ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Research on Accounting and Financial Risk Management of the Government's Secure Housing Construction Funding Model

Zhen Li[1,*]

[1] School of Intellectual Accounting, Henan Vocational College of Agriculture, Zhengzhou, Henan, 451450, China

SUMMARY: *The study addresses the accounting and financial risk issues under the housing construction financing model, uses the decision tree algorithm to process the accounting data, and establishes the decision tree accounting data mining model, and adopts fuzzy correction processing to realize the data mining measurement. The XGBoost algorithm is introduced to establish a financial risk early warning model, and its hyperparameters are optimized with early warning accuracy and early warning efficiency as the optimization objectives, and multiple dimensional financial indicators are selected as the early warning features. Meanwhile, the prediction effects of different early warning models under different indicator sets are compared, and it is found that the XGBoost model performs optimally in each indicator, with accuracy, false positive rate, recall and precision of 0.9843, 0.1254, 0.9821 and 0.8921, respectively. Finally, by using the SHAP additive explanation algorithm, the key financial indicators that have an impact on the financial risk warning results were extracted. Indicators such as "asset return rate" and "investment return rate" have a positive impact on the financial situation of housing construction fund raising, while "total asset turnover rate" and "asset-liability ratio" have a negative impact on the financial situation of housing construction fund raising.*

KEYWORDS: *decision tree; XGBoost; SHAP; financial risk warning; accounting*

1 Introduction

With the acceleration of China's urbanization process, the construction of governmental guaranteed housing has become an important project to improve people's livelihood [1, 2]. However, guaranteed housing construction is characterized by large-scale investment, long construction period, and strong policy dependence, and its financing model faces challenges from accounting and financial risk management in multiple dimensions [3, 4].

In terms of accounting, the problems of accounting for China's secure housing construction are mainly concentrated in four aspects: (1) the dual accounting system of institutional accounting system and real estate development enterprise accounting system in parallel; (2) the asset-liability ratio of many enterprises is higher than the average; (3) it is difficult to balance the funds for the construction of real estate engineering projects of the enterprise; (4) the standard of the apportionment coefficient is not uniform. For the above problems in the accounting process should be carried out thematic discussions, combined with the actual situation to put forward scientific and reasonable solutions to improve the quality of accounting information [5, 6]. In terms of financial risk management, the first is the problem of fund

*lz15838238168@163.com
https://doi.org/10.65102/is2026559
Published: 30 April, 2026

allocation. As the funds raised mainly rely on the government's financial allocation, the lack of diversified sources of funds causes the regional distribution of funds to be unreasonable, with some regions and projects receiving more funds and some regions and projects receiving less funds [7-9]. Second is the problem of financial supervision. The construction of sheltered housing involves the cooperation of many departments and units, and the division of responsibilities of the supervisory departments is not clear enough, leading to the blurring of responsibilities and weak supervision [10, 11]. The means of financial supervision are not perfect enough to detect irregularities and unstandardized financial behavior in time [12]. In addition, there are risk prevention and control problems. There are risk factors such as fluctuations in market demand, project construction and other risk factors in the construction of protective housing, but in terms of financial management, there is a lack of risk assessment and early-warning mechanisms, which leads to a lack of awareness of risk [13-16]. For these problems, corresponding countermeasures should be taken to strengthen the scientific management of financial management, the effectiveness of financial supervision and the systematization of risk prevention and control, so as to realize the sustainable development of safeguarded housing construction [17-19].

The research is carried out by integrating various advanced machine learning algorithms and theories, such as XGBoost, NSGA-II, SHAP and so on. Firstly, a multi-objective approach is adopted to construct a multi-order financial data mining accounting process and establish a decision tree informationized accounting data mining model. Financial risk analysis of housing construction financing is carried out by mining accounting data. Then the financial risk early warning model is constructed based on the XGBoost algorithm, multiple indicators are selected as early warning features, and the sample imbalance problem faced when conducting financial risk early warning research is solved through the SMOTE oversampling method. Comparative studies are conducted in four aspects: accuracy, false positive rate, recall rate and precision to verify the accuracy and reliability of the model. Finally, SHAP is introduced to visualize and analyze the model and explore the relationship between early warning indicators and financial risk of housing construction financing model from the perspective of prediction.

2 Data mining for accounting for the financing of shelter construction

2.1 Pre-processing of financial data

In order to improve the efficiency of financial accounting, it is necessary to pre-process the collected financial data. The article can use dimensionless processing and convergent processing, etc., to carry out basic classification, screening and classification processing of financial data. For data of different nature, convergent integration is carried out according to the inverter indicators, and the same type of data is merged. Design the data preprocessing structure, as shown in Figure 1.

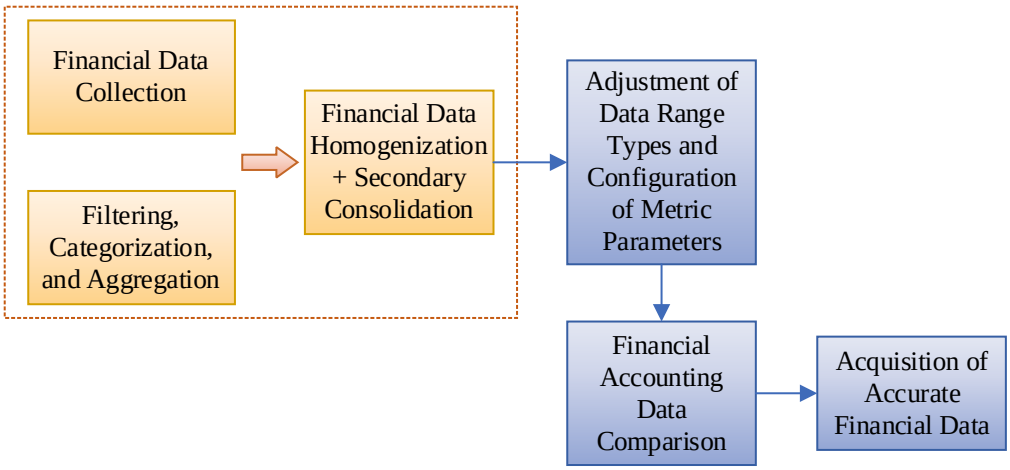


Figure 1: Fnancial data preprocessing structure

2.2 Constructing a multi-order financial data mining accounting process

Compared with the traditional financial data mining accounting process, the multi-order financial data mining accounting process is more targeted, and the calculation efficiency is higher, which can more accurately control and mine the data in the financial accounting process. According to the decision tree principle, the accounting process is practically applied. Data mining algorithm is a continuous process, mostly in the absence of clear assumptions, set up collaborative processing of financial processing tasks, the need to clarify the key points of the accounting process, the establishment of interactive relationships between the various accounting links.

The article combines the actual financial processing needs of power supply enterprises to set corresponding measurement goals. The same type of target is an independent data mining measurement class, which can be divided into the basic data collection and processing class, financial data accounting and classification process and comprehensive analysis and reflection process. Calculate the current expectation value, the specific formula is:

$$D=\sum_{t=1}^{T}\left(wt-\sqrt{1+m-n^2}\right)\times\frac{1}{w} \tag{1}$$

where: D denotes the data mining measurement expectation; w denotes the mining range; t denotes the number of mining times $\left(t=1,2,\cdots,T\right)$; m denotes the confidence coverage region; n denotes the repeated mining region.

2.3 Establish decision tree informationized accounting data mining model

Based on the decision tree principle and the actual financial situation of the power supply enterprises, the informationized accounting data mining algorithm model is established. Based on the change of the expectation value of the data mining measurement, the objective function of the model measurement is calculated, and the specific formula is:

$$P=\sum_{Y=1}^{y}\varepsilon_Y-\frac{V_2}{V_1+\tau_Y}\times V_1\varepsilon_Y \tag{2}$$

where: P denotes the data mining algorithm model objective function; ε_Y denotes the mining unit difference; Y denotes the number of digging times $(Y=1,2,\cdots,y)$; V_1 and V_2 denote the data mining pickup orders; τ_Y denotes the mining controllable difference.

Combined with the current objective function, calculate the model decision tree data mining algorithm controllable accounting error, the specific formula is:

$$M=a-b\times\sum_{d=1}^{D}v_d^2+c-\frac{q_d+b-a^2}{cr} \tag{3}$$

where: M denotes the mining algorithm controllable accounting error; a denotes the accounting region; b denotes the repeated accounting region; v_d denotes the permutation ratio; d denotes the number of data mining times $(d=1,2,\cdots,D)$; c denotes the value of weights; q_d denotes the training error; and r denotes the independent mining range.

Based on the principle of decision accounting accounting data mining algorithm model, combined with the design of the financial situation of housing construction enterprises and data mining algorithms, multi-dimensional accounting financial data, through the decision tree to strengthen the control of the measurement error, so as to improve the corresponding data mining algorithm model, the output of real and reliable accounting results.

2.4 Fuzzy correction processing to achieve data mining measurements

Fuzzy correction processing refers to the second auxiliary correction measurement for the measurement results processed by the algorithmic model, so as to ensure the real stability of the final test. The article uses the fuzzy calculation structure and data mining algorithm to construct a correction space and clarify the judgment rules and standards of fuzzy calculation. By comparing the measurement results from the model with the initial measurement standard, the article clarifies the link where the error occurs. Then the model is used to re-measure the link, transform the fuzzy value into a clearly corrected value, and establish the mapping of the algorithm on the financial processing software or platform to realize the complete and specific fuzzy correction processing. It should be noted that the current correction standard is not fixed, and can be combined with the actual needs of the extended extension, more flexible, able to cover the changing financial scope of the housing construction enterprises, to ensure that the results of the financial measurement is true and reliable, and to enhance the accuracy of the financial data processing calculations.

3 Early warning model on the risk of financing the construction of sheltered housing

3.1 XGBoost model

The idea of the XGBoost algorithm is to continuously generate trees, and for each tree growth, this is done by continuously performing feature splitting. To be precise, each time a tree is generated, the predicted residuals need to be fitted. In the training process, after generating k trees, if we want to predict the score of a sample, we will sum up the scores on the leaf nodes of each tree according to its features, i.e., the predicted value of this sample.

First, the expression of the XGBoost model is as follows:

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i) \tag{4}$$

where $F = \left\{ f(x) = \omega_{q(x)} \right\} \left(q : R^m \to T, \omega \in R^T \right)$, where $\omega_{q(x)}$ is the score of the leaf node q and $f_k(x_i)$ is one of the regression trees and the k th base decision tree.

The XGBoost objective function (loss function) is defined as:

$$Obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{5}$$

As in (5), there are two items, the one used to compare the gap between the predicted and real scores of the samples is the empirical risk loss function, it is easy to see that the smaller the gap is the better the model performs, and the overfitting problem of the model is often caused by optimizing this item only; the one used to measure the complexity of the model is the structural risk loss function, which is also a regularization term, in general, the larger the value of this item the better it performs, but the overfitting problem of the model can be also caused by its excessively large value. Thus, the two terms of Eq. (5) need to be balanced to make the model perform better while avoiding the overfitting problem.

As mentioned above, the idea of the XGBoost algorithm in the newly generated tree (i.e., the newly learned function) to fit the predicted residuals. Then, after generating t trees, the predicted scores can be written as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{6}$$

Therefore, the objective function for the t th iteration is:

$$Obj^t = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + cons\tan t \tag{7}$$

The next step is to find the f_t that minimizes the objective function and minimizes the fitting error of the residuals.The idea of XGBoost is to approximate the objective function by a Taylor's second order expansion at $f_t = 0$. Thus the objective function is approximated as:

$$\begin{aligned} Obj^t = & \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] \\ & + \Omega(f_t) + constant \end{aligned} \tag{8}$$

Here, the first order derivatives and second order derivatives are denoted by g_i and h_i respectively:

$$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right), h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right) \tag{9}$$

In optimizing the objective function, the predicted scores of the first $t-1$ trees with y residuals have been determined, and the objective function can be simplified as:

$$\sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{10}$$

The regularization term contains the number of leaf nodes T and the fraction of leaf nodes ω, the coefficients γ and λ of both of them control the oversize of T and ω respectively, so as to avoid overfitting. In XGBoost algorithm, the regularization term is defined as follows:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{11}$$

In the simplified objective function (10), the loss function values of all samples are accumulated. Also, each sample can correspond to a leaf node, i.e., each leaf node corresponds to a sample, so reorganize all the samples of the same leaf node:

$$\begin{aligned} Obj^t &\approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \\ &= \sum_{i=1}^{n}\left[g_i\omega_{q(x_i)} + \frac{1}{2}h_i\omega_{q(x_i)}^2\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\omega_j^2\right] + \gamma T \end{aligned} \tag{12}$$

The final objective function (12) equation is rewritten as a quadratic function on ω, then at this point it can be solved in a variety of ways, for example, according to the vertex formula can be solved for the optimal ω and the corresponding optimal objective function value. Define $G_j = \sum_{i\in I_j} g_i$ and $H_j = \sum_{i\in I_j} h_i$, where $I_j = \{i \mid q(x_i) = j\}$ denotes the set of sample labels in the sample that are assigned to the j th leaf node. Therefore, the optimal ω and the optimal value of the objective function are, respectively:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda},\ Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{13}$$

Next describe how each tree is split during the training process of the XGBoost model. Define the information gain index of the feature A_i of the t th tree at the cut-off point a as $Gain(D, A_i = a)$, which is expressed as:

$$Gain(D, A_i = a) = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{14}$$

The above equation (14) consists of four terms: the fractions on the newly split left and right leaves, the fractions on the leaf nodes before splitting and the regularization term coefficients. Thus, the splitting principle is that the gain of the objective function generated after the splitting of a node is greater than the regularization term γ before the splitting is allowed. Compared with the CART regression tree, the parameters γ and λ that control the complexity of the model are added, and the optimal features and the optimal cut-off point can be obtained by ordering the values of each feature $Gain$.

3.2 Selection of research indicators

The selection of financial indicators will directly or indirectly affect the performance of the early warning model, and the most direct reason for a company to have a financial crisis is that the various financial conditions of the enterprise are abnormal. The following principles for the selection of indicators are formulated:

(1) Comprehensiveness. Financial indicators as a reflection of an enterprise's financial situation and the evaluation of various operating conditions, in the case of failure to determine the key financial indicators that have an impact on the enterprise's financial crisis, the selection of financial indicators must be fully integrated with the characteristics of China's securities market, and should be representative of the premise of integrated, comprehensively reflecting the overall financial situation of the enterprise.

(2) Practicality. The selected financial indicators should have a strong degree of recognition, and at the same time ensure that they can accurately and comprehensively reflect the actual operation of an enterprise.

(3) Comparability. In the selection of financial indicators as a predictive feature, should also consider the comparability of the indicators in different stocks, and the same company in different periods of the indicators are also comparable, while ensuring that the selection of indicators in the unit of measurement of horizontal comparability, but also to ensure that the vertical comparability in different periods.

(4) Operability. First of all, when making the selection of indicators is often the first thing to consider is the operability of the indicators, that is, the ease of understanding of the financial indicators and the feasibility of data collection, the selected financial indicators can be accurately measured in real life.

Based on the above principles, as shown in Table 1, six aspects, specifically including solvency, ratio structure, operating ability, profitability, development ability and relative value, were finally selected.

Table 1: Financial index

Classify	Financial index	Code
Debt paying ability	Current ratio (%)	A1
	Quick ratio (%)	A2
	Cash ratio (%)	A3
	Interest Coverage Ratio (%)	A4
	asset-liability ratio (%)	A5
Structure of rate	Current assets ratio (%)	B1
	cash to assets ratio (%)	B2
	Working Capital Ratio (%)	B3
	Fixed assets ratio (%)	B4
	Current Liabilities Ratio (%)	B5
	Operating profit percentage (%)	B6
Capax negotii	Average accounts receivable turnover ratio	C1
	Inventory turnover ratio	C2
	Turnover of payable	C3
	Working capital turnover	C4
	Turnover of current assets	C5
	Fixed asset turnover	C6
	Turnover of total capital	C7
Profitability	Return on Assets (%)	D1
	Total net profit margin (%)	D2
	Operating Gross Margin (%)	D3
	Operating Profit Margin (%)	D4
	Rate of return on investment (%)	D5
Development capacity	Capital preservation and appreciation rate (%)	E1
	Rate of capital accumulation (%)	E2
	Total Asset Growth Rate (%)	E3
	Return on equity growth rate (%)	E4
	Net profit growth rate (%)	E5
	Sustainable Growth Rate (%)	E6
Relative value	Pe ratio	F1
	Market sales ratio	F2
	Market rate	F3
	Price-to-book ratio	F4

3.3 Model Training and Optimization

This study comprehensively selects a number of different dimensions of financial indicators as features, through the organic integration of XGBoost, NSGA-II, SHAP and other advanced machine learning algorithms, to carry out a financial risk early warning model based on interpretable machine learning algorithms, and the overall flow of the proposed financial risk early warning model is shown in Figure 2.

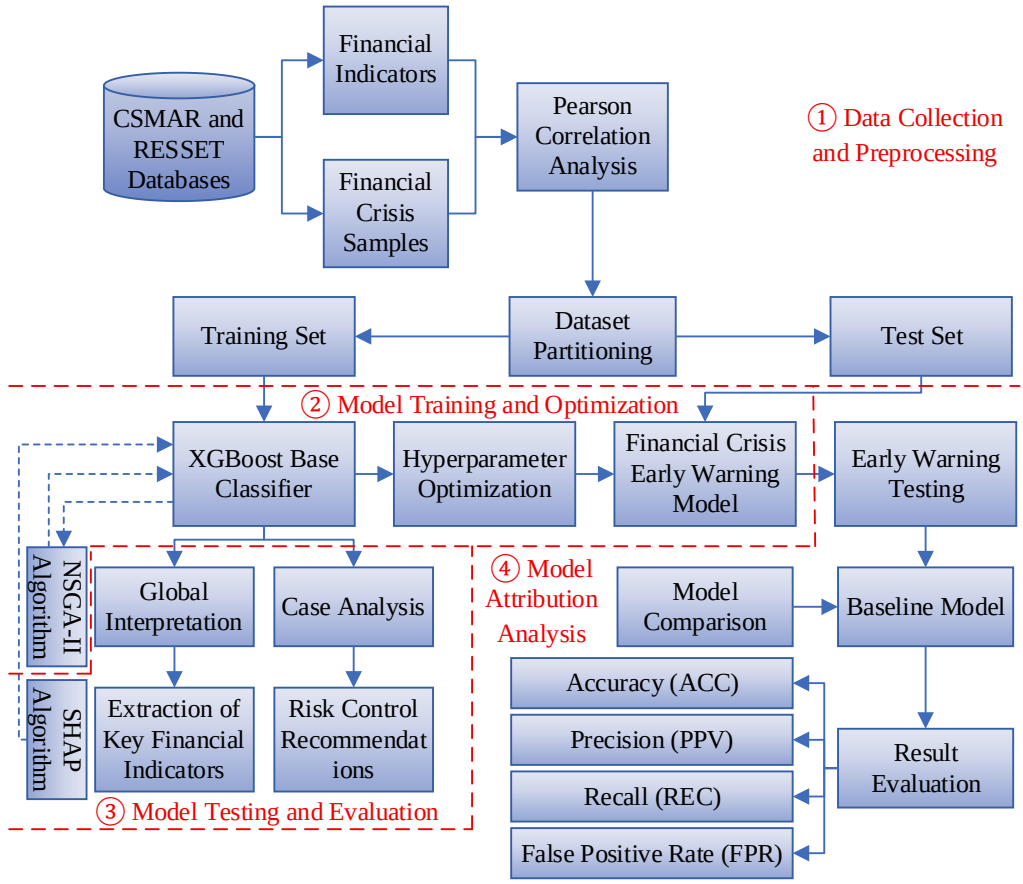


Figure 2: Flow chart of financial risk early warning model

The XGBoost algorithm is used as the base model for model training. Then, due to the numerous hyperparameters of the XGBoost model, the values of these hyperparameters will have a large impact on the warning performance of the trained model, so we simultaneously take the maximization of the warning accuracy and the maximization of the warning efficiency of the warning base model as the optimization objectives, and use the NSGA-II multi-objective optimization algorithm to optimize the hyperparameters of the discriminative base model.

(1) Optimization objective 1: Maximization of warning accuracy (ACC) index:

$$Obj_1 = \max\left(\frac{TP+TN}{TP+FN+FP+TN}\right) \tag{15}$$

For the first optimization objective of the NSGA-II multi-objective optimization algorithm, we choose the model early warning precision (ACC), which is calculated based on the results of the confusion matrix obtained from the model prediction, and this indicator represents the correctness of the model to correctly warn the samples with and without financial crises.

(2) Optimization objective 2: Maximization of early warning efficiency recall (Recall):

$$Obj_2 = \max\left(\frac{TP}{TP+FN}\right) \tag{16}$$

The second optimization objective of the NSGA-II multi-objective optimization algorithm that we chose is the maximization of Recall, Recall is also calculated based on the results of the model confusion, which measures the number of correct model warnings as a proportion of the

number of samples in which all the financial crises occurred, and is used as a measure of the efficiency of the model when it performs the early warning of financial crises.

4 Empirical results and analysis

4.1 Funding accounting data mining results

4.1.1 Data acquisition

The data are mainly obtained from the databases of Guotai Junan database, Wande database and National Bureau of Statistics, etc. Most of the indicators can be obtained directly from the above databases, and some other proportionality indicators need to be constructed, so that the basic data can be obtained from each of the above libraries, and the corresponding results can be obtained after the statistics and arithmetic operations of EXCEL. In this paper, data preprocessing, feature screening and model construction are partly realized by Python language, mainly involving Pandas, Numpy and Sklearn.

4.1.2 Data pre-processing

(1) Data exploration

This paper firstly needs to carry out exploratory analysis of the data, and the results show that some features belong to the category features need to be coded uniquely hot, some features belong to the number of variance is too large and need to be standardized, and all other features are numerical features. There is a significant difference between the very small amount and the very large amount of some features, and the variance of some features is especially significant. Therefore, standardization is required. In addition, the number of individual features also has its own variance, and for missing values, it also needs to be processed.

In realistic financial risk prediction, the weight of positive and negative types of samples is not balanced as only a small percentage of firms experience financial risks. In the sample of this paper, there are only 95 financial crisis samples, while there are 1995 total samples, and the difference in the number of financial abnormal samples and financial normal samples is too large, therefore, this paper needs to deal with the imbalance of the samples.

(2) Missing value processing

Even though the XGBoost model can deal with missing values accordingly, the accuracy of the model can be improved after taking appropriate and effective interpolation method to fill the missing values. In this paper, for the missing rate of less than 20% of the feature values, this paper selects the random forest filling method for processing; for the missing rate of more than 20% of the feature values, this paper is deleted. The specific missing situation of all data is shown in Figure 3.

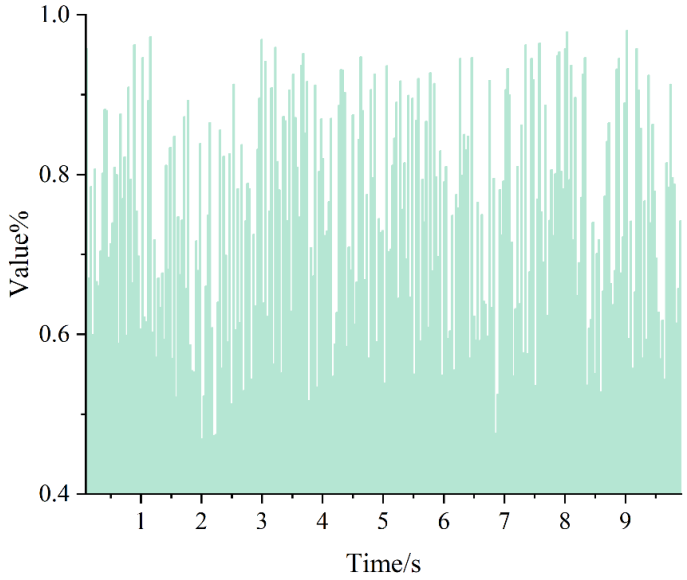


Figure 3: Missing values

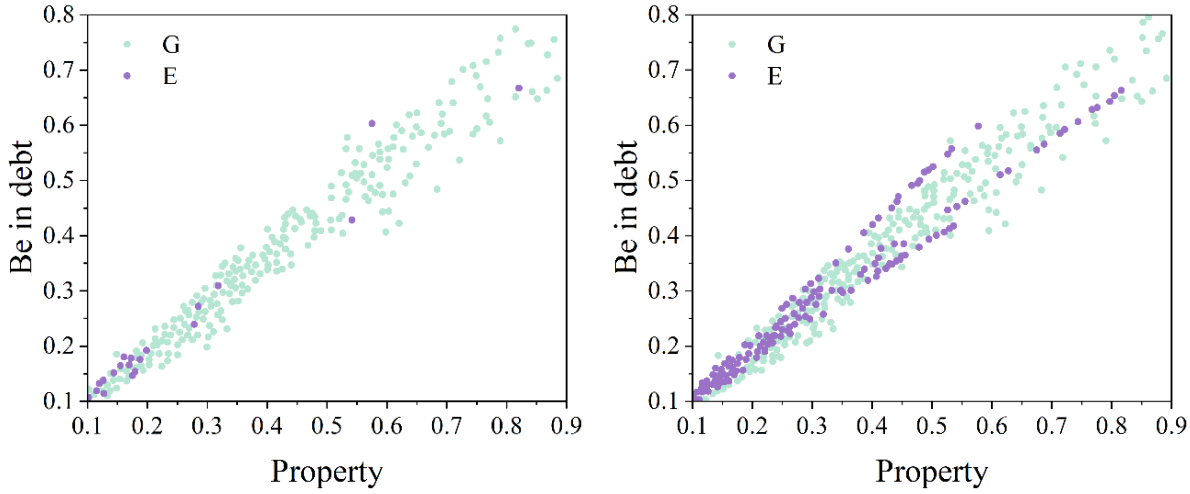
(3) Normalization

This subsection needs to select the Z-score method to standardize the processing for some features. The processed data conforms to the standard normal distribution with mean 0 and standard deviation 1, fluctuating up and down around 0. This step can be accomplished directly through the third-party sklearn library in python.

(4) Sample balance processing

In the specified time window, there are relatively few events in which financial crises occur. Therefore, this paper performs synthetic sampling through SMOTE+ENN. Interpolation is performed between the minority class samples to generate samples, and samples in the majority class samples where none of the K nearest neighbor points belong to the majority class are eliminated. This takes care of the overlapping samples generated by SMOTE. First, the random number random_state=7 is set, and the training set, validation set, and test set are randomly assigned in the ratio of 3:1:1, obtaining 1197 samples for the training set, 399 samples for the validation set, and 399 samples for the test set; and then, the training set, as well as the test set, is sampled with SMOTE-EN in order to facilitate the training of the model, and the total number of samples of the two amounts to 2067 respectively, 685 samples, respectively.

The two-dimensional comparison before and after sampling is shown in Figure 4, which presents the change of sample size in the form of a two-dimensional scatter plot, reflecting the original data and the sampled data from left to right, respectively. It can be clearly observed that the sample size increases, and the sample size increases mainly at the bottom as well as the edge position. On the one hand, the increasing samples in the bottom position indicate that financial risks are more likely to appear in micro and small enterprises, whose total assets and liabilities are lower; on the other hand, the samples appearing in the edge position indicate that financial risks are also likely to erupt in enterprises with disproportionate assets and liabilities ratios.



(a) Before sampling (b) After sampling

Figure 4: Comparison of samples before and after sampling

4.1.3 Characterization

Feature screening is an important part in constructing a model. Previous studies have often used the default feature importance method for feature screening. However, there are two drawbacks of the feature importance method: first, it prefers continuous variables because it is easier to find the cut-off point for continuous variables, in other words, it is easier to overfitting; second, its essence is the degree of dependence of the trained model on the variables, which doesn't represent the generalization ability of the variables on the unknown test set. In this paper, we choose the sorting method, the essence of which is to disrupt the original relationship between X and Y by randomly disrupting the variables. If disrupting a variable significantly increases the model's LOSS on the validation set, it means that the variable is important. If disrupting a variable has no effect on the model's loss in the validation set, or even decreases the loss, then the variable is not important to the model, or even harmful. In this paper, we repeat the screening 10 times by sklearn's own permutation_importance library to pick the features that have a large impact on the model.

4.2 Early-warning analysis of the risk of financing the construction of sheltered housing

4.2.1 Importance scores for model features

After tuning the hyperparameters of the XGBoost model, the financial risk early warning model was obtained, which got a score of AUC of 0.9122 on the training set. At the same time, this paper ranks the contribution of each feature to the model prediction results in accordance with the degree of contribution of each feature, and takes the top 10 features in terms of the degree of influence to be plotted, and the results are shown in Figure 5. As can be seen from the figure, the feature of return on capital (D1) has the highest ranking, as a split feature for 1057 times; followed by return on investment (D5), as a split feature for 974 times; The next eight higher ranked characteristics are: fixed capital turnover (C6), current ratio (A1), cash ratio (A3), gearing ratio (A5), operating profit as a percentage (B6), net profit growth rate (E5), return on investment (D5), and managed price-to-earnings ratio (F1).

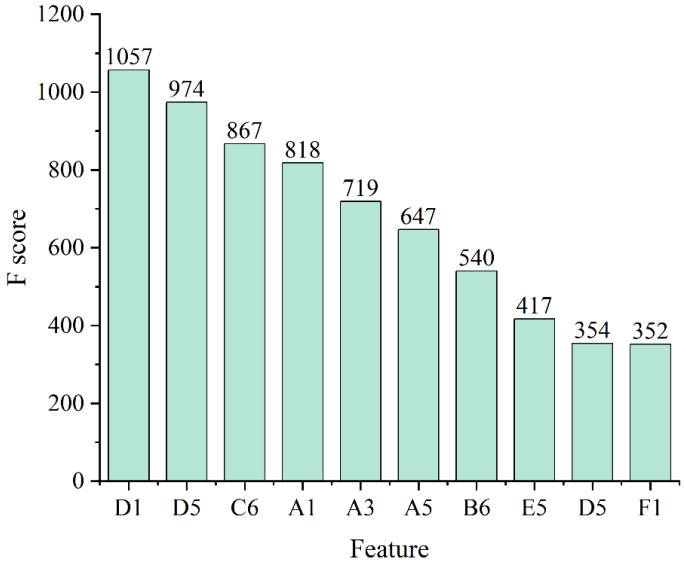
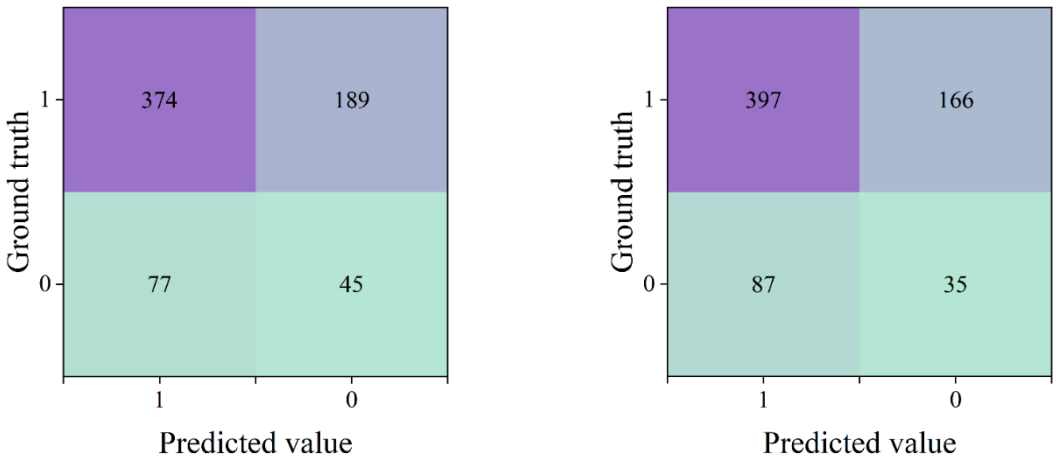


Figure 5: XGBoost model feature importance score

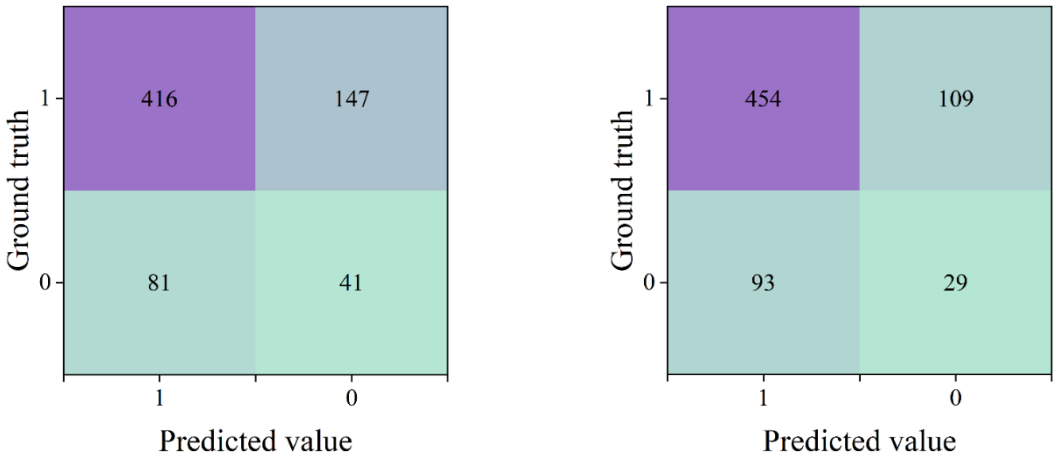
4.2.2 Assessment of the model's learning ability and generalization capacity

The purpose of this section is to compare financial risk early warning models based on different algorithms and evaluate multiple models based on four scoring metrics: accuracy (ACC), false positive rate (FPR), recall (REC) and precision (PPV). The risk early warning models constructed based on classification methods such as support vector machine (SVM), random forest (RF), and GBDT are selected as the reference models for model performance comparison.

In order to understand the performance differences of the above models, this paper judges the advantages and disadvantages by comparing the classification results of each classifier and the corresponding evaluation indexes. Figure 6 shows the confusion matrix of each classifier based on the test dataset, while Table 2 shows the results of four scoring indexes, namely, accuracy, false positive rate, recall and precision of the models. It can be found that the XGBoost algorithm-based stock crash warning model constructed in this paper outperforms other classification models in terms of accuracy (ACC), false-positive rate (FPR), recall (REC), and precision (PPV) metrics, and the highest value is 0.9821.



(a) RF (b) SVM



(c) GBDT (d) XGBbosst

Figure 6: Confusion matrices of each classifier

Table 2: Performance Comparison of Different Financial Risk Early Warning Models

Model	ACC	FPR	REC	PPV
SVM	81.09%	26.86%	80.94%	77.51%
RF	73.27%	34.51%	76.35%	59.04%
GBDT	85.15%	19.66%	89.58%	79.62%
XGBoost	93.43%	12.54%	98.21%	89.21%

4.3 SHAP-based model interpretability analysis

4.3.1 Single-sample interpretation

SHAP's force diagrams can be used to interpret the predictions of individual samples, where each indicator's Shapley value is treated as a “force” and used to show whether each indicator is contributing to an increase or decrease in the predicted value, while SHAP's waterfall diagrams are used to visualize the contribution of individual samples to the model's predictions. It shows the impact of each indicator in the sample as an arrow, and the number of each arrow indicates the contribution of that indicator to the predicted value, based on the principle of additivity of Shapley values to explain the model.

For a sample, its prediction result can be expressed as the sum of the Shapley values of all indicators plus the base value. For a binary classification problem, if the prediction result of a certain sample x is $f(x)$, the base value is y_{base}, and the Shapley value of the j th metric is $f(x_j)$, then the prediction result can be expressed as:

$$f(x) = y_{base} + \sum_{j=1}^{T} f(x_j) \tag{17}$$

where T is the total number of indicators, and $\sum_{j=1}^{T} f\left(x_j\right)$ is the computational process to be shown in the waterfall plot.

This article selects the first sample from the test set to draw the SHAP waterfall chart, as shown in Figure 7. From the figure, it can be seen that the predicted value of the sample is -1, and the baseline value is -0.43. After data standardization, the value of the feature "total asset turnover rate" is 12.842, which has a negative impact on the prediction result with an impact

intensity of -0.75. The value of the feature "investment return rate" is -0.487, which has a positive impact on the prediction result with an impact intensity of 0.18. Therefore, the financial risk warning model believes that this sample is in a financially healthy state. A higher "total asset turnover rate" is beneficial to the company's financial health, while a lower "investment return rate" is detrimental to the company's financial health.

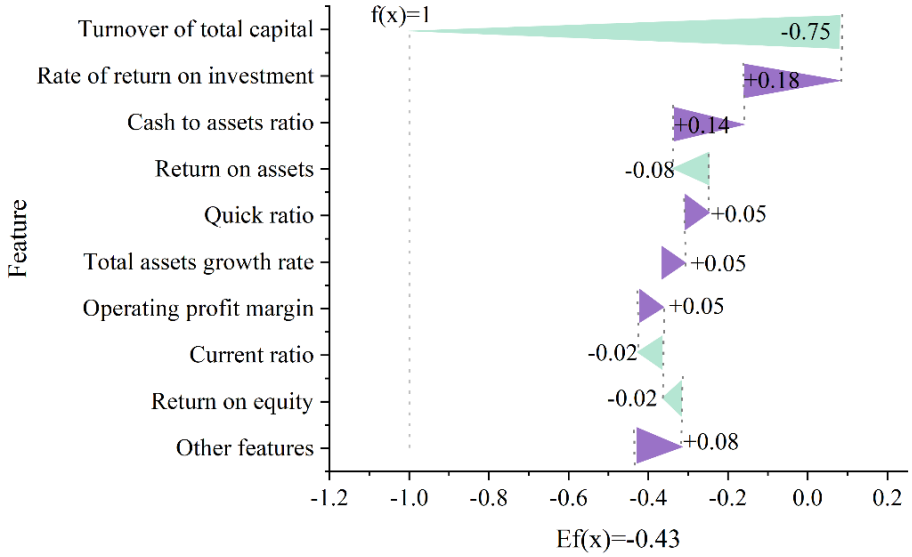


Figure 7: SHAP waterfall chart

4.3.2 Global interpretation

In terms of the local interpretability analysis of the SHAP model, for each sample of predictions, the model produces a predicted value, and the Shapley value for a given metric is the marginal contribution of that metric to the predicted value of the sample. The global interpretability analysis of the SHAP model, on the other hand, refers to interpreting the extent to which each metric contributes to the predicted output in the entire model over the entire dataset of the given model, specifically by using the Tree SHAP computation method, traversing each decision tree in the model, combining the samples from each leaf node, and then computing the Shapley value for that combination, i.e., the contribution of each sample in that combination to the contribution to the target output. In this process, each metric contributes to different combinations of samples, and the Shapley value is the average contribution of each metric in different combinations. The Shapley value corresponding to each indicator j is calculated as:

$$f(x_j)=\frac{1}{M}\sum_{S\subseteq\{1,2,..,N\},i\notin S}\frac{(N-|S|-1)!|S|!}{N!}\left[f_{S\cup\{j\}}(x)-f_S(x)\right] \tag{18}$$

where x is the sample to be interpreted, N is the number of samples in the dataset, M is the number of subsets that satisfy the condition $|S|<M$, $f_{S\cup\{j\}}(x)-f_S(x)$ is the difference in the output of the indicator subset S when indicator j is added to the subset with indicator j versus when it is not added to the subset with indicator j, and $f(x_j)$ is the average contribution value of indicator j over all possible indicator subset S, i.e., the average contribution of indicator j in all possible indicators, i.e., the Shapley value.

(1) Feature Importance

This paper calculates the mean of the absolute values of each feature's Shapley value to

measure the importance of individual features. The larger the mean, the more important the feature. The features are sorted from the most important to the least important, and are visualized as shown in Figure 8. Overall, "asset return rate" plays the most crucial role in predicting financial risks, averaging an increase of nearly 19 percentage points in the absolute probability of financial distress in the prediction. The second is "investment return rate".

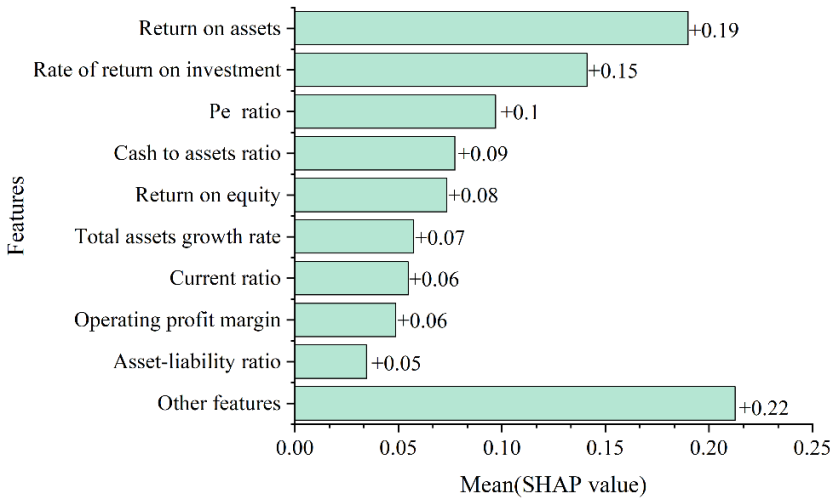


Figure 8: Shap feature importance of XGBoost model

The two indicators with the highest importance were then selected to visualize the classification effect of the financial risk prediction model, as shown in Figure 9. The sample points for financial risk are shown in purple color, while the sample points for financial health are shown in green color. The results show that the probability of a company's financial health increases when the net asset margin is greater than 1 and the earnings per share is greater than 0.5.

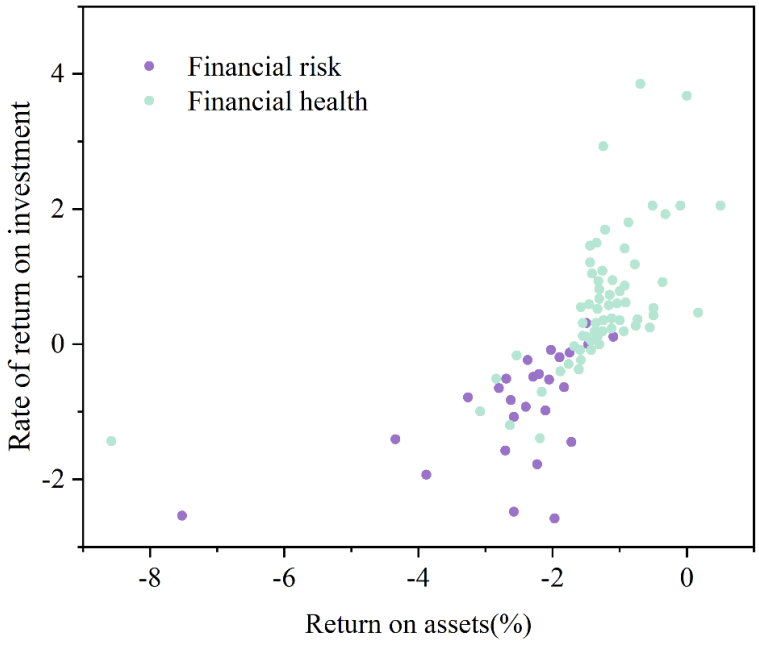


Figure 9: XGBoost model classification performance

Based on the above figure, we can conclude that the financial risk of financing housing

construction can be better distinguished by using only these two important indicators. This result not only verifies the importance of the two indicators in the financial risk early warning model, but also provides insights for the subsequent calculation of the threshold value of key features.

(2) Direction of the role of features

In order to comprehensively consider the importance and role direction of features, a summary diagram is drawn. In the summary plot, the vertical axis is to rank the features based on the sum of Shapley values of all samples, and the horizontal axis represents the Shapley values of the features, with each point representing a sample, as shown in Figure 10.

Take "asset return rate" as an example. A high "net asset return rate" (green) has a negative impact on the prediction results, while a low "asset return rate" (purple) has a positive impact. Specifically, in the fundraising for government-subsidized housing construction, those with a high "asset return rate" have less of their housing construction fundraising getting into financial risks (positive example), and more often exhibit financial health (negative example).

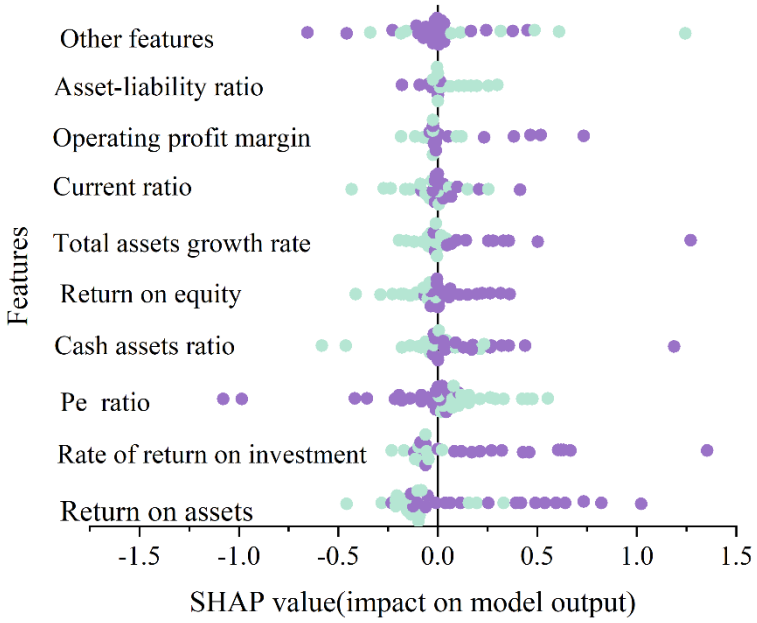


Figure 10: SHAP Summary Chart

Based on the above figure, this paper summarizes the direction of the role of important characteristics (top nine in terms of importance) on the financial health of the company as shown in Table 3.

“Return on Assets”, “Return on Investment” and “Operating Profit Margin” measure the profitability of housing finance. A higher “return on assets” means that housing construction financing is able to manage and utilize its assets efficiently and achieve a higher level of profitability with a relatively low investment in assets. Higher “return on investment” means that shareholders can get higher dividends or capital gains, indicating that housing construction financing has strong profitability. Higher “operating profit margin” means that housing construction financing can earn higher profits in providing services, which helps housing construction financing to operate and develop healthily. Therefore, the three indicators have a positive impact on the financial health of housing finance.

Table 3: The Impact of Key Features on the Financial Health of the Company

Order number	Index	Impact on financial health
1	Return on assets	Forward direction
2	Rate of return on investment	Forward direction
3	Pe ratio	Negative direction
4	Cash assets ratio	Forward direction
5	Return on equity	Forward direction
6	Total assets growth rate	Forward direction
7	Current ratio	Forward direction
8	Operating profit margin	Forward direction
9	Asset-liability ratio	Negative direction

(3) Early warning threshold of features

In this paper, we take “return on assets” as an example and draw a scatter plot to describe the relationship between the feature “return on assets” and its Shapley value, as shown in Figure 11. After normalizing the data, when the ROA is less than -0.1, this feature has a positive impact on the prediction results, i.e., it increases the probability of housing construction financing falling into financial risk. When the return on assets is greater than -0.1, the feature has a negative effect on the predicted outcome, i.e., it decreases the probability of housing construction financing being at financial risk.

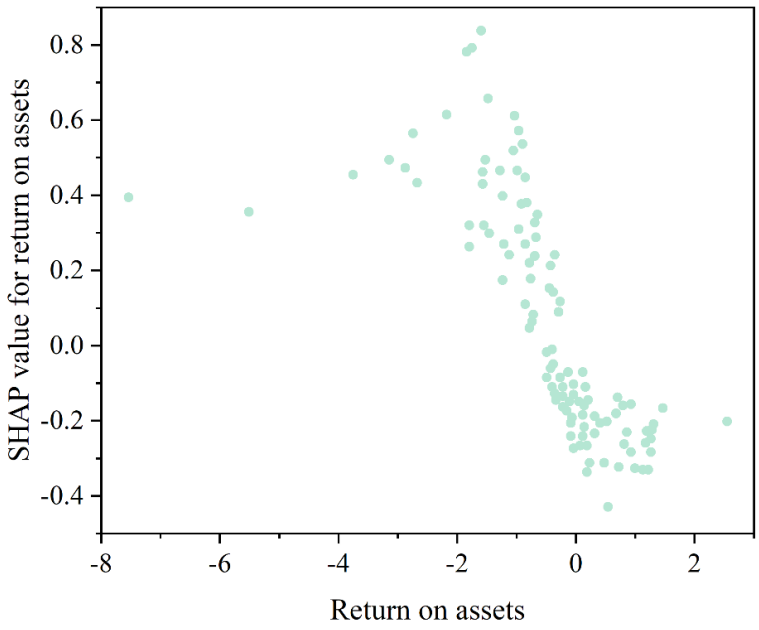


Figure 11: Scatter plot of return on assets

5 Conclusion

The accounting data mining algorithm based on decision trees processes the accounting data of housing construction fund raising, combines the XGBoost model to construct a risk warning model for housing construction fund raising, constructs a financial warning index system, and uses the NSGA-II multi-objective optimization algorithm to optimize the hyperparameters of the warning model. Using XGBoost, through sample over-sampling and parameter adjustment, the financial risk warning model obtained has an ACC score of 0.9343 and a REC score of

0.9821, which can effectively identify the risk characteristics in fund raising. At the same time, the prediction results of the model are visualized using the SHAP explanation framework. Combining the SHAP explanation framework and the importance ranking, it can be found that the indicators that have a positive impact on the company's financial health are: "asset return rate", "investment return rate", "cash asset ratio", "net asset return rate", "total asset growth rate", "current ratio", and "operating profit margin".

About the Author

Zhen Li was born in Zhengzhou, Henan, China in 1988. She obtained a bachelor's degree from Zhengzhou University in China. Her main research directions are financial risk management and accounting education.

References

[1] Jiang, H. T., & Ban, Q. C. (2013). A study on application of supportive housing in Chinese affordable housing. Applied Mechanics and Materials, 357, 2393-2397.

[2] Wang, Z., Zhang, H., Liu, S., & Chen, J. (2025). Strategic Interaction in the Supply of Affordable Housing Construction Land: Evidence from China’s Cities. Buildings, 15(10), 1684.

[3] Zou, Y. (2014). Contradictions in China's affordable housing policy: Goals vs. structure. Habitat International, 41, 8-16.

[4] Shi, W., Chen, J., & Wang, H. (2016). Affordable housing policy in China: New developments and new challenges. Habitat International, 54, 224-233.

[5] Gan, X., Zuo, J., Wu, P., Wang, J., Chang, R., & Wen, T. (2017). How affordable housing becomes more sustainable? A stakeholder study. Journal of Cleaner Production, 162, 427-437.

[6] Ling, C. S., Almeida, S. J., & Wei, H. S. (2017). Affordable housing: challenges and the way forward. BNM Quarterly Bulletin: Box Article, Fourth Quarter, 19-26.

[7] Zhang, T., & Hashim, A. H. B. (2011). Theoretical framework of fair distribution of affordable housing in China. Asian Social Science, 7(9), 175-183.

[8] Owens, A. (2015). Housing policy and urban inequality: Did the transformation of assisted housing reduce poverty concentration?. Social Forces, 94(1), 325-348.

[9] Howell, J., Whitehead, E., & Korver-Glenn, E. (2023). Still Separate and Unequal: Persistent Racial Segregation and Inequality in Subsidized Housing. Socius, 9, 23780231231192389.

[10] AlQahtany, A. M. (2022). Government regulation and financial support on housing delivery: lessons learned from the Saudi experience. International Journal of Housing Markets and Analysis, 15(3), 613-631.

[11] Ianchuk, S., Garafonova, O., Panimash, Y., & Pawliszczy, P. (2021). Marketing,

management and financial providing of affordable housing. Marketing i menedžment innovacij, (2), 213-230.

[12] Spaan, M., & Abraham, Y. S. (2023). Barriers to and enablers of affordable housing construction: insights from construction industry professionals. Engineering Proceedings, 53(1), 36.

[13] Akinsulire, A. A., Idemudia, C., Okwandu, A. C., & Iwuanyanwu, O. (2024). Strategic planning and investment analysis for affordable housing: Enhancing viability and growth. Magna Scientia Advanced Research and Reviews, 11(2), 119-131.

[14] Ashtari, M. A., Ansari, R., Hassannayebi, E., & Jeong, J. (2022). Cost overrun risk assessment and prediction in construction projects: A Bayesian network classifier approach. Buildings, 12(10), 1660.

[15] Osifo, E. O. (2024). Integrating compliance and cost control in public infrastructure and affordable housing construction contract management. International Journal of Engineering Technology Research & Management, 8(12), 439-443.

[16] Maleeva, T. V., & Selyutina, L. G. (2018, October). Analysis and evaluation of financial resources of social housing construction in city. In Materials Science Forum (Vol. 931, pp. 1118-1121). Trans Tech Publications Ltd.

[17] Mavlioutov, R. R., Egorova, E. V., & Pakhomova, O. Y. (2018, October). Financial maintenance of building of affordable housing on the basis of public-private partnership. In Materials Science Forum (Vol. 931, pp. 1107-1112). Trans Tech Publications Ltd.

[18] Reid, A. (2023). Closing the affordable housing gap: identifying the barriers hindering the sustainable design and construction of affordable homes. Sustainability, 15(11), 8754.

[19] Chan, A. P., & Adabre, M. A. (2019). Bridging the gap between sustainable housing and affordable housing: The required critical success criteria (CSC). Building and environment, 151, 112-125.