Dynamic Financial Distress Prediction Using Combined LASSO and GBDT Algorithms

Ziyi Jiao
School of Economics and Management, Henan Polytechnic Institute, Nanyang 473000, China
E-mail: m15237711330@163.com

Keywords: financial distress forecasting, lasso algorithm, gbdt algorithm, concept drift, similarity weighting

Received: June 25, 2024

With the global economy in a downward cycle under the influence of the epidemic, companies are facing a crisis in their business and financial conditions, and most companies are more likely to be in financial distress in a poor economic environment. The existence of concept drift problem makes the actual prediction of financial distress prediction poor or can only solve limited types of concept drift. Most existing research on financial distress prediction methods use machine learning methods, such as random forests, but there are limitations in dealing with concept drift problems, such as difficulty in model updating and data imbalance. Therefore, a study proposes a model that combines the minimum absolute shrinkage and selection operator with gradient boosting tree algorithm to solve the problem of dynamic concept drift and accurately predict the financial difficulties of enterprises. The study selected financial datasets from Chinese A-share listed companies from 2019 to 2022, with selection criteria including but not limited to the company's market value, industry representativeness, and financial information. In order to reduce potential sample bias caused by market structure changes, policy adjustments, and other factors, the study adopts time series and industry stratified sampling methods to ensure the representativeness of the samples. Firstly, conduct a thorough analysis of the two algorithms and apply them to dynamic financial indicator selection in financial samples. Secondly, a comprehensive prediction model is established using the sample similarity index. Experiments compare the performance of the model with a variety of basic classifiers, including random forests, support vector machines, naive Bayes, logistic regression, single decision trees, and ordinary feedforward neural networks. The results show that the accuracy of the model in dynamic environment is 92.47% and 92.31%, F value is 85.33% and 85.12%, G value is 91.78%, 91.65% and 91.92%. The gradient lifting tree classifier performs best in accuracy, F-value and G-value, with an average increase of 0.051 accuracy and 0.07 F-value, while the performance of G-value is stable but not significantly different. Through Wilcoxon test, it is found that similarity weighting significantly improves the prediction effectiveness of most classifiers. The study achieved effective processing of dynamic concept drift for the first time by combining two algorithms and using sample similarity index.

Povzetek: Raziskava predstavlja model za napoved finančnih težav podjetij, ki združuje algoritem LASSO in GBDT za učinkovito obvladovanje dinamičnih sprememb konceptov.

1 Introduction

Timely prediction of corporate financial distress is crucial for managers, investors, and creditors in financial risk management and corporate governance [1]. Timely identification of financial difficulties can avoid certain economic losses, reduce systemic risks in the financial market, and greatly benefit the stable development of enterprise decision-making [2]. However, with the complex changes in corporate governance structure and market environment, existing financial ratio analysis and qualitative judgment are no longer able to meet the needs of enterprises for accurate prediction of financial distress, and when facing high-dimensional and complex data, it is difficult to handle the nonlinear structure in the data and the complex interactions between variables [3]. But with the progress and development of science and technology, machine learning technology has been widely applied in the field of financial forecasting due to its excellent data-driven analysis ability [4]. Among them, LASSO algorithm, with its ability to handle high-dimensional data, can effectively handle variable multicollinearity problems, while Gradient Boosting Decision Tree (GBDT) algorithm has strong advantages in nonlinear modeling and feature interaction recognition [5]. With the increasing requirements for financial risk management and corporate governance, designing new predictive models to more accurately predict and respond to financial difficulties has become an urgent need. In response to this demand, the study proposes a combination of Least Absolute Selection and Shrinkage Operator (LASSO) algorithm and GBDT algorithm to construct a company's financial distress prediction model, aiming to improve the predictive ability of enterprise financial distress by optimizing feature selection and model training process. The innovation of the research lies in using LASSO algorithm to optimize feature selection,

combined with the efficient learning mechanism of GBDT algorithm, to construct a financial distress prediction model that can handle high-dimensional data features and adapt to concept drift. This study provides more accurate tools for risk assessment and prediction in enterprises, helping managers make wiser decisions in complex and ever-changing market environments.

The study is divided into four parts. The first part reviews the research status of Lasso algorithm and GBDT algorithm in various fields, as well as the implementation effects of financial distress prediction methods and the research status of many scholars in this field. The second part is the research and model construction of Lasso algorithm and GBDT comprehensive algorithm for dynamic prediction of corporate financial distress. The first section is the research and analysis of Lasso algorithm and GBDT algorithm, and the second section is the construction of a financial distress prediction model based on GBDT integrated algorithm. The third part verified the accuracy of the two algorithms and tested the performance of the prediction model. The fourth part is a summary and outlook on the research methods and results.

2 Related work

Currently, most financial distress prediction methods suffer from weak model generalization ability and poor adaptability to dynamic concept drift, resulting in unstable prediction performance. Therefore, this study requires the design of new prediction models to improve the accuracy and adaptability of predictions. The research aims to address the issue of low accuracy in predicting financial difficulties faced by enterprises during the global economic downturn cycle. The Lasso algorithm and GBDT algorithm used in the study have advantages in feature selection and prediction. Among them, the Lasso algorithm can effectively select variables and reduce dimensionality when processing high-dimensional data, while the GBDT algorithm has efficient learning ability when dealing with complex nonlinear relationships. However, there is limited research and application of these two algorithms in the field of financial distress prediction, and further exploration is needed on how to effectively combine their advantages to address the problem of dynamic conceptual drift in financial data.

Lasso regression algorithm is a compressed estimation method which not only automatically selects features to reduce the dimensionality but also has the advantage of high computational efficiency. The algorithm has been applied in different data prediction fields because of its good computational performance. Kang et al. applied the Lasso algorithm in tumor treatment. The study firstly processed slides containing tumor tissues to extract pathological data features in tumor tissues, and then used Lasso algorithm to construct a prediction model based on the collected data feature set. The experimental results showed that the model had a relatively excellent prediction accuracy [6]. Motamedi et al. concluded that the data prediction model constructed by the traditional deep learning algorithm had the disadvantages of overfitting and computational complexity, which made it difficult to apply the model constructed by this algorithm to activity prediction for drug analysis, and therefore proposed a feature prediction model constructed using the Lasso regression algorithm. Experimental results show that the model has high prediction accuracy because it can screen out irrelevant feature data [7]. Jiang and Jiang constructed a gene identification model by combining support vector machine and Lasso algorithm. The model was applied to screen and identify normal genes and genes of samples with pulmonary hypertension in a genetic sample set to improve the efficiency of diagnosis. The experimental results showed that the AUC area of the model could reach 0.924 and 0.962 in different sample sets, respectively, with a relatively excellent comprehensive application performance [8]. Miswan et al. proposed to improve the prediction accuracy of data with time-series nature by first classifying the data into ranks according to temporal features using gray relational analysis, and then determining the screening feature rank according to the target value, then use Lasso algorithm for data feature identification, and finally use ML classifier for prediction of the identified results. The experimental results show that the method has a high accuracy of data prediction [9].

The GBDT algorithm is a gradient decision tree for training multiple weak classifiers. The algorithm can solve the optimal value by iterating the classifier in the direction of the minimum loss value through the training results. The GBDT algorithm has been applied in different fields because of its low solution error. Arumugam and Kuppan combined the GBDT algorithm with the SOA algorithm and applied the combined algorithm to the management of grid-connected power generation systems. In the grid-connected generation system, the GBDT algorithm enables the demand monitoring of the grid load, and the SOA algorithm in turn proposes an optimal management scheme based on the monitored results. The implementation results show that the grid-connected generation system with this algorithm can achieve an operational efficiency of up to 95.9375% [10]. Jing et al. used a machine learning algorithm to optimize the GBDT model and used the model in a dynamic differential pricing system for train class seats. The experimental results show that the dynamic differential pricing system proposed in the study can not only balance the occupancy rate among trains, but also determine the update nodes of differential pricing according to the fluctuation of occupancy rate, which has high application value [11]. Huang et al. constructed a feature extraction method using the entropy minimum description length principle and GBDT algorithm in order to can efficiently identify the type of surface water pollution, which can obtain tree transformation features by establishing a nonlinear relationship between water quality indicators and pollution levels, and perform classification identification. Experimental results show that the method has high effectiveness [12]. Ma et al. constructed a GBDT algorithm with strong generalization ability in order to reduce the consumption of human and material resources

in lithology identification. The algorithm accurately identifies rock compositions by sensitivity analysis of lithological parameters such as porosity and saturation. The experimental results show that the accuracy of the algorithm can reach 92%, which provides a new solution direction for rock composition identification [13]. Table 1 summarizes the advantages and disadvantages of each research method.

Table 1: Performance of each research method

Document number	Study year	Research method	Research advantage	Underresearch
Literature [6]	2021	Lasso-based machine learning algorithm	It has high prediction accuracy and can automatically select features	Suitable for specific fields, such as oncology treatment; Low adaptability to dynamic data changes
Literature [7]	2022	LASSO-Random Forest algorithm	It can screen out irrelevant feature data and has high prediction accuracy	High computational complexity and overfitting problem
Literature [8]	2023	LASSO combined with support vector machines	The AUC area in different sample sets is large, and the diagnostic efficiency is high	Only for gene recognition applications, the scope of application is limited
Literature [9]	2021	Grey relation analysis and LASSO improve feature selection	Improve the prediction accuracy of time series data with high accuracy	The method is complicated and the implementation is difficult
Literature [10]	2021	GBDT-SOA combination algorithm	The operation efficiency of grid-connected power generation system is 95.9375%	Mainly used in power grid management, the scope of application is limited
Literature [11]	2021	Improved GBDT model	Dynamic differential pricing system has high application value	Only applicable to train class seat pricing, limited scope of application
Literature [12]	2021	The entropy minimum description length principle is combined with GBDT	The feature extraction method of water pollution identification is effective and the classification is accurate	Mainly for surface water pollution identification, limited application
Literature [13]	2022	GBDT algorithm	The accuracy rate is as high as 92%, and the rock composition identification is accurate	Mainly used in lithology identification, the scope of application is limited

In summary, Lasso algorithm and GBDT algorithm have played important roles in different fields due to their excellent recognition and prediction abilities. The Lasso algorithm automatically selects features through compression estimation in the biomedical field, and has high accuracy in gene recognition and tumor therapy. The GBDT algorithm has improved the accuracy of prediction and classification in grid connected power generation system management and train level seat dynamic price difference system. However, the application research of these two algorithms in financial distress prediction is not sufficient. In response to this research gap, the study combines Lasso algorithm and GBDT algorithm to construct a financial distress prediction model. Provide managers with tools that can predict the financial situation of the enterprise in a timely manner, so as to better control financial risks and formulate crisis response strategies in a timely manner.

3 Construction of a dynamic prediction model of corporate financial distress based on lasso and gbdt integrated algorithm

Affected by the epidemic, the global economy has been in a downward cycle, and predicting financial difficulties is particularly important for the operation of enterprises. Traditional prediction methods have poor adaptability when facing dynamic changes in financial data, and their accuracy is significantly reduced under the influence of concept drift. The Lasso algorithm can effectively select variables and reduce dimensions for high-dimensional data, but its sensitivity is low when dealing with dynamic changing data. The GBDT algorithm is very accurate in capturing nonlinear relationships, but there are problems such as sample imbalance. In response to the above issues, the study proposes to construct a dynamic prediction model for corporate financial distress by combining Lasso algorithm and GBDT algorithm, aiming to improve the accuracy of predicting corporate financial distress and adapt to the dynamic concept drift of financial data [14]. Firstly, the Lasso algorithm is used to dynamically select indicators at different time periods to address the dynamic and nonlinear characteristics of financial data. The GBDT algorithm constructs a prediction model based on selected indicators and integrates methods to improve the stability

and accuracy of the model in the face of concept drift. In order to help enterprise managers, establish reasonable and scientific decision-making, and promote the good development of the enterprise.

3.1 Research and analysis based on LAsso algorithm and GBDT algorithm

Enterprise financial situations are characterized by dynamic changes, and the resulting datasets are multidimensional and multi-period. This study proposes a combination of the GBDT integration algorithm and a combination of prediction weights based on sample similarity for dynamic prediction of the financial distress of enterprises in response to the characteristics of their financial situation. The financial distress prediction based on time lapse is to some extent subject to conceptual drift, and the prediction of conceptual drift needs to be carried out in three aspects. The first is the selection of dynamic indicators, however, there are adaptive dynamics in the dynamic forecasting of financial distress. Therefore, a combination of indicators is selected using the Lasso algorithm for different time periods to complete the compression and screening of variables. The Lasso regression model is shown in Figure 1.

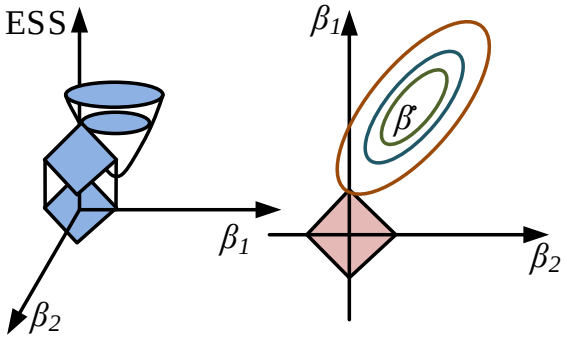


Figure 1: Lasso regression model.

The Lasso algorithm can effectively identify and select important features when processing multi-dimensional data, and remove unimportant features. In the face of complex economic environments, it is possible to dynamically select the most critical financial indicators for current predictions, reduce model complexity and improve prediction accuracy, and achieve variable selection and model simplification. Prior to the LAsso method, the parameters of the regression using OLS can be solved by the optimization problem, whose expression is shown in Equation (1).

$$\hat{\beta}_{ols} = \arg\min_{\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^{'}\beta)^2\right\} \quad (1)$$

In Equation (1), the coefficients are not zero, where the number of independent variables P is too large relative to the sample size, then there is no unique solution for the parameters estimated by OLS, which can be judged as overfitting. In this regard, some regularization is required as a penalty to eliminate the independent variables in order to solve the overfitting generated by the large number of independent variables relative to the sample size. The data distribution on each time has different characteristics, and different financial indicators have different ability to distinguish the degree of risk of the sample, so it is important to select indicators with strong distinguishing ability. Therefore, in order to accurately reflect the financial situation of enterprises in different periods, different periods need to be adopted, and indicators need to be selected using dynamic selection, and finally the final forecast results are obtained through a scientific weighting method. The enterprise financial data produce conceptual drift with time, and the conceptual drift has three types: gradual, repetitive and abrupt. Introducing Markov distance can enhance the sensitivity of the model to concept drift, in order to more accurately capture the subtle changes in financial data over time and reflect the latest financial situation in a timely manner. The sample principle of this method is shown in Figure 2.

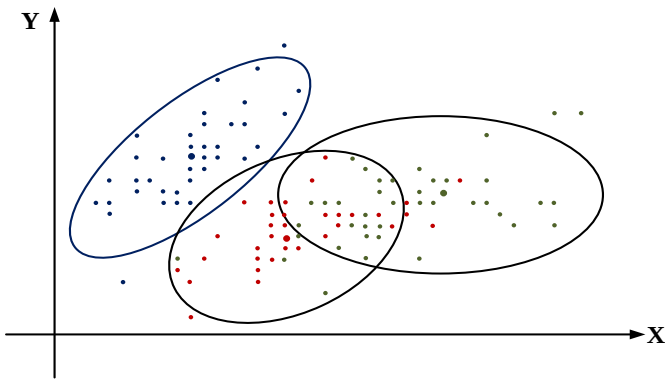


Figure 2: Markov distance principle in weighted sample similarity.

The Marschall distance is a metric that measures the distance from a point to a data division, which takes into account the correlation within the data set and better represents the data set characteristics. With $\vec{x} = (x_1, x_2, x_3, \ldots, x_M)^T$ denoting the indicator vector of the sample and x_j denoting the $j(j = 1,2,3,\ldots,M)$ th indicator of the sample, the Equation defining the Marschall distance from the sample point to the data set $\overleftrightarrow{x}$ D is shown in Equation (2).

$$Mdist(\vec{x}, D) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})} \quad (2)$$

In Equation (2), S denotes the covariance matrix of the data set D, and $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_M)^T$ denotes the vector of mean indicators of the data set D. The $\{D_t\}_{t=1}^{T}$ denotes the training set containing data batches, and T D_{tk} denotes the training subset of the $k\{k = 1,2,3,\ldots,K\}$ class. $X_t = \{\vec{x}_i\}_{i=1}^{N}$ denotes the indicator set of the $t(t = 1,2,3,\ldots,T)$ th training data batch. That is, the larger the Marcian distance between the test sample and the training data set, the smaller the corresponding similarity. Next is the construction of the prediction model based on the degree of conceptual drift, combined with the dynamic indicator selection method using GBDT as the base classifier [15]. The GBDT algorithm integrates multiple decision tree models and gradually reduces model errors through a gradient boosting strategy, solving the problem of overfitting in a single decision tree. The training process of this algorithm is shown in Figure 3.

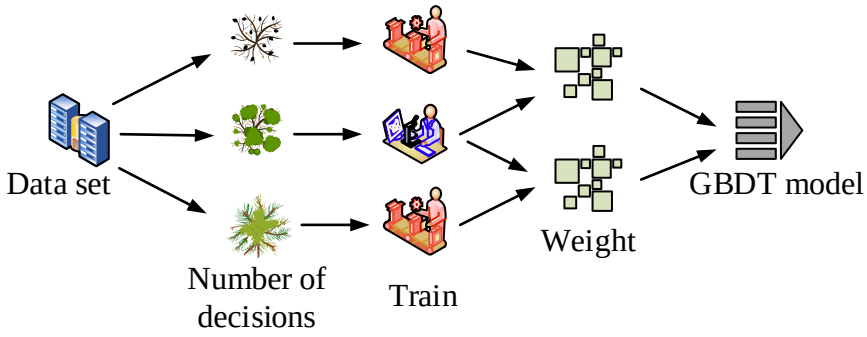


Figure 3: GBDT algorithm training process.

The GBDT algorithm improves prediction accuracy by constructing multiple decision trees and integrating prediction results, which can capture nonlinear relationships in the data. This algorithm continuously improves the model's fit to data, making it more flexible in adapting to changes in data when dealing with concept drift. In the regression problem with m training samples, the expression of the training samples is shown in Equation (3).

$$\left\{(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \ldots, (X^{(m)}, Y^{(m)})\right\} \quad (3)$$

In Equation (3), $X^{(i)}$ represents the data feature vector of each n dimension, and $Y^{(i)}$ represents the output of the i sample. GBDT can be understood as an additive model of decision tree, and its corresponding model Equation is shown in Equation (4).

$$F(x; w) = \sum_{t=0}^{T} \alpha_t h_t(x; w_t) = \sum_{t=0}^{T} f_t(x; w_t) \quad (4) \qquad (4)$$

In Equation (4), x denotes the training sample, w denotes the parameters of the categorical regression tree. The GBDT algorithm is based on increasing the complexity of the model by continuously extending the length of the overfitted model to reduce the overfitting bias, thus achieving the effect of improving the accuracy of the model fit [16].

3.2 Financial distress prediction model construction based on GBDT integration algorithm

To ensure the prediction of each classifier, the accuracy of each classifier needs to be checked using the corresponding training, and the classifiers with too poor discriminatory ability are eliminated. In the actual modeling process, training classifiers are used separately for each original training data, and GBDT is selected as the main model for the training set to better reflect the characteristics of the data set at each time [17]. Next, use classifiers to predict the test samples separately, selecting financial indicators such as profit margin, debt ratio, and cash flow as variables. Finally, the similarity between the test samples and the training set is used to weight the prediction results of each classifier to produce the final prediction results. The GBDT integration framework based on sample similarity is shown in Figure 4.

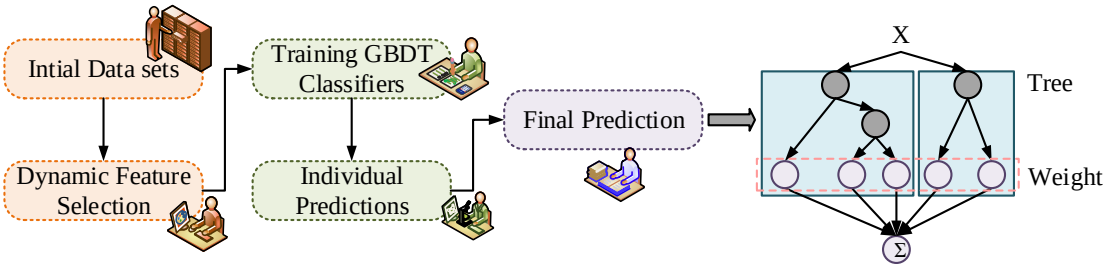


Figure 4: GBDT integration framework based on sample similarity.

The true risk types of the test samples are not predicted in advance, so the prediction results of the classifier on the training set for the test samples are used as the basis for discrimination, and the samples for the risk categories are used in the training set for calculating the similarity. In this model, each company conducts targeted analysis to determine its financial health status. By comparing data between different companies, similarities between them are identified to determine whether the company may encounter financial problems [18]. Then the similarity between the sample i and the training sample set $D_{tk'}^{'}$ sim_{it} is calculated by the formula shown in Equation (5).

$$sim_{it} = \frac{1}{\log(1 + Mdist(\vec{x_i}, X_{tk'}))} \quad (5)$$

In Equation (5), $k^{'}$ is the prediction of the classifier $GBDT_t$ on the test sample i, and the classifier trained on $D_t^{'}$ $GBDT_t$ has a voting weight w_{it} on the prediction of the sample i, whose expression is shown in Equation (6).

$$w_{it} = \frac{sim_{it}}{\sum_{t=1}^{T} sim_{it}} \quad (6)$$

By the larger the final voting weight in Equation (6), the greater the similarity between the training sample set $D_{tk'}^{'}$ and the test sample i, the more reliable the prediction result of the classifier for the sample $GBDT_t$ i. Then the final $T^{'}$ classifier's combined prediction result of for sample y^{*} i is calculated as shown in Equation (7).

$$y^{*} = sign(\sum_{t=1}^{T} w_{it} \hat{y}_{it}) \quad (7)$$

The GBDT algorithm analyzes the historical financial data of each company to predict potential financial problems that the company may face, taking into account the similarity between different companies and improving the accuracy of predictions. The integrated prediction model is a dynamic prediction method for corporate financial distress, and statically selected indicators are able to have certain limitations in the face of dynamic financial distress prediction. The training data of different periods are collected annually with a time interval of 1 year. This dynamic forecasting framework is shown in Figure 5.

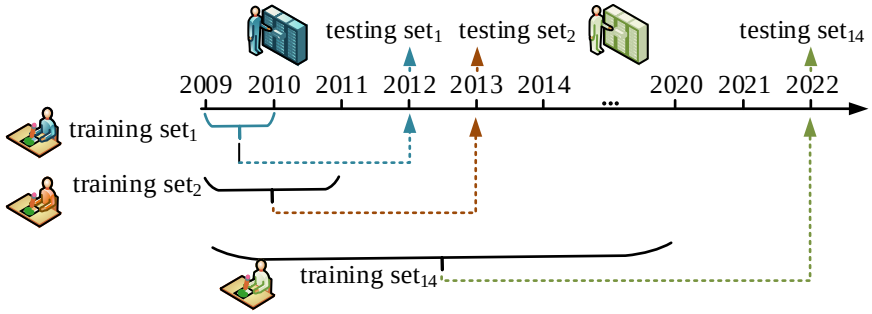


Figure 5: A dynamic prediction framework based on GBDT.

A test sample is taken from $(t-2)$ and the corresponding training sample uses all available data in the year $(t-2)$ to obtain a prediction of whether the sample will be in financial distress after two years. The training data set will cover more and more batches of data as time goes on. All samples are collected for the years 2009-2022 when the risk type is flagged, with 2009 being the first predicted test sample set. This prediction model is dynamic and constantly updated and adapted over time, in order to more accurately predict the financial prospects of the enterprise and provide decision-making information for managers [19, 20]. The structure of the financial distress prediction model based on LAsso algorithm and GBDT algorithm is shown in Figure 6.

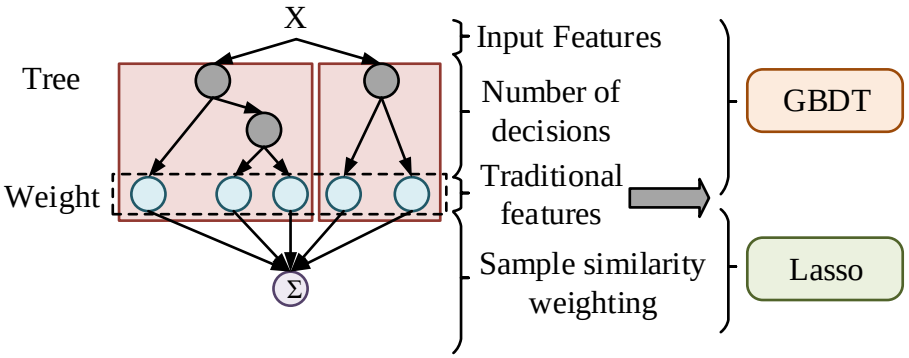


Figure 6: Financial distress prediction model based on Lasso and GBDT algorithms

The model, by predicting the enterprise's financial and material distress situation, allows for the rational allocation of financial and social resources based on the enterprise's risk profile, and promotes the enhancement of the financial economy and social operations. The reasonableness of the model is evaluated to enhance the validity and reliability of the model. The evaluation indexes used in the model proposed in this study are, F-value consisting of recall and accuracy, G-value consisting of recall and correct value, and accuracy rate. The formula for calculating the accuracy rate is shown in Equation (8).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (8)$$

In Equation (8), TP denotes the number of samples correctly predicted as positive, TN denotes the number of samples correctly predicted as negative, FP denotes the number of samples incorrectly predicted as positive, and FN denotes the number of samples incorrectly predicted as negative. And the formula for calculating the G-value is shown in Equation (9).

$$G = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}} \quad (9)$$

In Equation (9), the G-value is the geometric mean of the recall rate and the correct rate predicted by the negative samples. The larger the correct value, the larger the G value. The G value can effectively measure the predictive performance of the model on imbalanced samples, and can provide a more objective and comprehensive evaluation of the model [21]. The overall technical route of the research method is shown in Figure 7.

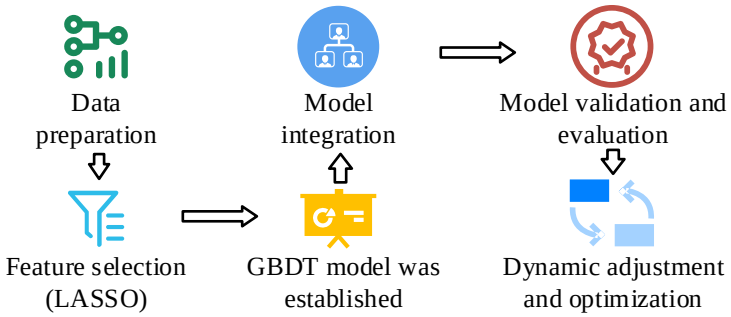


Figure 7: Research technology route

In Figure 7, the algorithm flow of the combined model based on LASSO and GBDT is as follows: Firstly, data collection is carried out to collect the financial data of China's A-share listed companies from 2019 to 2022. The data is then preprocessed, including cleaning the data, processing missing values, and standardizing and normalizing the data. Then the initial features are extracted from the original data set, and the LASSO algorithm is used to select the features, remove the redundant features, and retain the important features. LASSO's equation is shown in equation (10).

$$\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (10)$$

In equation (10), λ represents the regularization parameter. x_{ij} is the value of the j feature of the i sample. β represents the regression coefficient vector, including the intercept term and the coefficients of each feature. N and p represent the number of samples and the number of features respectively. y_i represents the target variable of the i sample. β_j represents the regression coefficient for the j feature. β_o represents the intercept term in the regression model. Finally, the selected feature set is output. Then, the GBDT model is established. Firstly, the data is divided into training set and test set. The GBDT algorithm was used to train the model. Initialize the GBDT model $F_0(x)$ and calculate the residual $r_{im} = y_i - F_{m-1}(x_i)$ for each stage $m = 1$ through M. A new base learner $hm(x)$ is fitted to minimize the residual.

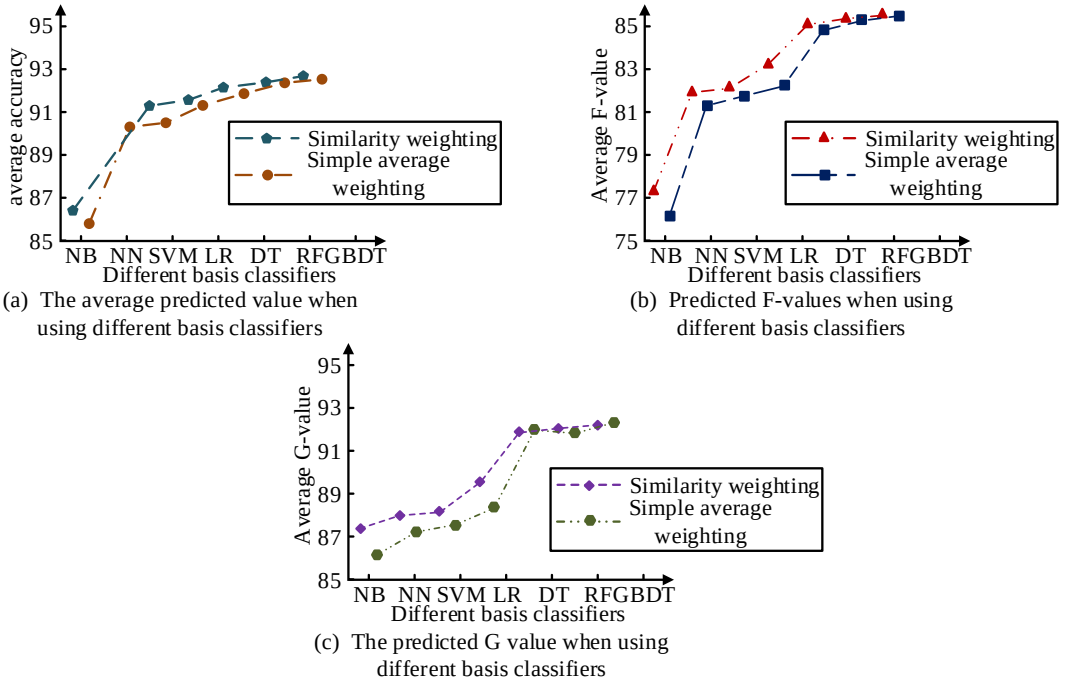
After updating the model, the trained GBDT model is output. The features selected by LASSO are combined with the GBDT model to form the final prediction model. The GBDT model is trained using LASSO-selected features. The model output is combined with the weights to generate comprehensive prediction results. Verify model performance using test sets. Compare the performance of different models and analyze the superiority of the comprehensive model. The model is updated regularly according to new data and market changes. The parameters of LASSO and GBDT were adjusted to improve the adaptability and prediction accuracy of the model.

4 Analysis of experimental results

The study selected the financial data of China's A-share listed companies from 2019 to 2022, covering multiple industries, and predicted the financial distress of enterprises through various financial indicators such as market value, profit margin, debt ratio, and cash flow. The data sets are divided into training sets and test sets in chronological order, simulating time changes in actual predictions, and stratified sampling by industry to ensure representativeness. Different sample balance ratios (1:1 and 1:2) were designed to test the stability and accuracy of the model under unbalanced data. Data preprocessing includes cleaning missing values and outliers, time series processing, Lasso algorithm feature selection and standardization processing. Through multi-step pre-processing and stratified sampling, the quality and representativeness of the data are ensured to reflect the financial situation of the enterprise from different perspectives.

4.1 Performance analysis of integrated algorithms based on Lasso and GBDT

The dynamic prediction framework proposed in this study is trained and predicted with the base classifier master model of GBDT for each time batch training, with a balance of 1:3 (number of financial distress samples: number of non-financial distress samples yet) data. To test the applicability of the dynamic prediction framework proposed in this study on different classifiers, it is compared with different base classifiers for experiments, including random forest (RF), support vector machine (SVM), plain Bayesian (NB), logistic regression (LR), single decision tree (DT), and ordinary feedforward neural network (NN). In order to test the robustness of this prediction model for realistic high-risk samples with unbalanced data, a comparison experiment with different sample balances was designed, including sample sets with balances of 1:1 and 1:2. Firstly, the dynamic prediction results of its three evaluation indexes of accuracy, F-value and G-value under the use of different base classifiers are shown in Figure 8.



(a) The average predicted value when using different basis classifiers

(b) Predicted F-values when using different basis classifiers

(c) The predicted G value when using different basis classifiers

Figure 8: Dynamic prediction results of prediction accuracy, F-value, and G-value under different basis classifiers.

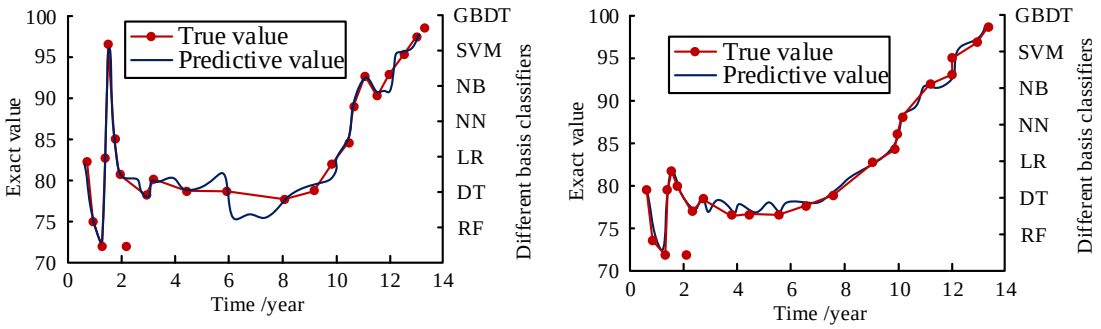
As can be seen from Figure 8, the best predictors for most years are concentrated on two base classifiers, GBDT and RF, with the highest curves represented by GBDT and RF, as well as DT as a base classification also has a better performance. The accuracy of the GBDT base classifier is 92.47 and 92.31, respectively, and the NB classification has the lowest values for the two metrics, with 87.38 and 86.84, respectively. The difference between the two types of base classifiers is large. It can be concluded from Figure 8(c) that the G-values of GBDT, DT, and RF base classifiers are almost close to each other, with G-values of 91.78, 91.65, and 91.92, respectively. Thus, it can be seen that the average values of the three evaluation metrics of accuracy, F-value,

and G-value are all obtained the maximum when GBDT is used as the base classifier, and the integrated prediction effect of GBDT as and classifier is better than other comparison models. For a more intuitive representation, the Wilcoxon test was applied to the gaps in the evaluation indexes of prediction effectiveness after using similarity weighting, and the results are shown in Table 2.

As shown in Table 2, the difference in prediction effectiveness before and after similarity weighting when using RF as the base classifier is not significant under any of the three evaluation metrics measures. For the GBDT classifier, the prediction accuracy of the integrated model after using similarity weighting improved by 0.051 on average and was significant at the 1.3% significance level. Model prediction as measured by F-values improved by 0.07 and was significant at the 10% significance level. There was no significant difference in the prediction effect of the models measured by G-value. The mean difference between the instrumental accuracy, F-value, and G-value before and after similarity weighting was positive for all other comparator-based classifiers when applying the integrated prediction strategy with similarity weighting, and the effect was significantly improved and significant at the 1% significance level. This shows that the sample similarity-weighted integrated prediction strategy has an enhanced effect on the prediction of most common classifiers. To verify the feasibility of the model, data from 2009-2022 were selected as training data for Company A and Company B according to different models. Among them, the Lasso regression model takes the weight of 25, the test data is 77%, and the remaining 27% is the prediction data, and the prediction trend results are shown in Figure 9.

Table 2: Wilcoxon test for the difference in prediction performance of different ensemble models

	Accuracy		F value		G value	
	Mean difference	Significance	Mean difference	Significance	Mean difference	Significance
RF	0.051	0.006	0.049	0.062	0.039	0.667
DT	0.034	0.831	0.006	0.951	0.074	0.998
NB	0.176	0.008	0.534	0	0.603	0
NN	0.224	0	0.314	0	0.091	0.005
LR	1.003	0	1.376	0	0.746	0
SVM	0.355	0	0.921	0	0.861	0
GBDT	0.478	0	1.068	0	0.879	0

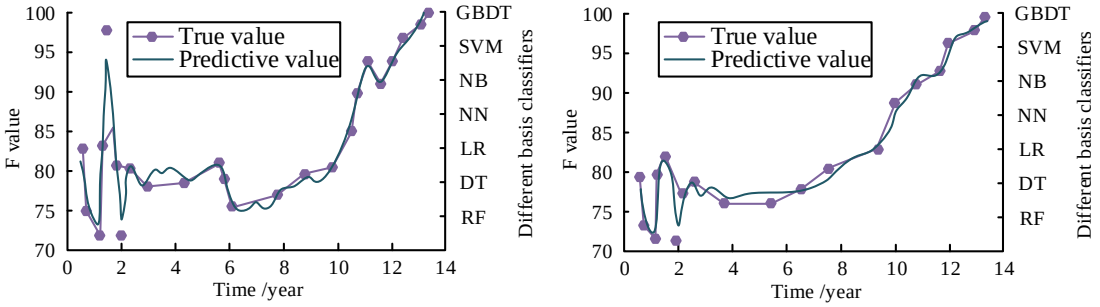


(a) Comparison between predicted results and true values of Company A under the Lasso regression model

(b) Comparison between predicted results and true values of Company B under the Lasso regression model

Figure 9: Company A and B's accurate values under different basis classifiers from 2009 to 2022.

As can be seen from Figure 9, the Lasso regression model prediction results are less stable, and the sample data fit the prediction trend graph better when there is little fluctuation. Through Figure 9 (a), it can be concluded that Company A's data fluctuates more between year 4 and year 8 of the test, when the fit is poor and the curve is more divergent. Through Figure 9 (b), it can be concluded that Company B has less data fluctuation and the fit is better under this model with less curve divergence. This shows that the stability of the Lasso regression prediction model is poor and does not meet the model prediction accuracy requirements. The prediction trend results for the GBDT regression model are shown in Figure 10.



(a) Comparison between predicted results and true values of Company A under the GBDT model

(b) Comparison between predicted results and true values of Company B under the GBDT model

Figure 10: Company A and B's conducted F-values under different basis classifiers from 2009 to 2022.

As can be seen from Figure 10, the GBDT regression model prediction results are roughly consistent with the actual values, and the prediction accuracy is good. Through Figure 10 (a), it can be concluded that Company A's data fluctuates less between the 6th and 14th year of the test, and the prediction results are less different from the actual values. Through Figure 10 (b), it can be concluded that the data fluctuation curve of Company B almost fits and the prediction results under this model basically belong to the actual value. It can be seen that the prediction accuracy of the GBDT regression prediction model still has a relatively large reduction, and the stability needs to be further improved. All curves are Receiver operating characteristic curve (ROC) curves, and the area enclosed by the coordinate axis below the ROC curve is the AUC. The performance comparison results of each model are shown in Figure 11.

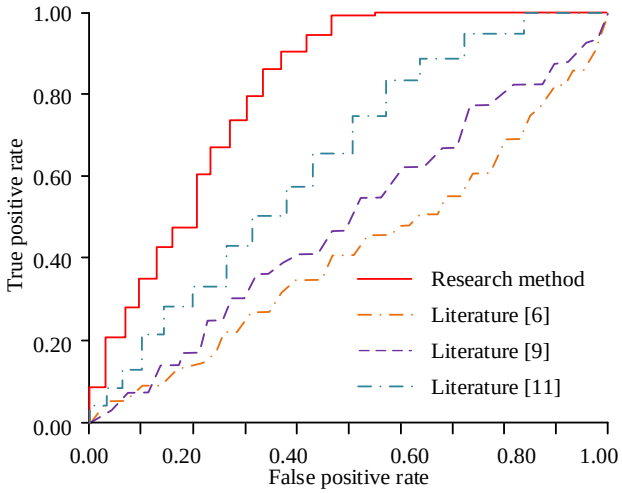
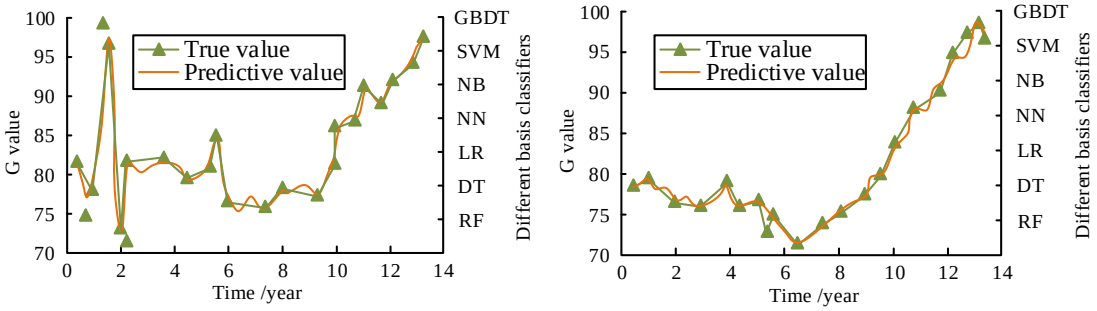


Figure 11: Statistical results of AUC indicators of each model.

As can be seen from Figure 11, the ROC curve of the prediction model based on LASSO-GBDT algorithm proposed by the research is always above the ROC curve of other models, and the AUC area is the largest, 0.90. It can be seen that the prediction performance of the studied model is more accurate than that of the existing classical model.

4.2 Performance analysis of the dynamic prediction model of enterprise financial distress

In addition, this experiment collects financial data of Chinese A-share listed companies from 2009 to 2022, and this dataset is used as the main research object. The prediction trend results for the Lasso-GBDT regression model are shown in Figure 12.



(a) Comparison between predicted results and true values of Company A under the Lasso-GBDT model

(b) Comparison between predicted results and true values of Company B under the Lasso-GBDT model

Figure 12: Company A and B's conducted G values under different basis classifiers from 2009 to 2022.

As can be seen from Figure 12, the stability and accuracy of the Lasso-GBDT regression model have been substantially improved. Through Figure 12 (a), it can be concluded that Company A has a better fit of the data curve during the test period, and the predicted results are almost close to the actual values, with a substantial improvement in accuracy and greater stability. Through Figure 12 (b), it can be concluded that the data fluctuation curve of Company B is almost consistent, and the prediction effect will not be affected by a small amount of pathological data under this model, and the accuracy is higher compared with the single prediction model. It can be seen that the Lasso-GBDT regression prediction model has better prediction accuracy and stability compared with the single-item model. To further test the robustness of the model, the effect of three different balances of the sample on the validity of the text-integrated prediction model was explored, and the results are shown in Table 3.

Table 3: Wilcoxon test for differences in model prediction performance on samples with different equilibrium degrees

	Accuracy		F value		G value	
	Mean difference	Significance	Mean difference	Significance	Mean difference	Significance
1:01	0.362	0.000	0.287	0.000	0.381	0.000
1:02	0.079	0.001	0.097	0.002	0.037	0.026
1:03	0.067	0.002	0.096	0.003	0.034	0.018

As can be seen from Table 3, the data balance corresponding to the highest evaluation index of different evaluation indexes is different, and different evaluation indexes have their own focus on different evaluation dimensions, reflecting the scientific nature of the model prediction effects. the prediction mean differences for the 1:1 data on all three evaluation indexes are significant at the 1% significance level, the prediction mean differences for the 1:2 and 1:3 data on accuracy and F-values are significant at the 1% The differences in the predicted mean values for the 1:2 and 1:3 data were significant at the 1% level of significance for both accuracy and F-values, and the differences in the predicted mean values for G-values were significant at the 4% level of significance. Thus, the mean differences of the evaluation indicators before and after similarity weighting are significantly positive for all three balance degrees, indicating that the prediction of the model becomes better after similarity weighting.

5 Discussion

In exploring the performance of different basic classifiers in ensemble models, by comparing the three key evaluation indicators of accuracy, F-value, and G-value, the results show that GBDT and RF classifiers generally outperform other algorithms. Among them, GBDT achieved an accuracy of 92.47%, and also demonstrated excellent performance in terms of F and G values, with a G value of 91.78%. In addition, GBDT improved its accuracy by 0.051 and F-value by 0.07 after similarity weighting. The research results were compared with relevant international studies, such as Zhang et al.'s detailed exploration of the application of RF and GBDT in financial market forecasting, and confirmed the effectiveness of these two algorithms in handling nonlinear and complex data [22]. A study by Nykamp K et al. on imbalanced classification datasets also showed that the predictive accuracy of the model can be significantly improved through appropriate sample weighting [23]. The consistency of the research results has strengthened the reliability of this study, further confirming that the sample similarity weighting strategy has a positive impact on the performance of multiple classifiers, which is of great significance for the improvement and application of ensemble models.

The proposed model combined with LASSO and GBDT shows excellent prediction ability under dynamic environment, and the prediction accuracy reaches 92.47% and 92.31% respectively. In contrast, traditional random forests and support vector machines perform poorly when dealing with concept drift problems, such as the accuracy of the naive Bayes classifier at 87.38% and 86.84%. The F-values of this model are 85.33% and 85.12%, which are significantly better than other basic classifiers such as random forest and single decision tree. In addition, the study model has a G-value of 91.78%, 91.65%, and 91.92%, respectively, which outperforms other comparator-based classifiers in handling unbalanced data and dynamic environments. The reason for the

performance difference is that the research solves the multicollinearity problem by dynamically selecting financial indicators of different time periods to adapt to the dynamic change of data. The sample similarity weighting mechanism is introduced to enhance the stability and accuracy of the model in the case of concept drift. The feature selection capability of LASSO combined with the nonlinear modeling capability of GBDT makes the model perform well when dealing with high-dimensional, multi-period enterprise financial data, avoiding the problems of easy overfitting and high computational complexity of traditional methods. Through the sample similarity index, the dynamic concept drift problem in financial distress prediction is effectively solved. This method significantly improves the accuracy and stability of the prediction, while reducing the complexity of the model.

6 Conclusion

At a time when the economic and social environment is constantly changing, finding effective methods for financial distress is crucial for financial companies to help business managers be more forward-looking in their assessment of financial conditions and promote more stable corporate development. However, for the existing prediction models, for many forms of conceptual drift, they cannot accurately predict the financial distress situation of enterprises. This study proposes a dynamic prediction of corporate financial distress based on Lasso and GBDT integrated algorithm for dynamic conceptual drift. The experimental results show that the data balance corresponding to the highest evaluation index of different evaluation indexes is different, and different evaluation indexes have their own focus on different evaluation dimensions, reflecting the scientific nature of the model's prediction effect. the prediction means differences for the 1:1 data on all three evaluation indexes are significant at the 1% significance level, and the prediction mean differences for the 1:2 and 1:3 data on both accuracy and F-value are significant at the predicted mean differences in accuracy and F-values for both 1:2 and 1:3 data were significant at the 1% significance level, and the predicted mean differences in G-values were significant at the 4% significance level. For the GBDT classifier, the prediction accuracy of the integrated model after using similarity weighting improved by 0.051 on average and was significant at the 1.3% significance level. The improvement in model prediction as measured by F-values was 0.07 and was significant at the 10% significance level. There was no significant difference in the predictive effect of the models measured by G-values. The Lasso regression model took a weight of 25, with 77% of the test data and the remaining 27% of the predicted data. This shows that the sample similarity weighted integrated prediction strategy has an enhancing effect on the prediction of most common classifiers. By using a financial distress prediction model under dynamic concept drift, enterprise managers can more accurately identify financial risks and prepare in advance. In practical applications, the research model can deal with high and dynamic data, and improve the prediction accuracy and stability. Managers and investors can periodically use the model to assess financial health and adjust strategies based on the projections. The adaptability and prediction accuracy of the model can be enhanced by dynamic selection of financial indicators and sample similarity weighting in different time periods. This helps managers and investors make more precise decisions in response to different market environments and business conditions. This study tested the prediction models on the financial status of Company A and Company B. However, in practice, the financial status of each company has multiple manifestations and can also be classified into multiple financial health classes, which need to be explored in more depth and prediction models established for multidimensional discriminations, for which further research and discussion are needed.

Data availability statement

The data used to support the findings of this study are all in the manuscript.

Conflict of interest

The authors declare that they have no competing interests.

Funding statement

No funding was received.

References

[1] Kuerten B G, Samuel B, Bonner M J, Ayuku D O, Njuguna F, Taylor S M, Puffer E S. Psychosocial burden of childhood sickle cell disease on caregivers in Kenya. *Journal of Pediatric Psychology*, 2020, 45(5):561-572. https://doi.org/10.1093/jpepsy/jsaa021

[2] Cuesta-González M, Paredes-Gazquez J, Ruza C, Fernandez-Olit B. The relationship between vulnerable financial consumers and banking institutions. A qualitative study in Spain. *Geoforum*, 2021, 119(3):163-176. https://doi.org/10.1016/j.geoforum.2021.01.006

[3] Lavikainen P, Aarnio E, Niskanen L, Mantyselka P, Martikainen J. Short-term impact of co-payment level increase on the use of medication and patient-reported outcomes in Finnish patients with type 2 diabetes. *Health Policy*, 2020, 124(12):1310-1316. https://doi.org/10.1016/j.healthpol.2020.08.001

[4] Ohishi M, Fukui K, Okamura K, Itoh Y, Yanagiharaa H. Coordinate optimization for generalized fused Lasso. *Communications in Statistics-Theory and Methods*, 2021, 50(24):5955-5973. https://doi.org/10.1080/03610926.2021.1931888

[5] Luo S, Zhao W, Pan L. Online GBDT with chunk dynamic weighted majority learners for noisy and drifting data streams. *Neural Processing Letters*,

2021, 53(5):3783-3799. https://doi.org/10.1080/03610926.2021.1931888

[6] Kang J, Choi Y J, Kim I, Lee H, Kim H S, Baik S H, Kim N K, Lee K Y. LASSO-based machine learning algorithm for prediction of lymph node metastasis in T1 colorectal cancer. Cancer Research and Treatment: Official. *Journal of Korean Cancer Association*, 2021, 53(3):773-783. https://doi.org/10.4143/crt.2020.974

[7] Motamedi F, Pérez-Sánchez H, Mehridehnavi A, Fassihi A, Ghasemi F. Accelerating big data analysis through LASSO-random forest algorithm in QSAR studies. *Bioinformatics*, 2022, 38(2):469-475. https://doi.org/10.1093/bioinformatics/btab659

[8] Jiang C, Jiang W. Lasso algorithm and support vector machine strategy to screen pulmonary arterial hypertension gene diagnostic markers. *Scottish Medical Journal*, 2023, 68(1):21-31. https://doi.org/10.1177/00369330221132158

[9] Miswan N H, Chan C S, Ng C G. Hospital readmission prediction based on improved feature selection using grey relational analysis and LASSO. *Grey Systems: Theory and Application*, 2021, 11(4):796-812. https://doi.org/10.1108/GS-12-2020-0168

[10] Arumugam P, Kuppan V. A GBDT-SOA approach for the system modelling of optimal energy management in grid-connected micro -grid system. *International Journal of Energy Research*, 2021, 45(5):6765-6783. https://doi.org/10.1002/er.6270

[11] Jing Y, Guo S, Chen F, Wang X, Li K. Dynamic differential pricing of high-speed railway based on improved GBDT train classification and bootstrap time node determination. *IEEE Transactions on Intelligent Transportation Systems*, 2021, 23(9):16854-16866. https://doi.org/10.1109/TITS.2021.3106042

[12] Huang P, Wang L, Hou D, Lin W, Yu J, Zhang G, Zhang H. A feature extraction method based on the entropy-minimal description length principle and GBDT for common surface water pollution identification. *Journal of Hydroinformatics*, 2021, 23(5):1050-1065. https://doi.org/10.2166/hydro.2021.060

[13] Ma L, Xiao H, Tao J, Su Z. Intelligent lithology classification method based on GBDT algorithm. *Editorial Department of Petroleum Geology and Recovery Efficiency*, 2022, 29(1): 21-29. https://doi.org/10.13673/j.cnki.cn37-1359/te.2022.01.003

[14] Yan P, Zhou Y. Application of recommendation algorithm based on matrix dimensionality reduction model in network information analysis model. Informatica, 2024, 48(9). https://doi.org/10.31449/inf.v48i9.5969

[15] Li R, Chang C, Justesen J M, Tanigawa Y, Tibshirani R J. Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. *Biostatistics*, 2022, 23(2):522-540. https://doi.org/10.1093/biostatistics/kxaa038

[16] Zhang N, Zhang Y, Sun D, Kim-Chuan T. An efficient linearly convergent regularized proximal point algorithm for fused multiple graphical lasso problems. *SIAM Journal on Mathematics of Data Science*, 2021, 3(2):524-543. https://doi.org/10.1137/20M1344160

[17] Luo S, Zhao W, Pan L. Online GBDT with chunk dynamic weighted majority learners for noisy and drifting data streams. Neural Processing Letters, 2021, 53(5):3783-3799. https://doi.org/10.1007/s11063-021-10565-z

[18] Zhu H, Li H. Predict prices of second-hand house using gbdt algorithm and PSO algorithm. *Frontiers in Economics and Management*, 2021, 2(11):513-524. https://doi.org/10.6981/FEM.202111_2(11).0069

[19] Slavova-Azmanova N S, Newton J C, Saunders C, Johnson C E. 'Biggest factors in having cancer were costs and no entitlement to compensation'-The determinants of out-of-pocket costs for cancer care through the lenses of rural and outer metropolitan Western Australians. *Australian Journal of Rural Health*, 2020, 28(6):588-602. https://doi.org/10.1111/ajr.12686

[20] Guo Y, Mustafaoglu Z, Koundal D. Spam detection using bidirectional transformers and machine learning classifier algorithms. *Journal of Computational and Cognitive Engineering*, 2023, 2(1):5-9. https://doi.org/10.47852/bonviewJCCE2202192

[21] Afrin S, Shamrat F M J M, Nibir T I, Muntasim M F, Moharram M S, Imran M M, Applicable A N. Supervised machine learning based liver disease prediction approach with LASSO feature selection. *Bulletin of Electrical Engineering and Informatics*, 2021, 10(6):3369-3376. https://doi.org/10.11591/eei.v10i6.3242

[22] Lei H. Financial index data prediction based on improved GBDT model. *IEEE International Conference on Artificial Intelligence and Computer Applications*, Dalian, China, 2021, 13(2): 697-702. https://doi.org/10.1109/ICAICA52286.2021.9498075

[23] Nykamp K, Anderson M, Powers M, Garcia J, Herrera B, Ho Y Y, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genetics in Medicine*, 2020, 22(1): 240-241. https://doi.org/10.1038/gim.2017.37