Research on the Model of Preventing Corporate Financial Fraud under the Combination of Deep Learning and SHAP

Yanzhao Wang*
Henan Institute of Economics and Trade, College of Finance,
Zhengzhou, 450000, China

***Abstract*—Preventing financial fraud in listed companies is conducive to improving the healthy development of China's accounting industry and the securities market, is conducive to promoting the improvement of the internal control system of China's enterprises, and is conducive to promoting stability. Based on the combination of SHAP (Shapley Additive explanation), a prediction and identification model should be built to determine the possibility of financial fraud and the risk of fraud for the company. The research model has effectively improved the identification accuracy of financial fraud in listed companies, and the research model has effectively dealt with the gray sample problem that is common in the forecasting model through the LOF algorithm and the IF algorithm. When conducting comparative experiments on the models, the overall accuracy rate of the research model is over 85%, the recall rate is 78.5%, the precision rate is 42%, the AUC reaches 0.896, the discrimination degree KS reaches 0.652, and the model stability PSI is 0.088, compared with traditional financial fraud Forecasting models FS model and CS model has a higher predictive effect. In the empirical analysis, selecting a company's fraud cases in 2020 can effectively analyze the characteristic contribution in the fraud process and the focus on fraud risks. The established model can effectively monitor the company's finance and prevent fraud.**

Keywords*—Financial fraud; deep learning; ensemble algorithm; feature selection

I. Introduction

With the rapid development of China's social economy, listed companies also occupy the majority of the market economy, and the quality of financial information of listed companies has an important impact on the efficiency of the capital market. The frequent occurrence of financial fraud in listed companies has seriously hindered the healthy development of the capital market, leading to the weakening of the company's control structure, reducing the effectiveness of corporate governance, and deteriorating the quality of its audit function [1-3]. At present, traditional financial fraud prevention and prediction systems are gradually unable to meet the needs of the market economy, and financial fraud prediction models are not yet sound. Relevant institutions and enterprises are difficult to effectively evaluate financial fraud phenomena, resulting in difficulties in effectively improving the financial risks of listed companies. How to enable different institutions and relevant enterprises to avoid financial fraud is the key to improving social economy and improving the market system. XGBoost integrated learning algorithm can more effectively predict financial fraud, which is currently an excellent classification algorithm. Therefore, in order to more accurately predict the existence of fraud in enterprises, a model based on deep learning combined with SHAP to prevent financial fraud in enterprises was studied, enabling regulators to conduct regulatory early warning for listed companies, creating a good ecological environment for the healthy growth of the capital market [4-6]. The research proposed combining ratio scaling with XGBoost model. The financial fraud prediction model can be quantitatively evaluated, and the design of a scoring card can not only have high prediction accuracy, but also have a negative sample capture rate. Derivative screening of the financial fraud risk indicator system, dimensionality reduction processing features make the model easy to use, combining different corporate governance perspectives with financial data as indicator variables. Reduce the cost of manual investigation, reasonably assess audit risks, and determine a more efficient audit scope. Reasonably evaluate the potential value and future growth space of different enterprises. At the same time, the model can be used to evaluate the potential fraud risks of debt companies, providing a more favorable analysis of the security of borrowing funds.

II. Related Works

Financial fraud means that the managers of the enterprise cover up the real financial status and cash flow of the enterprise by changing or falsifying the accounting information, so as to bring huge economic benefits to the fraudsters. How to predict or detect financial fraud has been researched and analyzed by various scholars. HWA et al. constructed a thinking map based on the relationship between enterprises and audit firms, and used the method of feature extraction to build a thinking framework to analyze whether enterprises have financial fraud [7]. Houssou et al. use homogeneous and non-homogeneous Poisson processes to detect financial fraud in imbalanced datasets. Experimental results show that applying the model to financial datasets shows better predictive ability than baseline methods especially in the case of high data imbalance [8].

The XGBoost algorithm is an integrated learning method, which is widely used in all walks of life, and is also deeply researched and developed by different professionals. C Zhao et al. established a prediction model of speed estimation and distance between other vehicles during vehicle operation by mixing neural network and limit gradient, and found that the model can help drivers effectively predict the speed and distance of other vehicles when changing lanes in different scenarios, with good prediction effect [9]. W niu et al. integrated a gradient lifting algorithm into computer traffic monitoring to protect computer security and monitor hacker attacks. XGBoost is mainly used to distinguish whether there is malicious traffic. The performance of the established model can effectively distinguish between malware traffic and normal traffic from normal data sets in actual use, and the false alarm rate is less than 1% [10]. Bao et al. combined AdaBoost ensemble learning with under sampling data processing methods, and introduced a model evaluation index that is more suitable for fraud prediction tasks. Experimental results prove that the model is superior to the financial ratio-based regression model and single-core support vector machine model in predicting financial fraud [11].

To sum up, in the research on the prediction and detection of financial fraud, there is relatively more research on the detection of financial fraud. However, there is little research on the prediction of financial fraud and economic fraud. The SGBoost algorithm is widely used in engineering, Internet security, data set classification, and system modeling, but there are relatively few related studies on the application of the XGBoost algorithm to predict financial fraud. Therefore, research on deep learning combined with SHAP to prevent corporate financial fraud models, Using Benford's law, LOF local anomaly detection, and isolated forest detection method to eliminate gray samples in the prediction model, establish a financial fraud prediction model for listed companies based on machine learning methods, and create a good and healthy ecological environment for the capital market.

III. Construction of Financial Fraud Prevention Model based on Deep Learning and SHAP

A. Fraud Prediction Model Index Construction

Financial fraud hides or changes the relevant financial status, operating conditions and capital flow of the enterprise through forgery, omission, fabrication, etc., which can bring huge economic benefits to the fraudsters. Therefore, choose the appropriate financial fraud early warning indicators to establish a suitable model foundation. In the process of selecting indicators, follow the principles of reference, systematization, operability, and loose before tightening to make the model have good versatility [12-14]. The premise of an efficient and accurate deep learning model is to select good financial fraud characteristic variables, and the index selection is shown in Fig. 1.

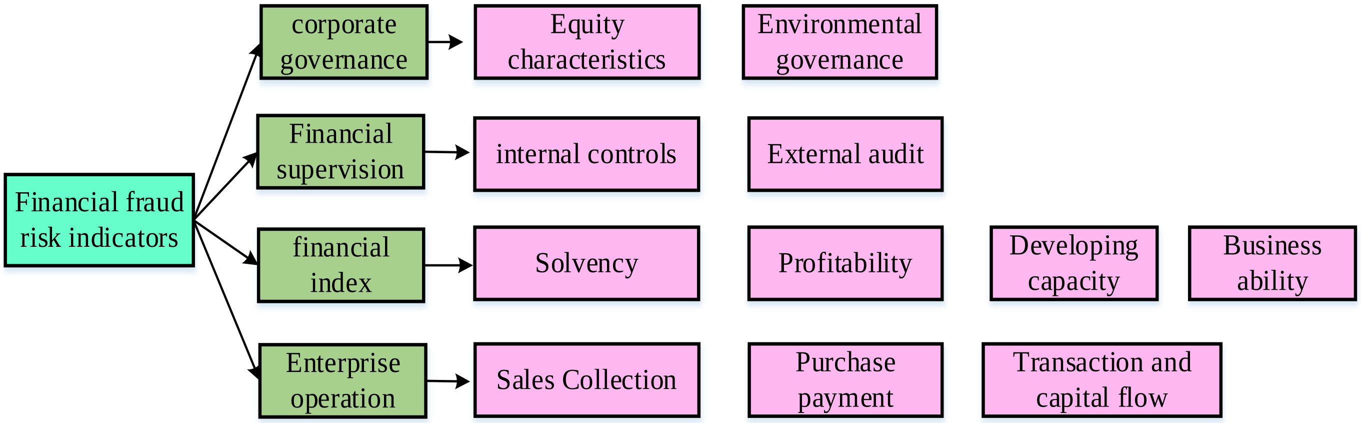


Fig. 1. Financial fraud risk indicator system.

Fig. 1 shows the financial fraud risk indicator system. Corporate governance indicators reflect the level of managers within the company and the degree of internal checks and balances. The secondary indicators are divided into equity characteristics and environmental governance. Financial and accounting supervision indicators divide supervision into internal control and external audit. Financial fraud is more likely to occur in companies that lack effective internal control systems. External audit indicators, such as firm size, auditors and other external factors, examine the factors affecting financial fraud. Financial indicators reflect the development of the company's economy, solvency reflects an important indicator of the company's long-term operation and development, and profitability is an indicator for evaluating the company's created value and company value. Operating ability reflects the turnover of different assets of the company and the level of management. Development ability is the basis for measuring whether the company has good growth potential in the future. The cash flow cycle is a risk indicator that reflects whether the company has abnormal operations. The sales collection indicator reflects whether the company's revenue is normal, the purchase payment indicator reflects whether the credit purchase and cash purchase at the purchasing end of the enterprise are normal, and the transaction and capital flow reflect whether the company has engaged in financial fraud through related transactions. After determining the financial fraud risk indicator system, research and select appropriate financial fraud samples. The financial fraud samples for this research are selected from the CSMAR China listed company financial annual report database and the violation information summary table of the violation event database, and select the handling documents issued by the regulatory agency. The data of illegal companies involving "fictitious profits" and "fictitious assets" between 2010 and 2021 are used as financial fraud samples in this paper. The financial fraud sample noun is divided into three different training sets before building the model: training samples, test samples, and out-of-time samples. The ratio of training samples to testing samples is 7:3. The out-of-time samples are the last samples of the time slice among all the samples. The training samples are used to train the model, while the test and out-of-time samples are utilized to evaluate the model's capability, stability, and generality [15]. In order to make the model more flexible and stable, and avoid the impact of extreme data values on the model, feature discretization is selected. The discretization formula is shown in formula [16-18] (1).

$$woe_i = \ln(\frac{p_{y_i}}{p_{n_i}}) \tag{1}$$

In formula (1), p_{y_i} is the ratio of fraudulent samples in the current group to all fraudulent samples in the sample, is the ratio p_{n_i} of all non-fraudulent samples in the current group to non-fraudulent samples in all samples, and woe is the difference between the two ratios, the difference Expressed logarithmically. Ensure the operational efficiency of the model and avoid overfitting, the overall features are screened. In the preliminary screening process, after determining the feature missing rate, the distribution difference of the sample labels with different values of the feature is measured. The calculation formula is shown in formula (2).

$$IV = \sum iv_i \tag{2}$$

In formula (2), iv_i the expression of is shown in formula (3).

$$iv_i = (p_{y_i} - p_{n_i})woe_i \tag{3}$$

In formulas (2) and (3), it IV represents the amount of feature information, which can reflect the contribution of a single feature to label distinction. When the amount of feature information is too large, it means that the model simulation effect is excellent. The characteristic correlation calculation between two characteristic variables is measured by the Spearman correlation coefficient. The measurement formula is as follows.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{4}$$

In formula (4), it n represents the quantity of data, which d_i is the difference between the order of two feature data. In feature correlation screening, feature indicators with a correlation greater than 0.7 are eliminated. The calculation formula for calculating the distribution difference of the same index on different data sets, measuring characteristics and model stability is shown in formula (5).

$$PSI = \sum_i (p_{t\arg et}^i - p_{base}^i) * \ln(\frac{p_{t\arg et}^i}{p_{base}^i}) \tag{5}$$

In formula (5), PSI is the characteristic stability, $p_{t\arg et}^i$ is the ratio of the samples in the first box of the target set to the i total samples, and is the total proportion of samples in the p_{base}^i first box of the basic distribution i. In the characteristic stable value, the characteristic items less than 0.02 are eliminated. After screening the information quality of the features, Use Chichi information criterion and Bayesian information criterion to select the characteristics of the model to ensure that the model has sufficient complexity and data set description. The calculation formula of AIC is as follows.

$$AIC = 2k - 2\ln L \tag{6}$$

The calculation formula of BIC is as follows.

$$BIC = 2\ln n - 2\ln L \tag{7}$$

In formula (6) and formula (7), k is the number of model parameters, n is the number of samples and L is the likelihood function. The fraudulent companies in the non-fraudulent samples are called "gray samples". Eliminating gray samples can ensure the quality of training data of the deep speech learning model and improve the reliability of the model; balance problem. The study uses Binford’s law, LOF local anomaly factor method and isolated forest algorithm (IF) to eliminate gray samples.

The calculation formula of probability distribution and correlation coefficient in Binford’s law is shown in formula (8).

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}} \tag{8}$$

In formula (8), $\mathrm{Cov(X,Y)}$ is the covariance, $Var(X)$ and $Var(Y)$ is the variance respectively. In Binford’s hypothesis test, the chi-square test is selected to test the degree of conformity of the numerical distribution of the observed samples. The calculation formula of statistics is shown in formula (9).

$$\chi_n^2 = \sum_{i=0}^{9} \frac{(F_n(i) - F_0(i))^2}{F_0(i)} \tag{9}$$

In formula (9), $F_n(i)$ is the actual observed value, and $F_0(i)$ is the theoretical distribution value of the law. The LOF anomaly detection method introduces local reachability density to measure the degree of sample anomaly, which can output anomaly scores very well and has strong interpretability. The calculation formula is shown in formula (10).

$$lrd_k(p) = \frac{1}{\dfrac{\sum_{o \in N_{k(p)}} reach\_dis_k(p,o)}{|N_k(p)|}} \quad (10)$$

In formula (10), p and o are two different sample points, and the reachable distance is $reach\_dis_k(p,o)$, which $N_k(p)$ means that the distance between the sample points is less than or equal to the k distance between the nearest point and the sample point. The local anomaly factor is calculated using the following formula.

$$LOF_k(p) = \frac{\dfrac{\sum_{o \in N_{k(p)}} lrd(o)}{|N_k(p)|}}{lrd(p)} \quad (11)$$

The IF algorithm is an integrated algorithm is based on the idea of random partitioning of space. The principle of a single isolated tree in the two-dimensional space of the IF algorithm is shown in Fig. 2.

The blue point in Fig. 2 means normal sample, and the outlier point is indicated by red. The straight lines that divide the space are randomly selected along the two coordinates respectively. The probability of the red dot being isolated before the blue dot is much greater than that of the blue dot. The probability of being isolated before the red point is uncertain, however, the presence of a single tree can increase the chance. Therefore, the combination of multiple isolated trees can enhance the stability of the model, leading to the final IF model. The final IF abnormal score is defined as formula (12).

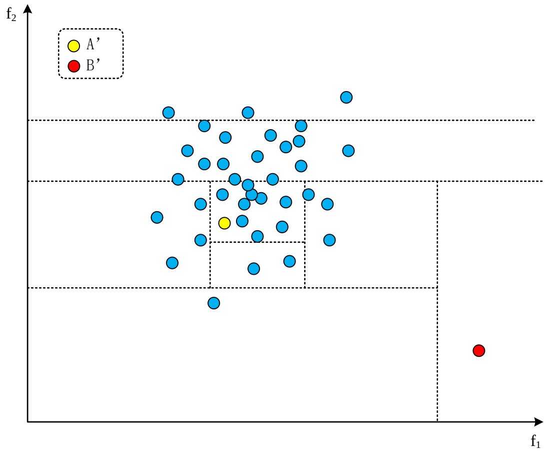


Fig. 2. Schematic diagram of IF algorithm.

$$Score(x_i) = 2^{\frac{E(h(x_i))}{C(n)}} \quad (12)$$

In the formula (12), it $E(h(x_i))$ represents the average path length of the blue point on the isolated tree, n is the number of training samples, and $C(n)$ is n the average path length of the binary tree trained by samples. After the gray sample is proposed, SMOTE is selected for oversampling. The basic principle of SMOTE is retracted as shown in the Fig. 3 below.

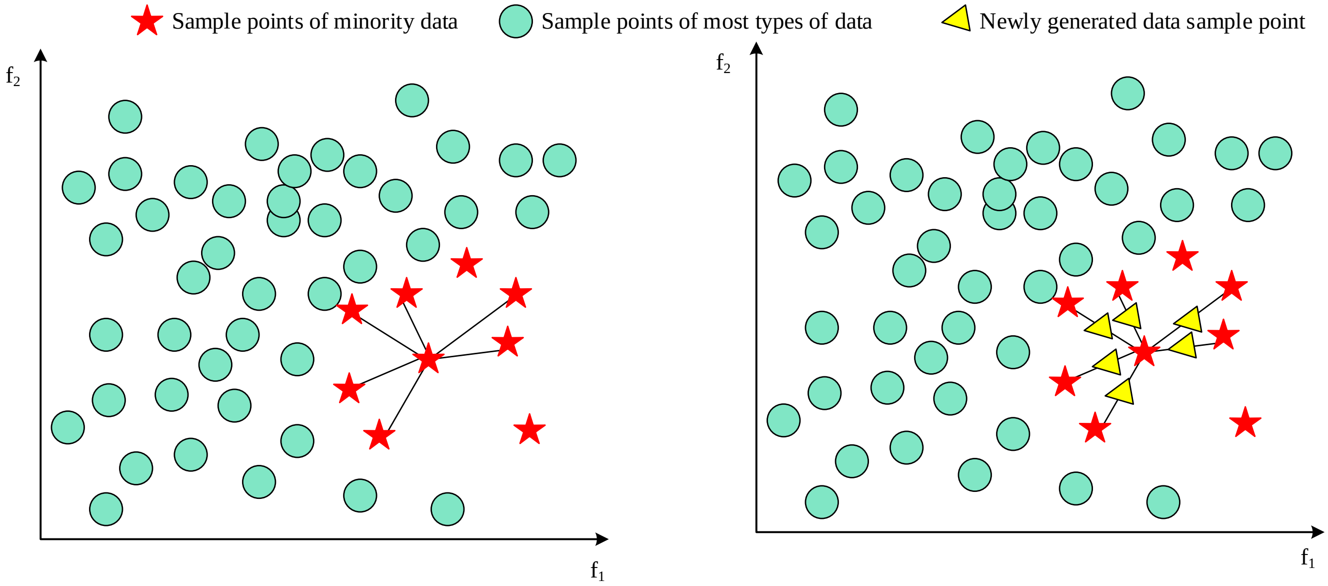

Sample points of minority data
Sample points of most types of data
Newly generated data sample point
f₂
f₁
f₂
f₁

Fig. 3. Schematic diagram of SMOTE algorithm.

SMOTE algorithm interpolates and synthesizes existing minority samples by analyzing minority samples, and then adds the samples to the dataset for training. The Light GBM algorithm was selected in this study to clean the samples. This was achieved by reducing the weight of samples with poor head and tail prediction results, so that they would not participate in the interpolation process of the SMOTE algorithm. The SMOTE algorithm has a high reliability for classification and oversamples the samples. Finally, the sampling results are combined.

B. Construction of Financial Anti-Fraud Model based on Deep Learning

After preprocessing the data, establish an anti-fraud model, and use the XGBoost model to improve the gradient boosting regression tree (GBRT) algorithm. The XGBoost model can support CART trees and linear classifiers at the same time. The study method is based on the forward distribution algorithm to achieve the integration of the additive model. The iterative process of the traditional GBRT algorithm is shown in formula (13).

$$\hat{y}^{(T)} = v\sum_{j=1}^{T} f_j(x;\Theta_j) = \hat{y}^{(T-1)} + vf_T(x;\Theta_T) \tag{13}$$

In formula (13), T is the number of basic regression trees, Θ_j is the corresponding regression tree structure, v is the scaling weight factor, $\hat{y}^{(T)}$ is the prediction result of the regression tree, and $f_j(x;\Theta_j)$ is the output result when the scaling weight is not considered.

The gradient boosting process of the XGBoost model is shown in Fig. 4.

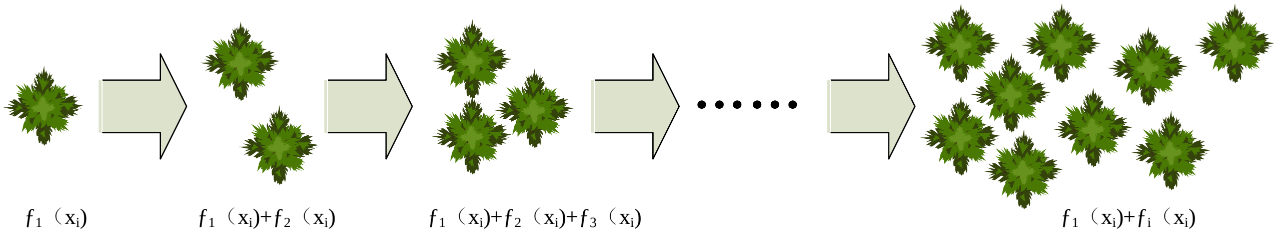


Fig. 4. XGBoost model principle.

The principle of a regression or classification model based on gradient lifting is shown in Fig. 4: First, establish a tree, and then gradually iterate. Each iteration process adds a tree, gradually forming a strong evaluator that integrates multiple tree models. The XGBoost model based on the tree model approximates the negative gradient of the model through the second-order Taylor expansion of the loss function, and learns it as the residual of the previous model. A higher learning weight is assigned to the samples with insufficient accuracy during the previous training process to improve the model accuracy, and serial iterations of multiple different models are implemented to gradually correct the deviation until the loss meets the convergence condition. The objective function after introducing the second-order Taylor expansion of the loss function is shown in Formula (14).

$$\hat{\Theta}_J \approx \arg\min_{\Theta_J}\left\{\sum_{i=1}^{N}\left[\hat{y}_j^{(j-1)} - y_i + vf_j(x_i;\Theta_j)\right]^2 + \Omega(\Theta_j)\right\} \tag{14}$$

In formula (14), N is the sample size, y is the minimum loss per additional node branch, and y_i is the given sample, and $\hat{y}^{(j-1)}$ has been determined. Therefore, $L(y_i, \hat{y}_j^{(j-1)})$ can be considered as a constant term to offset, and it is the regular term of the $\Omega(\Theta_j)$ th regression tree of j. After finding all the optimal single regression trees, the training is completed for enterprise fraud prevention prediction. The principle of $\Omega(\Theta_j)$ is shown in Formula (15).

$$\Omega(\Theta_j) = \gamma M_j + \frac{1}{2}\lambda\sum_{k=1}^{M_j}(w_k^{(j)})^2 \tag{15}$$

IV. Model Analysis and Empirical Application of Preventing Corporate Financial Fraud under Deep Learning

A. Model Prediction Effect Analysis

Effectively evaluate the actual put-to-use influence of the research model, the Fscore (FS) model and the Cscore (CS) model were selected to carry out a control experiment with the research model (RS), and the out-of-time samples were selected to test the model to test the stability of the model. In the control experiment, the application effects of different methods are compared by matrix. The matrix results are shown in the Table I.

TABLE I. Research Model Prediction Results

/	RS		CS		FS	
	Forecast		Forecast		Forecast	
Actual	Negative	Positive	Negative	Positive	Negative	Positive
FALSE	3146	513	2004	1763	1265	2574
TRUE	130	457	169	310	155	252

The research model correctly predicted 3146 cases of non-fraud samples and 457 cases of fraud samples, which is the model with the most correct prediction rate among the three models. The CS model correctly predicted only 2004 cases of non-fraud samples. Only 310 fraud samples were predicted, the FS model has the lowest prediction accuracy, only 1265 non-fraud samples were correctly predicted, and only 379 fraud samples were correctly predicted. The overall accuracy of the research model reaches 85%, the overall accuracy of the CS model reaches 54.5%, the prediction accuracy of FS model is relatively lower, reaching 36%. When analyzing the confusion matrix, we need to pay attention to the accuracy of the model, but also need to comprehensively evaluate the recall rate and accuracy rate of the model. The recall rate indicates the proportion of all companies that the model correctly predicts fraudulent companies, and the precision rate reflects the credibility of the model. The recall rate-precision rate results are shown in Fig. 5.

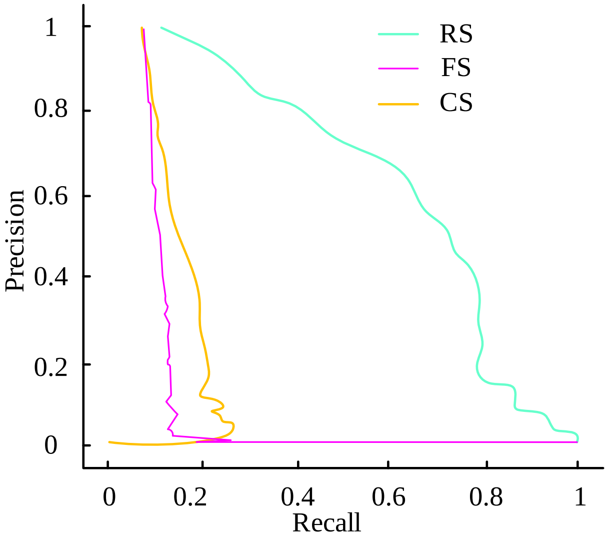


Fig. 5. PR curves of three models.

From the recall rate-precision rate result graph, it can be seen that the PR curve of the research model is the best, and the AP area is the largest, reaching 0.64. The precision rate and recall rate of the research model on samples out of time are higher than those of the CS model and the FS model. The indicators are superior, the AP area of the CS model is 0.16, and the AP area of the FS model is the smallest, which is 0.11, and the FS model is a commonly used forecasting model in the western capital market, and it has the worst forecasting effect on Chinese companies. The ROC curve can simply and intuitively observe the accuracy of different experimental models and make judgments through illustrations. This curve can accurately reflect the internal specificity relationship of the model and is a comprehensive representation to ensure the accuracy of the model. The results of the ROC experiment are shown in Fig. 6.

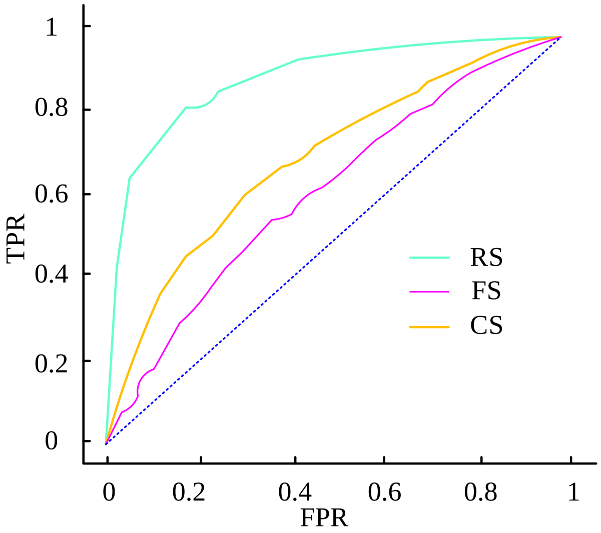
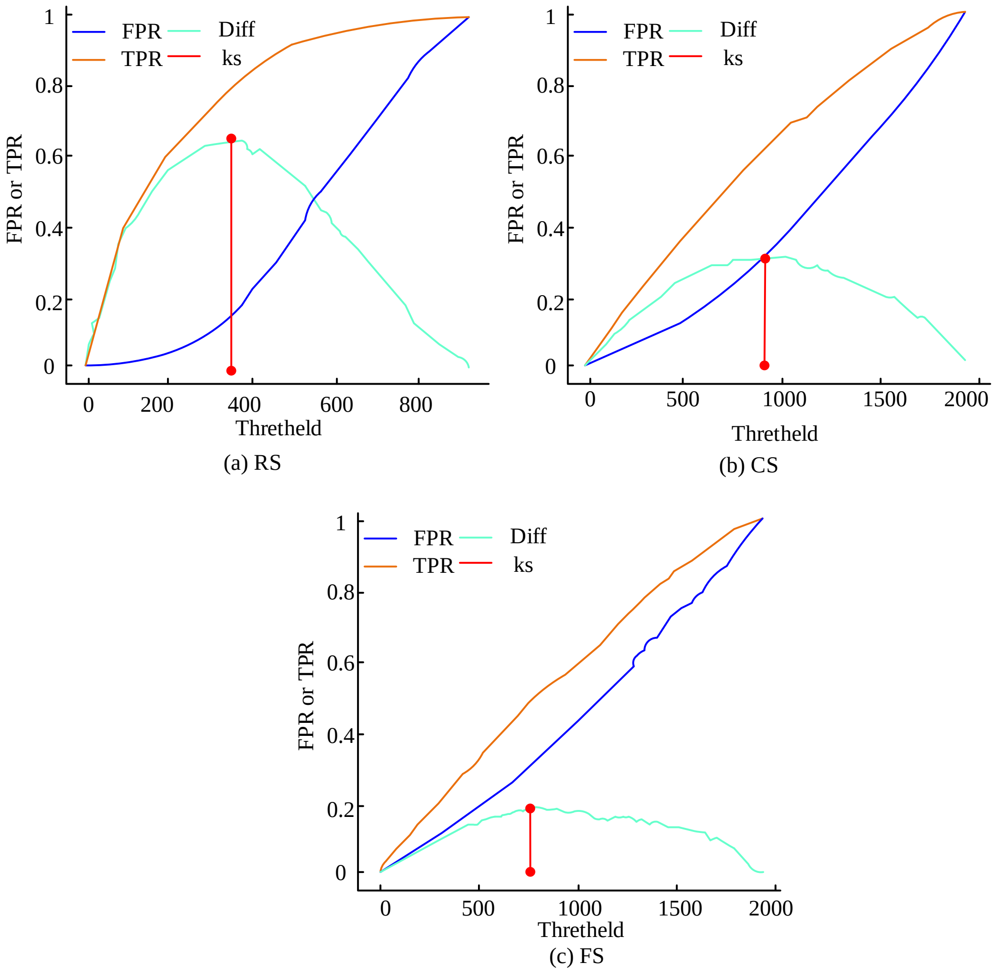


Fig. 6. ROC curves of different models.

When the area under the curve of the pure random classifier is 0.5, the larger the area under the ROC curve, the more realistic the fitting effect of the model. The research model's ROC curve performs best among the three models, with the largest area under the curve reaching 0.9. The ROC curve of the CS model is lower than that of the research model, and the area under the curve is only 0.7. The FS model had the lowest curve magnitude and the smallest area under the curve, reaching a value of 0.62. Combining Fig. 5 and Fig. 6, the research model can achieve a high accuracy rate when predicting fraud samples. The KS curve reflects the distinguishing ability of the model. After the model predicts the scores of all samples, it is divided into different parts according to TPR and FPR, and the distribution of the scores of different sample groups is tested by KS statistics. The KS curves of the three models are shown in Fig. 7.



(a) RS

(b) CS

(c) FS

Fig. 7. KS curves of three models.

Fig. 7 shows the comparison of KS curves for different models. The abscissa is arranged in descending order with thresholds of 1-0, and the ordinate is the difference between TPR and FPR under different thresholds. The higher the KS value is, the stronger the distinguishing ability of the model is. However, if it is too large, there is the problem of over-fitting. It can be seen from the above figure that the KS value of the research model is 0.65, which is within the range of the ideal model and has excellent positive and negative sample differentiation ability. Although the KS value of the CS model is in the ideal state, it is 0.35 lower than the KS value of the research model. The FS model has the worst ability to distinguish positive and negative samples among the three samples, only 0.18, and has a lower prediction effect on Chinese companies. The out-of-time sample test results of the three models combined with different evaluation indicators are summarized in Table II.

TABLE II. COMPREHENSIVE TEST RESULTS OF DIFFERENT MODELS

Model	Accuracy	AUC	Recall	Precision	AP	KS	PSI
RS	0.859	0.896	0.785	0.422	0.641	0.6 52	0.088
CS	0.545	0.697	0.564	0.143	0.159	0.30 5	0.115
FS	0.357	0.619	0.564	0.102	0.11 5	0.18 2	0.116

From the comprehensive test results in Table II, the accuracy of the research model reaches 0.859, the accuracy of the CS model is only 0.545, and the accuracy of the FS model is only 0.357. In AUC, the research model reaches 0.896, while the AUC of CS and FS Both are lower than the research model, reaching 0.697 and 0.619 respectively. In terms of recall rate, the recall rate of the research model reaches 0.785, the recall rate of the CS model is only 0.697, and the recall rate of the FS model is only 0.619. In terms of accuracy, the two models compared in the experiment are lower than the research model, the research model reaches 0.422, the CS model is only 0.142, and the FS model is only 0.102. In the PR curve comparison, the research model has the highest AP among the three models, which is 0.641, the CS model is only 0.159, and the FS model is only 0.115. Also in the KS value comparison, the research model reached 0.652, which has a good discrimination ability, the KS value of the CS model is only 0.305, and the FS model is only 0.182. PSI measures the distribution difference between test samples and model training samples. When PSI is less than 0.1, it can be considered that the model stability is very high. When PSI is 0.1-0.2, the stability of the model is average. When PSI is greater than 0.2, the model is stable. It can be seen that the PSI of the research model is the lowest among the three models, only 0.0883, and the stability is the highest, while the PSI of the CS model and the FS model are both higher than 0.1, and the stability is average, so the research model has a strong generalization ability.

B. Practical Application and Analysis of the Model

In the practical application and analysis of the research model, the SHAP method is used to explain the research model by using the SHAP library of Python, and a real fraud case in the training set is selected for analysis and prediction. The prediction results of the research model can be visualized with a single-sample SHAP graph. In the SHAP diagram, blue indicates that the contribution of the feature is negative, and orange indicates that the contribution of the feature is positive. The longer the orange color, the higher the probability of the predicted result being fraud. The specific results are shown in Fig. 8.

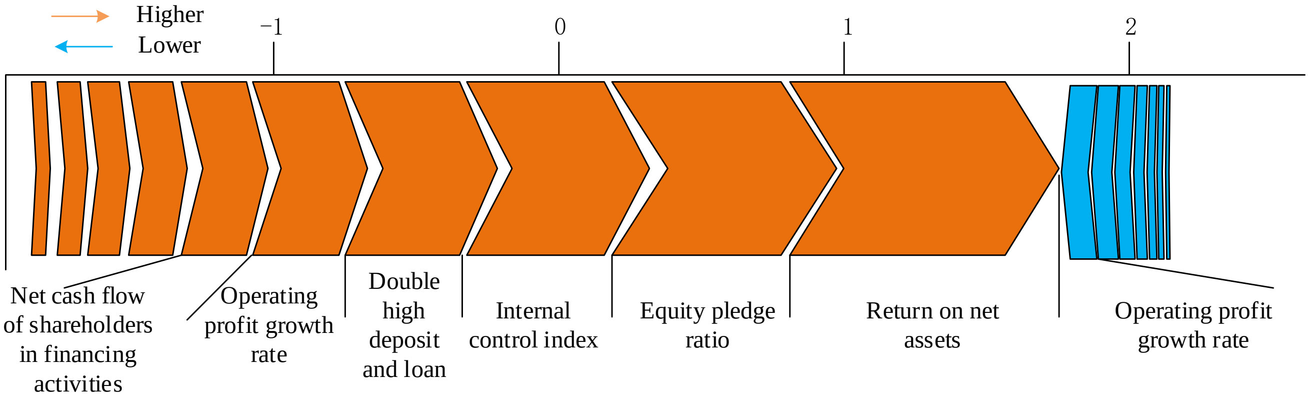


Fig. 8. SHAP chart of X enterprise’s research model forecast in 2020.

Fig. 8 shows the prediction results of the financial fraud scorecard. To improve the usability of the research and prediction model, a credit scoring mechanism is introduced into the model, and the model output is normalized through a unified scoring map to make the financial fraud risk readable. Among them, the high-risk score is ranked according to the feature contribution degree from high to low. To analyze the degree of dependence on related transactions, etc. In the prediction results of Company X in 2020, the important contributing variables are the proportion of equity pledges, internal control index, and high deposit and loan ratios, which reflect that Company X has risks in funding sources, risks in financial and accounting supervision, and abnormal business operations. The annual financial report disclosed by Company X contains major false information. First, there is a false increase in deposits. When there is a huge amount of monetary funds in the financial statements, there is a shortage of funds at the same time, and there is a phenomenon of high deposits and loans. Secondly, the forged business certificates lead to an inflated rate of return on net assets; in addition, part of the funds is transferred to the accounts of related parties for stock transactions, reflecting the illusion of a closed-loop capital flow. In addition, in terms of financial and accounting supervision, in the annual report, Company X reported that the company had major deficiencies in internal control, and its internal control audit agency maintained a negative audit opinion on the company’s internal control in 2020 [19-22].

V. Conclusion

With the development of the social economy, the quality of financial information in market information is more and more important to the efficiency of the market economy, but the problem of financial fraud is the most serious problem affecting the social economy. How to accurately predict financial fraud for listed companies is becoming more and more important. In this study, the integrated learning in deep learning is used to construct a risk warning model with more complete identification, more accurate and more robust, to score the financial fraud risk of listed companies, and to judge the prediction model of the company’s financial fraud possibility. The results show that the research model effectively solves the gray sample problem commonly faced when machine learning is applied to financial fraud identification research. The research model can improve the prediction effect of the model prediction. The overall accuracy rate of the research model is over 85%, and the recall rate is 78.5%. The accuracy rate reached 42%, AUC reached 0.896, the discrimination KS reached 0.652, and the model stability PSI was 0.088. Compared with the traditional financial fraud prediction models FS model and CS model, it has a higher prediction effect. In the empirical analysis, choose a certain company. The analysis of fraud cases in 2020 can effectively analyze the characteristic contribution in the fraud process and the focus of fraud risks. The research model is more suitable for the financial fraud prediction of listed companies in my country. However, in view of the availability of data, the research model does not sufficiently screen text analysis variables, there may be missing variables, and no specific analysis is carried out for the characteristics of different industries. The model also has a certain room for improvement, hoping to provide direction for future research.

REFERENCES

[1] H. Xia, H. Ma, P. Cheng. PE-EDD: An efficient peer-effect-based financial fraud detection approach in publicly traded China firms. CAAI Transactions on Intelligence Technology, 2022, 7(3):469-480.

[2] L. Liao, G. Chen, D. Zheng. Corporate Social Responsibility and Financial Fraud: Evidence from China. Accounting and Finance, 2019, 59(5):3133-3169

[3] Y. Jiang, Y. Zhao. Financial fraud contagion through board interlocks: the contingency of status. Management Decision, 2020,58(2):280-294.

[4] M. Wang, W. Zhao, W. Zhang, "Can Reform of Information Disclosure by an Exchange Restrain Corporate Fraud? Evidence from China." Asia-Pacific journal of financial studies, 2022,51(2):223-255.

[5] Z. M. Sanusi, A. Hudayati, T. K. Nisa. "Financial Pressure and Related Party Transactions on Financial Statements Fraud: Fraud Triangle Perspective." International Journal of Business and Emerging Markets, 2022, 14(2):213-230.

[6] R. Cao, G. Liu, Y. Xie, C. Jiang. "Two-Level Attention Model of Representation Learning for Fraud Detection." IEEE transactions on computational social systems, 2021,8(6):1291-1301.

[7] B. Hwa, C. Yca, D. Jl, "XZA Envelope. Financial fraud risk analysis based on audit information knowledge graph." Procedia Computer Science, 2022, 199:780-787.

[8] R. Houssou, J. Bovay, S. Robert. "Adaptive Financial Fraud Detection in Imbalanced Data with Time-Varying Poisson Processes." Journal of Financial Risk Management, 2019, 08(4):286-304.

[9] C. Zhao, X. Zhao, Z. Li, Q. Zhang. "XGBoost-DNN Mixed Model for Predicting Driver's Estimation on the Relative Motion States during Lane-Changing Decisions: A Real Driving Study on the Highway." Sustainability, 2022, 14(11):1-23.

[10] W. Niu, T. Li, X. Zhang, T. Hu, H. Wu. "Using XGBoost to Discover Infected Hosts Based on HTTP Traffic." Security and Communication Networks, 2019, 2019(1):1-11.

[11] Y. Bao, B. Ke, B. Li, Y. J. Yu, J. Zhang. "Detecting accounting fraud in publicly traded US firms using a machine learning approach." Journal of Accounting Research, 2020, 58(1): 199-235.

[12] R. H. Davidson. "Who did it matter: Executive equity compensation and financial reporting fraud." Journal of accounting & economics, 2022,73(2/3):101453.1-101453.24.

[13] A. Ys, B. Cg, A. Hl, B. JC, C. YG, XQ A. "Financial Feature Embedding with Knowledge Representation Learning for Financial Statement Fraud Detection." Procedia Computer Science, 2021, 187:420-425.

[14] A. Lotfi, M. Salehi, M. L. Dashtbayaz. "The effect of intellectual capital on fraud in financial statements." TQM Journal, 2022,34(4):651-674.

[15] A. Kumar, G. S. Mishra, P. Nand, M.S. Chahar, S.K. Mahto. "Financial Fraud Detection in Plastic Payment Cards using Isolation Forest Algorithm." International Journal of Innovative Technology and Exploring Engineering, 2021, 10(8):132-136.

[16] LaliSransrdjan.lalic.efb@gmail.comJoviieljanazeljana.jovicic@ef.unibl. orgBonjakoviTanjabosnjakovict@gmail.comUniverzitet u Istonom SarajevuEkonomski fakultet Banja LukaPoreska Uprava Republike PJ Bijeljina. The most common examples of financial fraud in Bosnia and Herzegovina: A practical insight. Journal of Forensic Accounting Profession, 2021, 1(2):80-88.

[17] Gepp A , Kumar K , Bhattacharya S . Lifting the numbers game: identifying key input variables and a best erforming model to detect financial statement fraud. Accounting and Finance, 2021, 61:4601-4638.

[18] Zhou H, Sun G, Fu S, Internet Financial Fraud Detection Based on a Distributed Big Data Approach With Node2vec. IEEE Access, 2021, PP(99):1-1.

[19] Swa B, Jl C, Xz A, Envelope MLA. Analysis of financial fraud based on manager knowledge graph. Procedia Computer Science, 2022, 199:773-779.

[20] Geng X., Yang D. Intelligent Prediction Mathematical Model of Industrial Financial Fraud Based on Data Mining. Hindawi Limited, 2021,2021(34):1-8.

[21] Verykios V S , Stavropoulos E C , Zorkadis V , et al. Sensitive data hiding in financial anti-fraud process. International journal of electronic governance, 2022,14(1/2):7-27.

[22] Ys A, Cg B, Hl A, JC B, YG C, XQ A. Financial Feature Embedding with Knowledge Representation Learning for Financial Statement Fraud Detection. Procedia Computer Science, 2021, 187:420-425.