization while fitting generalized linear models. Therefore, regardless of whether the target dependent/response varaible is continuous, binary or multivariate discrete, it can be modeled and predicted using LASSO regression[3-5]. Variable screening here means not putting all variables into the model for fitting, but selectively putting variables into the model to get better performance parameters. Complexity adjustment refers to controlling the complexity of the model through a series of parameters to avoid Overfitting. For linear models, complexity is directly related to the number of variables in the model, and the more variables, the higher the complexity of the model. More variables can often give a seemingly better model when fitting, but at the same time there is the danger of overfitting. At this point, if you use completely new data to validate the model, it is usually very ineffective[6-7]. In general, when the number of variables is much larger than the number of data points, or when a discrete variable has too many unique values, it is possible to overfit. The degree of LASSO regression complexity adjustment is controlled by the parameter λ, and the greater the λ, the greater the penalty on the linear model with more variables, so that a model with fewer variables is finally obtained. LASSO regression and Ridge regression belong to a family of generalized linear

models called Elastic Net. Models of this family have a second parameter, α, in addition to the same parameter λ, which controls the behaviour of models on highly correlated data. LASSO regression α=1, Ridge regression α=0, and the general Elastic Net model 0<α<1.

3. METHODOLOGY

3.1 Data Collection and Preprocessing

Collect data related to diabetes risk, such as age, weight, family medical history, lifestyle habits, etc. This step also includes data cleaning, such as dealing with missing values, outliers, etc.

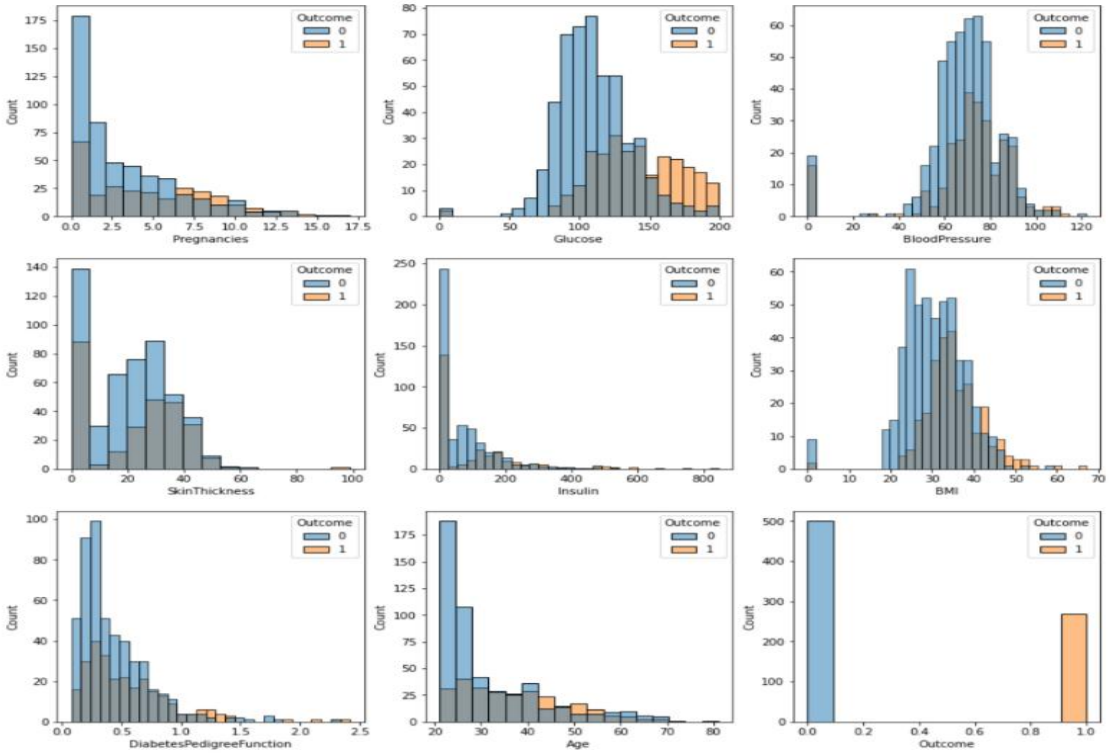

Figure 1: Data feature map

3.2 Feature Selection

The features were analyzed using LASSO regression model. LASSO achieves feature selection and complexity control by adding a L1 regularization term to the loss function. This regularization term penalizes the complexity of the model (i.e. the number of non-zero coefficients).

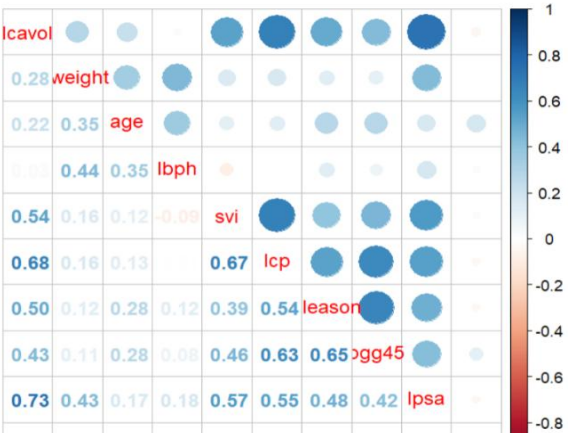

Figure 2: LASSO regression model was used to analyze the features

When using the LASSO regression model for feature selection, there are several key points to note:

1) Normalized/Normalized data: [8]LASSO is sensitive to features in different numerical ranges, so before applying LASSO, it is usually necessary to standardize or normalize features to ensure that all features are in the same order of magnitude.

2) Choose the right regularization parameter (λ) : The choice of the regularization parameter λ is crucial, it controls the strength of the penalty. A large λ may result in too many features being excluded (underfitting), while a small λ may result in too much model complexity (overfitting). The optimal λ value is usually selected by cross-validation and other methods.

3) Understand the sparsity of feature selection[9-11]: LASSO tends to produce sparse models, i.e. it tends to make some coefficients exactly zero. This means that some features may not be considered by the model at all. This is a double-edged sword that can either enhance the interpretability of the model or cause important features to be ignored.

3.3 LASSO Regression Model Construction

The LASSO regression algorithm can be simplified to the following form:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (2)$$

Where y i is the response variable, xij is the characteristic variable, βj is the model coefficient, λ is the regularization parameter, n is the sample number, and p is the characteristic number. The core of this algorithm is to determine the value of the coefficient βj by minimizing the loss function (i.e. the sum of the squares of the prediction error plus the regularization term)[12-15]. The regularization term's purpose is to prevent the model from overfitting and to enable feature selection to some extent (because some coefficients will become zero).

3.4 LASSO Model Training

From the above feature selection, we can see that the train column can layer the training set and test set, so we can separate the data and test it after the model is established. This data pattern is very thoughtful. Next, we will cut the data and find that the data is cut in a 7:3 way, so it is necessary to conduct model detection and training on the selected feature data.

The LASSO model is trained on the training set using selected features and parameters. Where the format file required for glmnet to read is constructed, x is a numerical matrix, y is a response variable, where y is a continuous independent variable, and modeling and cross-validation (cv.glmnet) are as follows:

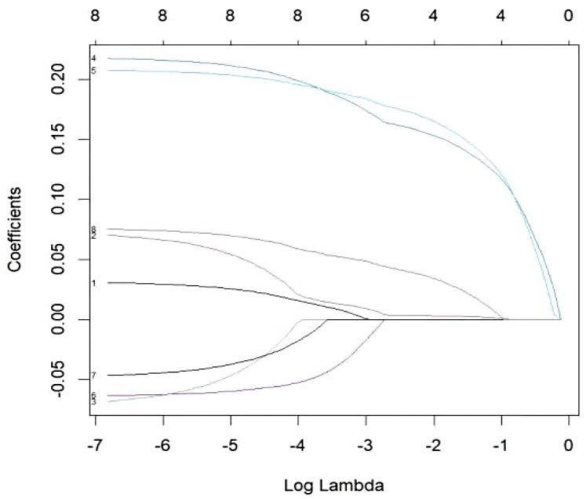

Figure 3: Patient disease prediction model training results

As can be seen from the figure, when three parameters tend to 0 with the change of alpha value, the coefficient

tends to 0 in linear regression is equivalent to having no influence on the corresponding variables, that is, it is meaningless. Therefore, the alpha value when we select six parameters is 0.045 according to the result. In this case, age, lcp, pgg45 can be removed to achieve the optimization of the model[16].

We will cross-validate the above models, and outline the concept of cross-validation:

Cross-validation is mainly used in modeling applications, such as PCR and PLS regression modeling. In the given modeling samples, take most of the samples to build the model, leave a small part of the sample to use the newly established model to forecast, and find the prediction error of this small part of the sample, record their squares and sums. As follows:

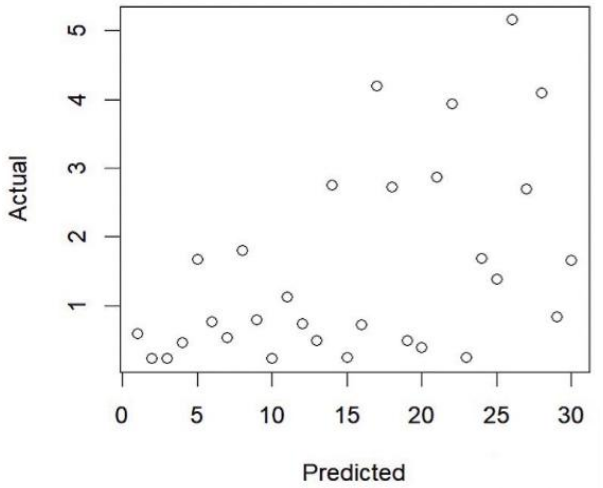

Figure 4: Cross-validation model

The results of decision curve analysis show that the decision curve of the random forest prediction model is higher than the all-positive line and all-negative line in the range of threshold probability 0 ~ 0.85, which indicates that the model has good clinical net return and prediction value. And the effectiveness evaluation results of the two methods are similar, indicating that the model has high stability. The final AUC value of the ten-fold cross-validation is 84.8%, which indicates that the LASSO prediction model has good classification and identification ability.

This study is a study on the detection and validation of diabetes disease prediction model based on database. LASSO regression is applied to the analysis of diabetes risk factors[17-18]. The results of LASSO regression model are relatively effective through data demonstration. However, the sample size included in the study was small and only internal verification was carried out, which affected the generalization ability of the model to a certain extent. In future studies, we will seek opportunities for large samples and external population verification, so as to further improve the prediction performance and generalization ability of the model.

4. CONCLUSION

Diabetes is a common chronic disease, its high incidence and serious consequences have caused more and more people to close note. With the continuous progress of machine learning technology, more and more scholars are applying machine learning technology to the medical field to help people identify and intervene in diseases earlier and more accurately[19]. The purpose of this study is to explore the LASSO regression model for predicting blood glucose concentration through physical examination information and optimize the LASSO model based on the diabetes data set of Alibaba Cloud Tianchi Competition.

The main contribution of regression models in disease prediction is their ability to analyze various biological, clinical and environmental data of patients to build mathematical models to predict whether a patient has a certain disease or the degree of their disease risk. These models are based on historical data, and by learning patterns and associations in the data, they can help healthcare professionals make early disease diagnosis, risk assessment, and personalized treatment decisions[20]. Regression models usually use a variety of statistical and machine learning algorithms, such as linear regression, logistic regression, support vector machines, random forests, etc., to make predictions by fitting relationships in the data, thus providing reliable tools and methods for disease prediction. The

implementation of the algorithm usually includes steps such as data collection, feature engineering, model training, and evaluation to ensure the accuracy and effectiveness of the model in disease prediction.

ACKNOWLEDGEMENT

In the process of writing this article, I am deeply impressed by Shen, Z., Wei, K., Zang, H., Li, L., & Wang, G. In their work, The Application of Artificial Intelligence to The Bayesian Model Algorithm for Combining Genome Data, published in 2023, The theory and method proposed in the Academic Journal of Science and Technology, Vol. 8, No. 3, pp. 132-135) are inspired. The paper provides valuable insights and algorithmic foundations for understanding and applying machine learning, particularly LASSO regression models, to diabetes risk prediction. I'm Shen, Z., Wei, k., Zang, H., Li, L., & Wang, G. Their research results have an important impact on the conception and implementation of this paper. I also thank them for their contributions to academic progress in related fields.

REFERENCES

[1] Friedman, J., Hastie, T. and Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent (2010), Journal of Statistical Software, 2008.Vol. 33(1): 1-22,
[2] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent, Journal of Statistical Software, 2011,Vol. 39(5): 1-13.
[3] Xinyu Zhao, et al. “Effective Combination of 3D-DenseNet’s Artificial Intelligence Technology and Gallbladder Cancer Diagnosis Model”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 81-84, https://doi.org/10.54097/iMKyFavE.
[4] Shulin Li, et al. “Application Analysis of AI Technology Combined With Spiral CT Scanning in Early Lung Cancer Screening”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 52-55, https://doi.org/10.54097/LAwfJzEA.
[5] Liu, Bo & Zhao, Xinyu & Hu, Hao & Lin, Qunwei & Huang, Jiaxin. . Detection of Esophageal Cancer Lesions Based on CBAM Faster R-CNN. Journal of Theory and Practice of Engineering Science. 2023,3: 36-42. 10.53469/jtpes.2023.03(12).06.
[6] Yu, Liqiang, et al. “Research on Machine Learning With Algorithms and Development”. Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 7-14, doi:10.53469/jtpes.2023.03(12).02.
[7] Xin, Q., He, Y., Pan, Y., Wang, Y., & Du, S. . The implementation of an AI-driven advertising push system based on a NLP algorithm. International Journal of Computer Science and Information Technology, 2023,1(1): 30-37.0
[8] Zhou, H., Lou, Y., Xiong, J., Wang, Y., & Liu, Y. . Improvement of Deep Learning Model for Gastrointestinal Tract Segmentation Surgery. Frontiers in Computing and Intelligent Systems, 2023,6(1): 103-106.6
[9] Implementation of an AI-based MRD Evaluation and Prediction Model for Multiple Myeloma. . Frontiers in Computing and Intelligent Systems, 2024,6(3): 127-131. https://doi.org/10.54097/zJ4MnbWW.
[10] Zhang, Q., Cai, G., Cai, M., Qian, J., & Song, T. . Deep Learning Model Aids Breast Cancer Detection. Frontiers in Computing and Intelligent Systems, 2023,6(1): 99-102.3
[11] Xu, J., Pan, L., Zeng, Q., Sun, W., & Wan, W. . Based on TPUGRAPHS Predicting Model Runtimes Using Graph Neural Networks. Frontiers in Computing and Intelligent Systems, 2023,6(1): 66-69.7
[12] Wan,Weixiang, et al. "Development and Evaluation of Intelligent Medical Decision Support Systems." Academic Journal of Science and Technology 8.2 (2023): 22-25.
[13] Tian, M., Shen, Z., Wu, X., Wei, K., & Liu, Y. . The Application of Artificial Intelligence in Medical Diagnostics: A New Frontier. Academic Journal of Science and Technology, 2023,8(2): 57-61.7
[14] Shen, Z., Wei, K., Zang, H., Li, L., & Wang, G. . The Application of Artificial Intelligence to The Bayesian Model Algorithm for Combining Genome Data. Academic Journal of Science and Technology, 2023,8(3): 132-135.2
[15] Zheng He, et al. “The Importance of AI Algorithm Combined With Tunable LCST Smart Polymers in Biomedical Applications”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 92-95, https://doi.org/10.54097/d30EoLHw.
[16] Prediction of Atmospheric Carbon Dioxide Radiative Transfer Model based on Machine Learning. . Frontiers in Computing and Intelligent Systems, 2024,6(3): 132-136. https://doi.org/10.54097/ObMPjw5n
[17] Liu, Y., Duan, S., Shen, Z., He, Z., & Li, L. . Grasp and Inspection of Mechanical Parts based on Visual Image Recognition Technology. Journal of Theory and Practice of Engineering Science, 2023,3(12): 22-28.1

[18] Xinyu Zhao, et al. “Effective Combination of 3D-DenseNet’s Artificial Intelligence Technology and Gallbladder Cancer Diagnosis Model”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 81-84, https://doi.org/10.54097/iMKyFavE.
[19] Liu, B. . Based on intelligent advertising recommendation and abnormal advertising monitoring system in the field of machine learning. International Journal of Computer Science and Information Technology, 2023,1(1): 17-23.
[20] Pan, Linying, et al. “Research Progress of Diabetic Disease Prediction Model in Deep Learning”. Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 15-21, doi:10.53469/jtpes.2023.03(12).03.