Dean&Francis
ISSN 2959-6157

Analysis of Diabetes Prediction Models Based on XGBoost and LightGBM

Weihao Zhou

School of Mathematics, Shandong University, Jinan, Shandong, 250100, China
202300091049@mail.sdu.edu.cn

Abstract:

Predicting diabetes more effectively and accurately is becoming more and more important, and this research helps find the best method to predict diabetes. With the development of the economy, the incidence of diabetes is gradually increasing. Research on comparing eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) in diabetes predictions and further comparing improved results remains limited. This study compares XGBoost and LightGBM's performances in diabetes predictions, and gives suggestions on the best algorithm to predict diabetes mellitus. By using a dataset collected from Kaggle, this research uses XGBoost and LightGBM to create prediction models and compares the results to find out which performs better in diabetes predictions. Furthermore, this research optimizes the data used in building models and again compares the second results to excavate deeper information. The test set's accuracy of XGBoost is 97.2% and the test set's precision of XGBoost is 97.2% using the whole data, while the test set's accuracy of LightGBM is 97.3% and the test set's precision of LightGBM is 97.3%. The test set's accuracy of XGBoost is 97.3% and the test set's precision of XGBoost is 97.4% using the optimized data, while the test set's accuracy of LightGBM is 94.9% and the test set's precision of LightGBM is 94.8%. In conclusion, LightGBM has slight advantages compared to XGBoost using the whole data, and on the other hand, XGBoost performs better using the optimized data. Because the two algorithms have different advantages, the selection of the algorithms in diabetes predictions needs to depend on specific circumstances.

Keywords: XGBoost, LightGBM, diabetes predictions

ISSN 2959-6157

1. Introduction

Diabetes is a chronic disease marked by high blood sugar, which is caused by absolute or relative deficiency of insulin secretion and utilization disorders. With the progress of the times and economic development, the increase in high-calorie foods and changes in lifestyle, the incidence of diabetes is rising at an alarming rate. However, if diabetic patients are not diagnosed in time, it will be an extremely heavy burden for society, families and individuals. There is an urgent need for further development and optimization of the screening and diagnostic methods for diabetes.

Based on the previous studies on the characteristics and diagnosis of diabetes, combined with machine learning, it is possible to predict diabetes more quickly [1]. Li et al. compared five machine learning models and the results showed that the Logistic Regression algorithm performed well in diabetes prediction, with an accuracy rate of 76.6% and an precision rate of 74% [2]. However, most machine learning models tend to exhibit a decline in performance when dealing with high-dimensional data, making it difficult to improve the accuracy of the prediction model by adding more physiological feature variables. Similarly, there was also a prediction of diabetes using the K-Nearest-Neighbors (KNN) algorithm, with an accuracy rate of 97.01%, which was excellent [3]. However, the KNN algorithm has low computational efficiency when dealing with large-scale data, often resulting in low performance and low model accuracy. Diabetes prediction models were built using decision trees and support vector machines. The results showed that the accuracy rates of both methods were 77%, and the precision rate of the decision tree was 83%, while that of the support vector machine was slightly lower at 80% [4]. Rufo et al. constructed a diabetes prediction model using LightGBM [5]. The results showed that the accuracy and AUC of LightGBM were both 98.1%, the sensitivity was 99.9%, and the specificity was 96.3% [5]. It can be concluded that compared to other algorithms, LightGBM is more accurate in diabetes prediction [5]. Zhang and Gong developed prediction models for acute liver failure diseases using two algorithms, XGBoost and LightGBM [6]. They obtained satisfactory results [6]. The accuracy rate of XGBoost was 75.014%, while that of LightGBM was 75.811% [6]. Both algorithms had relatively high accuracy rates and the difference was not significant [6]. This provides inspiration for this study. As high-performance gradient boosting tree models, XGBoost and LightGBM can demonstrate greater advantages in diabetes prediction [7]. By using machine learning algorithms to build diabetes prediction models, the prediction can be made more accurate, which can better assist in diabetes diagnosis and avoid delays and errors in the diagnosis process.

This article mainly builds diabetes prediction models using two algorithms, XGBoost and LightGBM, and compares their performance. Using the indicators for diagnosing diabetes and the diabetes patient dataset on the Kaggle platform as the training set and test set, two algorithms, XGBoost and LightGBM, were used to build diabetes prediction models respectively, and then the results were evaluated and compared. This article aims to develop a more optimal diabetes prediction algorithm to assist in the early diagnosis of diabetes.

2. Data and Methodology

2.1 Data Source and Explanation

The dataset used in this article is selected from the diabetes prediction classification section on the Kaggle platform [8]. This dataset is a set of population medical statistics, and it contains 100,000 samples [8]. And in the original data, hypertension, having heart disease and having diabetes were recorded as 1, while not having hypertension, not having heart disease and not having diabetes were recorded as 0. In data analysis, to make the variable diabetes a qualitative item, and the variables hypertension and heart disease quantitative items, “having diabetes” and “not having diabetes” should be changed to yes and no respectively, while hypertension and heart disease do not need any processing. In the model construction process, the first 70% of the data is used as the training set, and the remaining 30% is used as the test set.

2.2 Indicator Selection and Explanation

The dataset indicators are divided into gender, age, hypertension (HTN), heart disease (HD), body mass index (BMI), glycated hemoglobin (HbA1c), blood glucose (BG), and finally, the target indicator of diabetes.

As shown in Table 1, the types of each indicator are presented. If it is a quantitative indicator, it also includes the range of values, average value, standard deviation, median and variance. If it is a qualitative indicator, it mainly explains the different classifications of this indicator.

Table 1. Basic Information of Variables

Variable Name	Variable Explanation	Variable Type	Range	Average Value	Standard Deviation	Median	Variance
gender	gender	qualitative item	“Male” or “Female”	/	/	/	/
age	age	quantitative item	0.08-80	41.886	22.517	43	507.008
HTN	whether have hypertension	qualitative item	1 = yes 0 = no	/	/	/	/
HD	whether have heart disease	qualitative item	1 = yes 0 = no	/	/	/	/
BMI	body mass index	quantitative item	10.01-95.69	27.321	6.637	27.32	44.047
HbA1c	glycated hemoglobin level	quantitative item	3.5-9	5.528	1.071	5.8	1.146
BG	blood glucose level	quantitative item	80-300	138.058	40.708	140	1657.152
diabetes	whether have diabetes	qualitative item	1 = yes 0 = no	/	/	/	/

2.3 Introduction to the Methodology

In this article, BMI, HbA1c and BG are quantitative variables, while gender, HTN, HD and diabetes are qualitative variables. The two machine learning algorithms used in this article, XGBoost and LightGBM, both operate based on the gradient boosting decision tree framework and are capable of effectively handling the nonlinear relationships and mixed types of features in medical data.

This article uses two algorithms to construct diabetes prediction models. The first type of algorithms, XGBoost, is an ensemble learning algorithm based on gradient boosting decision trees. Its core lies in iteratively constructing multiple decision trees to achieve precise predictions. The logic it follows is to correct errors step by step. After the initial tree prediction is completed, each subsequent tree focuses on fitting the errors of the previous model. Finally, the weighted sum of the results of all trees is used as the final output. The second type of algorithms, LightGBM, is an efficient gradient boosting decision tree algorithm. Its core lies in constructing decision trees one by one through iterative processes and accumulating the prediction results to achieve high-precision modeling. The key innovations include: (1) employing a histogram-based algorithm for the discretization of continuous features; (2) applying Gradient-based One-Side Sampling (GOSS) to retain instances with large gradients and randomly sample those with small gradients; and (3) adopting a leaf-wise growth strategy that prioritizes splitting the leaf node with the highest loss reduction, which enhances predictive accuracy while mitigating overfitting.

The process of this article is shown in Fig. 1.

ISSN 2959-6157

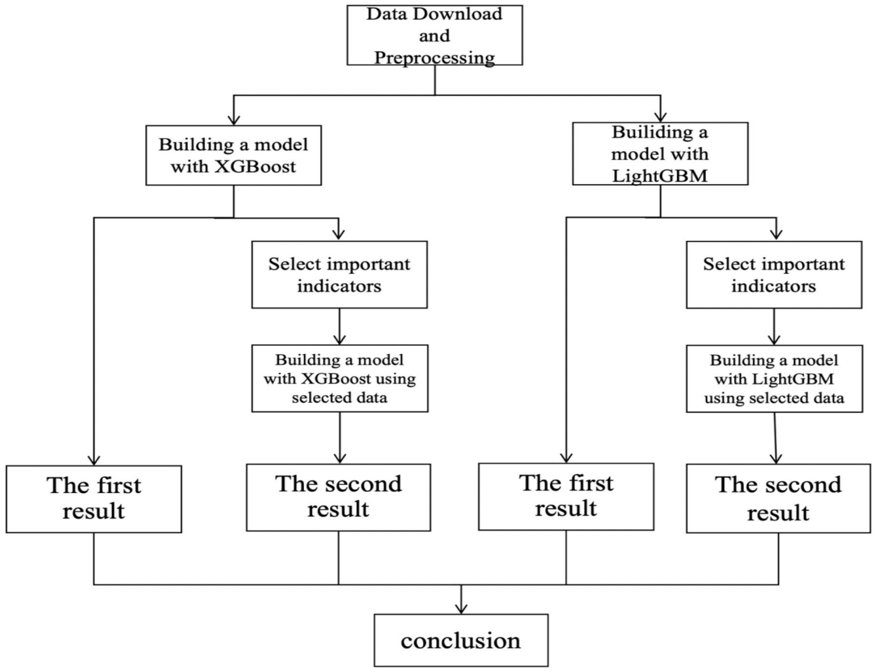

Fig. 1 Flow Chart of This Study (Original)

3. Results and Discussion

3.1 Correlation Analysis

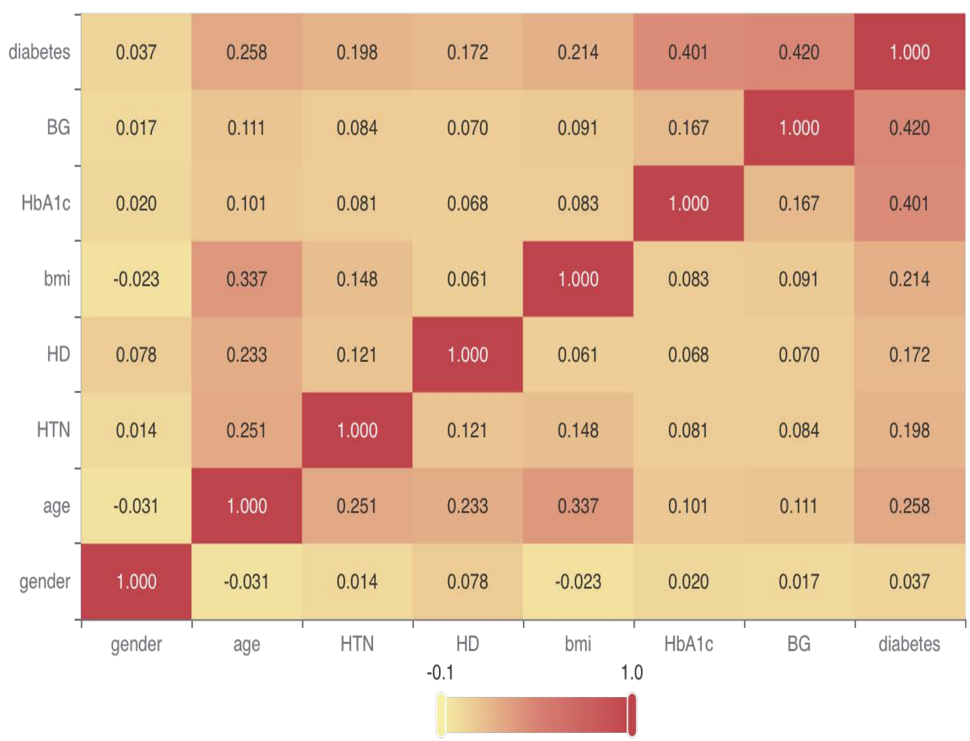

Fig. 2 Heatmap of Correlation Coefficients (Original)

As can be seen from Fig. 2, diabetes is moderately positively correlated with BG and HbA1c; HTN is weakly to moderately positively correlated with age and HD; age is weakly to moderately positively correlated with BMI; and BMI is weakly correlated with HD and HTN.

3.2 Results

Table 2. The First Results

		Accuracy	Recall	Precision	F1	AUC
XGBoost	Training Set	0.974	0.974	0.975	0.972	0.997
	Test Set	0.972	0.972	0.972	0.97	0.996
LightGBM	Training Set	0.973	0.973	0.974	0.971	0.997
	Test Set	0.973	0.973	0.973	0.97	0.996

As can be seen from Table 2, the accuracy and recall rate of the training set for the XGBoost algorithm are 0.974, and the precision rate is 0.975. The accuracy, recall rate and precision rate of the test set are all 0.972. This model demonstrates excellent adaptability to the data and strong generalization ability. The accuracy and recall rate of the training set for the LightGBM algorithm are 0.973, while the precision rate is 0.974. The accuracy, recall rate and precision rate of the test set are all 0.973. It also indicates that this model has excellent adaptability to the data and strong generalization ability.

As can be seen from Table 2, XGBoost and LightGBM both demonstrate excellent performance in both the training set and the test set. They are close to each other in all indicators and are both at a high level. In comparison, XGBoost outperforms LightGBM slightly in terms of accuracy, recall rate, precision rate and F1 score on the training set. For the test set, LightGBM outperforms XGBoost slightly in terms of accuracy, recall rate and precision.

Overall, the values of each indicator in the LightGBM test set are more stable, and the differences between the training set and the test set values are relatively small.

3.3 Method Improvement

Table 3. Proportions of Feature Importance

Feature Names	Feature Importances of XGBoost	Feature Importances of LightGBM
gender	1.30%	3.30%
age	5.00%	23.10%
HTN	2.70%	3.00%
HD	2.80%	2.30%
BMI	2.30%	29.50%
HbA1c	60.80%	17.00%
BG	25.00%	21.90%

From Table 3, it can be concluded that in the prediction model obtained using XGBoost, the importance ratios of the HbA1c indicator and the BG indicator are relatively high. Therefore, it is possible to simplify the prediction model and construct an XGBoost prediction model using only the HbA1c and BG as the two indicators. In the prediction model obtained using LightGBM, the importance ratios of indicators such as age, BMI and BG are relatively high. One possible approach is to construct a LightGBM model using three indicators: age, BMI and BG.

Table 4. The Second Results

		Accuracy	Recall	Precision	F1	AUC
XGBoost	Training Set	0.971	0.971	0.972	0.969	0.991
	Test Set	0.973	0.973	0.974	0.97	0.991
LightGBM	Training Set	0.949	0.949	0.95	0.939	0.988
	Test Set	0.949	0.949	0.948	0.94	0.987

ISSN 2959-6157

As can be seen from Table 4, the values of the new experimental results of XGBoost remain relatively high, being similar to or even higher than the values of the original experiments. However, the amount of data used has been significantly reduced. The new experimental results of LightGBM show high values for each item, but the performance has significantly declined compared to the original experimental results.

3.4 Discussion

In this study, the XGBoost and LightGBM algorithms were used without prior optimization of the hyperparameters. This might result in a large amount of ineffective exploration during model establishment, significantly increasing the consumption of computing resources. If Bayesian optimization is used to optimize the hyperparameters in advance, it can significantly accelerate the convergence to the optimal solution and reduce the computational cost [9].

This paper only uses two gradient boosting tree algorithms, XGBoost and LightGBM, to construct the model for diabetes prediction. By using CATBoost, another gradient boosting tree algorithm closely related to XGBoost and LightGBM, to build a diabetes prediction model, it is possible to more clearly observe the advantages and disadvantages of gradient boosting tree algorithms in diabetes prediction, and at the same time, more accurately determine which algorithm is truly more suitable for diabetes prediction [10].

Furthermore, in this paper, only the XGBoost and LightGBM algorithms were used separately to predict diabetes, and the two algorithms were not combined to make predictions for diabetes. By using soft voting, the two algorithms XGBoost and LightGBM were combined to form a new hybrid algorithm LIGB, which significantly improved its performance in classification. Just as Donepudi et al. demonstrated, when using the LIGB hybrid model to make predictions on the same dataset, it showed a more balanced performance in terms of high accuracy and high recall rate compared to XGBoost and LightGBM [11].

4. Conclusion

This paper, through two machine learning algorithms, XGBoost and LightGBM, constructed two corresponding diabetes prediction models based on machine learning and evaluated and compared the performance of the two models in diabetes prediction. Both algorithms demonstrated excellent performance in diabetes prediction. The prediction accuracy of XGBoost on the test set was 97.2%, while that of LightGBM was 97.3%. LightGBM had a slight edge over XGBoost. However, by analyzing the importance of each indicator in the model's prediction, it can be seen that LightGBM has a greater number of indicators with relatively high importance, while XGBoost only has two indicators with high importance. In the optimized mathematical model, the prediction accuracy of XGBoost and LightGBM remains extremely high, while the amount of data used by XGBoost is much smaller than that used by LightGBM. Therefore, both have their advantages and disadvantages. When choosing a method, it should be determined based on specific circumstances.

The combination of machine learning and diabetes diagnosis has obvious disciplinary advantages, and it has broad development space and application prospects. The research in this paper provides theoretical preparation for the integration of machine learning and diabetes diagnosis. This paper did not optimize the algorithm parameters in advance, which led to excessive ineffective exploration during the modeling process and waste of computing resources. The gradient boosting tree algorithm is selected only in two forms, which makes it difficult to clearly present the advantages and disadvantages of the gradient boosting tree algorithms and the optimal choice of the algorithms when building diabetes prediction models. The effect of the XGBoost and LightGBM hybrid model has not been explored, failing to fully reflect the characteristics of their relationship. The hyperparameters can be optimized using Bayesian optimization. Various algorithms can be compared horizontally, and they can be aggregated to form a new algorithm for exploring performance.

References

[1] Genuth SM, Palmer JP, Nathan DM. Classification and Diagnosis of Diabetes. In: Diabetes in America. 3rd ed. National Institute of Diabetes and Digestive and Kidney Diseases (US), Bethesda (MD); 2018. PMID: 33651569.

[2] Li T., Sun YY., Li XL., et al. Research on the Auxiliary Diagnosis of Diabetes Based on Machine Learning Classification Algorithms. Computer Knowledge and Technology. 2024, 20(10):27-29.

[3] Xiang JJ. Application of Machine Learning in Predicting Diabetes. China Science and Technology Information. 2025, (14):77-80.

[4] Gan FW., Zhang LX. Analysis of Prediction Model for Type 2 Diabetes Based on Machine Learning Classification Algorithm. Digital Technology and Application. 2025, 43(06):217-219.

[5] Rufo DD., Debelee TG., Ibenthal A., Negera WG. Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM). Diagnostics. 2021, 11(9):1714.

[6] Zhang DY., Gong YC. The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute

Liver Failure. in IEEE Access, 2020, 8, 220990-221003.
[7] Chen TQ., Guestrin C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). 2016, Association for Computing Machinery, New York, NY, USA, 785–794.
[8] Mohammed Mustafa Diabetes prediction dataset 2023.01.01 2025.07.22 https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/discussion
[9] Khurshid MR., Manzoor S., Sadiq T., Hussain L., Khan MS., et al. Unveiling diabetes onset: Optimized XGBoost with Bayesian optimization for enhanced prediction. PLOS ONE. 2025, 20(1): e0310218.
[10] Jaiswal S., Gupta P. Ensemble Approach: XGBoost, CATBoost, and LightGBM for Diabetes Mellitus Risk Prediction. 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2022, 1-6.
[11] Donepudi S., Nakka R., Thota KK., Ajmeera M., Praveen SP., Sindhura S. (2025). Enhancing Person-Centric Health Care for Diabetes Prediction: A Comparative Study of LightGBM, XGBoost, and Hybrid LIGB Model. In: Barsocchi, P., Naga Srinivasu, P., Kumar Bhoi, A., Palumbo, F. (eds) Enabling Person-Centric Healthcare Using Ambient Assistive Technology, Volume 2. Studies in Computational Intelligence, vol 1191. Springer, Cham.