Research of Prediction Diabetes Risk Using Logistic Regression Models

Gujie Li

Northeastern University, Coll of Soc Sci & Humanities, Boston, MA, 02115, USA

li.guj@northeastern.edu

Abstract. Diabetes has become a major global health concern, with its prevalence steadily rising. Early prediction and prevention are crucial to reducing the disease burden. This study explored the relationship between HbA1c level, age, smoking history, BMI, and diabetes risk by analyzing electronic health records (EHR) of more than 10,000 individuals and logistic regression models. By establishing a prediction model, the impact of these variables on the likelihood of an individual developing diabetes was evaluated. The results showed that the overall accuracy of the logistic regression model reached 89%, and the ROC-AUC score was as high as 0.9624, showing excellent discrimination between diabetic and non-diabetic cases. Among them, HbA1c level (coefficient 2.49), blood glucose concentration (coefficient 1.3), and age (coefficient 1.16) were confirmed to be key predictors for diabetes diagnosis, especially HbA1c level (coefficient 2.49) was the most influential factor. The study also discussed the potential limitations of the model performance and future improvement directions.

Keywords: Diagnosing diabetes, HbA1c, BMI, Diabetes prediction, Logistic Regression.

1. Introduction

Diabetes mellitus is a chronic metabolic disease characterized by persistent hyperglycemia, which poses a significant health risk and economic burden worldwide. Early detection and prediction of diabetes are crucial for effective management and prevention of its complications. The disease is influenced by a combination of genetic, behavioral, and physiological factors, so it is critical to identify and quantify the relationship between these variables and diabetes risk.

According to the research of some scholars, it was found that diabetes is caused by many factors. Davidson et al. used a multivariate linear regression model to control for confounding variables and analyzed the relationship between HbA1c levels and age, race/ethnicity, gender, and blood glucose concentration using NHANES III data. They found that age and race/ethnicity had a significant effect on HbA1c. The HbA1c level increases by 0.07% every 10 years [1]. Eliasson examined many clinical and experimental studies and investigations that strongly demonstrated a significant association between smoking, diabetes development, glycemic control, and diabetic complications (microvascular and macrovascular). Most of these effects are likely caused by nicotine, possibly in combination with other substances in cigarette smoke [2]. Narayan et al. found that being overweight, especially obesity significantly increases the risk of diabetes at a young age [3]. Similar findings were found in the study by Gregg et al., which showed that the proportion of people with diagnosed diabetes in the United States with higher BMI increased significantly from 41% to 83% between 1976-1980 and 1999-2000 [4]. In his study, Gale found that among European people aged 15 to 40, it was generally observed that the number of males suffering from the disease exceeded that of females, with a male-to-female ratio of approximately 3:2 [5]. These researchers' articles have proposed many factors that affect diabetes. Since most of these articles were published around 2000, people's lifestyles, eating habits, and environmental factors have changed over time. These changes Affects the incidence and development of diabetes.

This study aimed to predict the risk of diabetes using a logistic regression model. Logistic regression is a statistical method widely used for classification problems. It estimates the nonlinear relationship between independent variables and categorical outcomes (such as diabetes) by maximizing the likelihood function. The model uses HbA1c level, blood sugar level, age, heart

disease history, smoking history, and other health indicators as input variables to quantify the impact of these factors on the probability of disease, thereby identifying the most important predictive factors and providing data for the implementation of targeted intervention measures. support.

2. Methods

2.1. Dataset and Preprocessing

First, this paper selects a set of personal data of more than 10,000 individuals on Kaggle, including whether they have diabetes, as well as characteristics such as gender, age, BMI, smoking history, and blood sugar level. This dataset is based on electronic health records (EHR) from multiple medical service providers, and the consistency, relevance, and integrity of the data are ensured through cleaning and preprocessing steps. In addition, the widespread application of EHR is closely related to real medical scenarios, providing reliable data support for the development of efficient diabetes prediction models.

Data preprocessing involves handling potential missing values. For categorical variables, missing values are imputed with the most common category (mode), while for numerical variables, missing values are imputed with the mean. Categorical variables are converted to numerical format using one-hot encoding, and one category is removed to avoid multicollinearity.

After preprocessing, the data is split into a training dataset and a test dataset in an 80/20 ratio. To address the class imbalance in the target variable, the synthetic minority oversampling technique (SMOTE) is applied to the training dataset to ensure a balanced representation of diabetic and non-diabetic cases.

2.2. Experimental Design

After data preparation, a Logistic Regression model is selected to predict the likelihood of having diabetes. Logistic regression is a widely used statistical method for classification problems that estimate the nonlinear relationship between independent variables and classification outcomes (such as whether or not having diabetes) by maximizing the likelihood function.

The model is trained using the oversampled training dataset and evaluated on the test dataset using three metrics. The first metric is the confusion matrix, which evaluates the classification performance by measuring true positives, false positives, true negatives, and false negatives. The second metric is the classification report, which provides precision, recall, F1 score, and support for each class. Finally, the model is evaluated using the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) score was used to evaluate the ability of the model to effectively distinguish between diabetic and non-diabetic cases.

3. Results

The logistic regression model provided insights into factors influencing diabetes based on the coefficients associated with each characteristic. The magnitude and direction of these coefficients reflect the relative importance and impact of each factor on the likelihood of a diabetes diagnosis. As shown in Fig. 1, the HbA1c level had the highest coefficient of 2.49 in the logistic regression model, which was the most influential factor, and a higher level significantly increased the risk of diabetes. Blood sugar levels, with a coefficient of 1.3, were also a significant predictor. The coefficient for age is 1.16, which indicates that older age is more likely to develop diabetes. The analysis confirmed that A1C levels, blood sugar levels and age were the most critical factors, having a significant impact on the likelihood of developing diabetes. Lifestyle factors such as BMI and smoking history also play an important role, although their effects are less pronounced. Notably, people with no history of smoking had a lower risk compared with people with other smoking statuses.

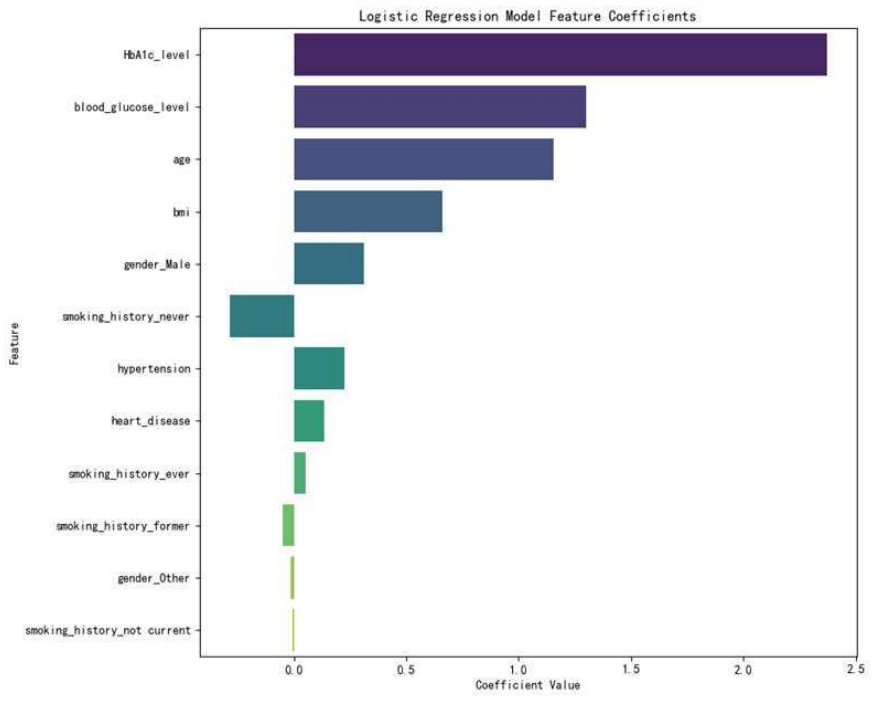

Fig. 1 Importance of features in the logistic regression model (Photo/Picture credit: Original).

According to the results in Fig. 2, in the evaluation model, 16,245 cases were correctly classified as not having diabetes and 1,509 cases were correctly classified as having diabetes. 2,055 cases were incorrectly classified as having diabetes. 191 cases were incorrectly classified as not having diabetes. The overall accuracy of the model reached 89%. The ROC curve shows strong model performance with an AUC score of 0.9624, indicating excellent discrimination between diabetic and non-diabetic cases.

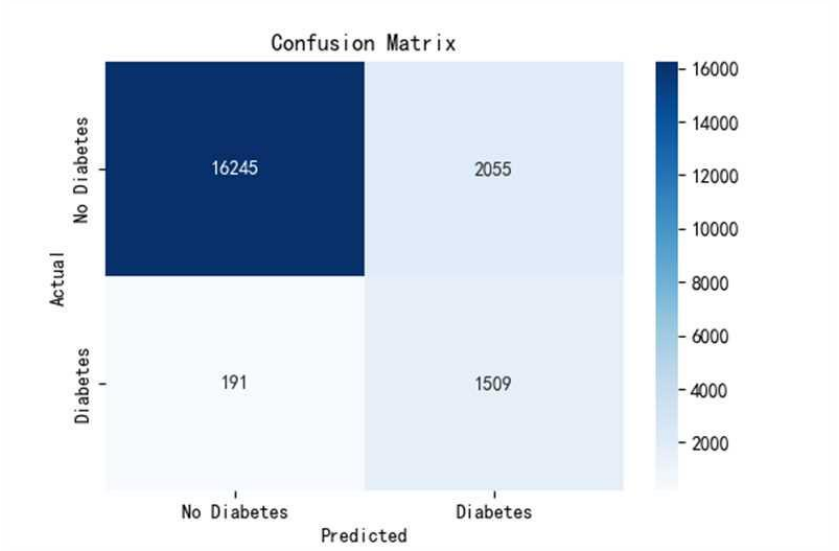

Fig. 2 Confusion matrix of the model (Photo/Picture credit: Original).

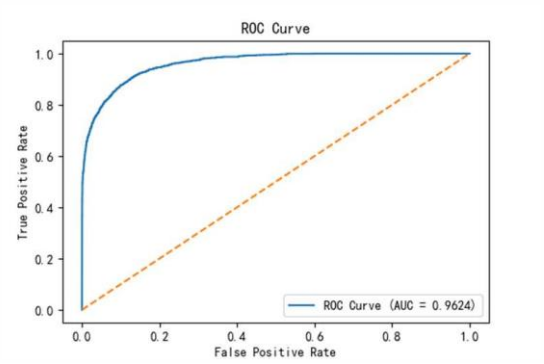

Fig. 3 ROC curve of the model, with an AUC value of 0.9624 (Photo/Picture credit: Original).

The results in Fig. 3 show that the logistic regression model provides a reliable framework for predicting diabetes, in which specific features such as HbA1c level and blood glucose play a key role in diagnosis. However, the accuracy of the diabetes category is relatively low, indicating that there is room for improvement, which may need to be achieved through additional data or alternative modeling techniques. Although the logistic regression model showed high accuracy (89%) and good discrimination ability (ROC AUC = 0.9624) in predicting diabetes, there are still several defects and room for improvement. The training data was oversampled by SMOTE technology, but in the test set, the number of non-diabetes categories (0) was still much higher than that of diabetes categories (1), which may cause the model to tend to predict the majority category more accurately. From the classification report, it can be seen that the model has a low accuracy (42%) in predicting the diabetes category (1), which indicates that the model is prone to misclassifying some diabetic patients as non-diabetic.

4. Conclusion

The results of this study show that HbA1c level (coefficient 2.49), blood glucose level (coefficient 1.3), and age (coefficient 1.16) are the most important factors affecting the risk of diabetes. In addition, BMI and smoking history also have some influence on diagnosis, but the effect is weaker. However, the classification report shows that the model has a low accuracy (42%) in predicting the diabetes category, and there is still a classification bias of the diabetes category in the test dataset, which indicates that there is still room for improvement in the model in dealing with imbalanced data and improving the accuracy of diabetes classification.

Through this study, more and more evidence is provided for diabetes prevention strategies, enabling healthcare providers to prioritize high-risk groups and improve early diagnosis. The results will also help to raise awareness of the importance of maintaining a healthy lifestyle and managing modifiable risk factors such as smoking, BMI, and blood glucose levels.

References

[1] Davidson M. B., Schriger D. L. Effect of age and race/ethnicity on HbA1c levels in people without known diabetes mellitus: Implications for the diagnosis of diabetes. Diabetes Research and Clinical Practice, 2010, 87(3): 415-421.

[2] Eliasson B. Cigarette smoking and diabetes. Progress in Cardiovascular Diseases, 2003, 45(5): 405-413.

[3] Narayan K. V., Boyle J. P., Thompson T. J., Gregg E. W., Williamson D. F. Effect of BMI on lifetime risk for diabetes in the US. Diabetes Care, 2007, 30(6): 1562-1566.

[4] Gregg E. W., Cadwell B. L., Cheng Y. J., Cowie C. C., Williams D. E., Geiss L., et al. Trends in the prevalence and ratio of diagnosed to undiagnosed diabetes according to obesity levels in the U.S. Diabetes Care, 2004, 27: 2806-2812.