ISSN: 2688-9323

Financial Risk Intelligent Identification from the Perspective of Machine Learning

Wei Li, Chengshu Wu*

School of Finance, Anhui University of Finance and Economics, Anhui 233030, China

Abstract

In the era of digital economy, the use of machine learning algorithms can effectively improve the intelligent level and identification accuracy of financial risk identification. However, a single machine learning method has limited ability to obtain key financial impact factors. This paper uses a family of ensemble learning algorithms for modeling, based on 180 financial indicators, and innovatively constructs a financial risk intelligent identification model that integrates XGBoost, LightGBM and CatBoost, three machine learning methods. The experimental results further show that the financial risk intelligent identification model based on decision tree integration can accurately find the key financial impact factors that cause the deterioration of financial risks. The financial risk intelligent identification model based on ensemble learning constructed in this study can efficiently and accurately identify key financial indicators, providing investors and regulators with a new approach to identifying the financial risks of Chinese listed companies that is easier to understand.

Keywords

Financial Risk Identification; Intelligent Finance; Machine Learning; Ensemble Model.

1. Introduction

The occurrence of financial default behavior for listed companies will cause irrecoverable losses to investors, and with the deepening of corporate financialization, financial risks will also increase (Yuan Linlin et al., 2021). Therefore, accurate and fast identification of financial risks for listed companies has become a real dilemma facing many investors. At the same time, intelligent identification of financial risks is an important means for banks and financial institutions to conduct risk control and one of the main goals for constructing intelligent financial risk management systems, which is also an important manifestation of financial intelligence. As an important part of financial technology innovation and development, intelligent financial risk management based on machine learning techniques can effectively and greatly improve the efficiency of identifying financial problems of listed companies, which is an important element for the high-quality development of China's financial management industry and financial industry. Based on the above analysis, the rapid development of artificial intelligence continues to overturn people's economic and social lives (Li Bin et al., 2019). The 2019 Financial Technology (FinTech) Development Plan (2019-2021) issued by the People's Bank of China clearly stated that artificial intelligence should be gradually applied to explore the application paths and methods of relatively mature artificial intelligence technologies in fields such as risk prevention and control, and build a full-cycle intelligent financial ecosystem. Against this background, how to establish an accurate financial risk identification model to better corporate governance has become a hotspot in academic research and industry applications (He Ying et al., 2020; Liu Meiling et al., 2020).

In the big data environment, with the significant improvement of computer parallel computing power, artificial intelligence technologies represented by machine learning and deep learning

ISSN: 2688-9323

have achieved relatively excellent performance in credit risk management and other fields (Su Zhi et al., 2017). Machine learning takes data as the research object, and the current data-driven prediction model is characterized by high dimensionality, complexity and nonlinearity, making traditional low-dimensional, abstract and linear empirical models unable to well adapt to the above characteristics (Tasoulis et al., 2020). Machine learning, as an extension of statistical knowledge, explains data, processes data, extracts value from it, demonstrates and communicates data results through the cross use of scientific knowledge to solve specific financial and financial problems (Gan et al., 2020). Through automated learning, machine learning methods can accurately extract valuable information that can effectively be used for financial risk identification. In addition, scholars have made many attempts in two key aspects of machine learning: feature selection and model building, among which feature selection includes financial (Persons, 2011; Liu Yunjing et al., 2022; Wu Bin et al., 2022) and non-financial characteristics (Wang et al., 2018).

According to the above analysis, machine learning models are learning algorithms that map inputs to predictions, which can be linear models represented by logistic regression or nonlinear models represented by neural networks, hence the above models can also be called predictors. The aforementioned nonlinear models are black box models, which are systems that do not reveal their internal mechanisms. In machine learning, a black box model describes a model that cannot be understood by viewing parameters (e.g. the parameters of a deep neural network). The opposite of a black box can be referred to as a white box, or an explainable model. In many studies, machine learning models are viewed as black boxes (even though these models themselves are not black boxes), meaning that they are unexplainable models. In this paper, explainable machine learning models are considered to be those whose discrimination process can be transformed into rules with logical relationships, i.e. when using machine learning models for risk identification and prediction, people can fully understand the risk factors and the underlying influencing factors in order to make better decisions. In statistical analysis, hypotheses are proposed and verified with a huge amount of data and rules are established to build corresponding models, for example, listed companies establish a set of machine learning models to associate financial data with marketing activity data to determine what constitutes effective marketing activities. In the aforementioned research, extracting and analyzing the influencing factors behind the model is the key to explainability.

Explainability does not have a mathematical definition, and Miller (2019) defines it as: explainability is the degree to which people can understand the reasons for decisions, which can also be understood as: the degree to which people can consistently predict model results. Therefore, the higher the explainability of machine learning models, the easier it is for people to understand the important basis for making decisions or predictions. When building an intelligent financial risk identification system, understanding "why" will help decision makers better understand the problems, data, and possible reasons for model failure. For this reason, this paper constructs an explainable intelligent financial risk identification system based on machine learning, which has well verified through experiments that explainable machine learning prediction models can accurately identify potential financial risks of listed companies and achieve some success in explainability. Compared to existing parameter methods, non-parameter models provide several important additional advantages.

Compared with linear statistical learning models, non-parameter models can learn nonlinear, discontinuous and complex interactions. Due to the tree-based learning nature of explainable non-parameter models, this class of models has good robustness to outliers in the predictor variables, and they have scale-invariant monotonicity to changes in the predictor variables. This means that there is no need to transform the predictor variables, which is an important advantage in practical applications. Other advantages are that missing values in the predictors

ISSN: 2688-9323

can be automatically adjusted without needing to be estimated, and predictive performance is not affected by multicollinearity problems.

2. Literature Review

Research achievements on intelligent identification of financial risks mainly focus on two methods: 1) identification methods based on statistics; 2) identification methods based on machine learning.

In the early studies of financial risk identification, researchers found the problem of value differences in financial ratios between failed and non-failed enterprises (Mselmi et al., 2017). It is known that the financial ratios of bankrupt enterprises are usually lower, and the diversified business behavior after the listed company's refinancing will obviously increase the financial risks of the enterprise (Liu Chao et al., 2022). Although companies in trouble have different characteristics, their common feature is unstable financial situation. In bankruptcy prediction, the ratios that measure profitability, liquidity and repayment ability are the most important (Antunes et al., 2017; Barboza et al., 2017). Considering that some studies cite different ratios as the most effective predictors of bankruptcy, the order of their importance is not obvious (Veganzones and Severin, 2018). Therefore, identification methods and models based on statistics have continued to develop for a long time.

E. Altman, an American scholar, first introduced the multiple linear discrimination method into financial risk identification. Based on a sample of 33 bankrupt enterprises in the United States from 1946 to 1965 and compared with normal enterprises, he established the classic Z-score model by selecting 5 financial indicators. In 1977, Altman used multiple discriminant analysis (MDA) to establish the Zeta model with 7 indicators, improving the accuracy of the original Z-score model (Huo Yuanyuan et al., 2019; Shen et al., 2020). Subsequently, the logistic regression (LR) model was introduced into the problem of financial risk identification and could accurately predict the probability of financial crisis occurring. Other statistical methods and models include Probit regression (Huo Yuanyuan et al., 2019; Meng Bin et al., 2019).

Compared to traditional logistic regression prediction (Kirkos et al., 2007; Persons, 2011), the widespread development of artificial intelligence technologies and machine learning algorithms has made many financial risk identification methods and models integrated with the aforementioned frontier intelligent technologies. For example, the artificial neural network (ANN) model was combined with the Z-Score model to construct a financial warning model by selecting 5 financial indicators. The results show that the accuracy rate of the ANN prediction model in the training set and test set is higher than the MDA model (Shen et al., 2020). Subsequent studies all confirm that artificial neural network models are superior to traditional LR models and MDA models (Zhou Ying, 2019). Subsequently, more and more scholars have tried to establish financial plight warning models using support vector machine (SVM) models, BP-NN models, logistic regression models and multiple discriminant analysis models, finding that the SVM model has the best predictive performance (Fang Kuannan and Yang Yang, 2018; Wang et al., 2020). Then random forest (RF) method was used to predict the bankruptcy of insurance companies (Sun Lingli et al., 2021), finding that the RF model as a financial warning is superior to SVM and MDA (Shen et al., 2020). In recent years, intelligent identification methods and models based on machine learning models have become an important branch and research hotspot, however, research work on the explainability of machine learning models is still insufficient and relevant literature is relatively scarce, so it needs to strengthen research in this area.

This research has certain theoretical significance and practical value: Firstly, in the field of economic management, this paper for the first time proposed explainable machine learning prediction models to better identify financial risks of listed companies, attempting to

ISSN: 2688-9323

understand "how the model makes predictions" and "how model subsets affect model decisions", these easily ignored but important issues. Secondly, machine learning prediction models can effectively and accurately identify financial risks of listed companies, which helps investors, financial institutions and financial regulators to strengthen credit risk control and improve the level of risk management. Finally, this paper constructs an interpretable predictive model based on machine learning algorithms to identify financial risks of listed companies, enriching the application of machine learning in the field of economic management research and extending the new paradigm of economic management research based on machine learning.

3. Research Design

The intelligent financial risk identification model constructed in this paper is essentially a binary classification problem, i.e. distinguishing listed companies with and without financial risks, whose core task is to classify (predict) sample data into the correct category. Another task of machine learning is regression, which is mainly used to predict numerical data, but this paper mainly solves the classification problem, i.e. accurately identifying listed companies with financial risks. In machine learning, classification and regression problems are both supervised learning. In supervised learning, the classification of training samples is known during model training, and after parameter tuning, the test data can be predicted accurately. Referring to the performance of machine learning models in classification problems, this paper intends to systematically test the explainability of machine learning models in the process of financial risk identification. In order to verify this, this paper selects two representative machine learning models, one is the linear logistic regression (LR) model and the other is the prediction model based on decision trees (DT). This section mainly introduces the prediction model based on decision trees.

3.1. Intelligent Financial Risk Identification Model based on Decision Trees

The logistic regression model fails in cases where the relationship between features and results is nonlinear or interactive. However, decision tree models are a class of machine learning algorithms that are commonly used for classification. Due to their simplicity and explainability, they have been widely used. Based on decision tree prediction models can split (Split, also called partition) data according to certain feature cut-off values multiple times. By partitioning, different subsets of the dataset can be created, and each instance belongs to one subset. The final subset is called a terminal (Terminal) or leaf node (Leaf Nodes), and the intermediate subset is called an internal node (Internal Nodes) or split node (Split Nodes). The average result of the training data for that node is used to predict the result for each leaf node.

Due to the decision tree model or its improved model, they can be well applied to prediction or classification problems, which is due to the fact that based on decision tree prediction models can well explain the key influencing factors that play an important role, making them occupy an important position in explainable machine learning. For financial risk identification and control, the above intelligent artificial intelligence models that are easy to explain are more needed, which can derive key financial impact factors through the calculation of the model, and the leaf nodes of the decision tree correspond to the prediction results, for example, can be used to predict loan granting based on a decision tree prediction model whether a financial institution ultimately extends credit to customers.

1) GBDT

Gradient Boosting Decision Tree (GBDT) model proposed by Friedman in 2001 is a typical ensemble technique based on decision tree models (Ke et al., 2017). As a sequential ensemble method, the goal of the GBDT model is to sequentially train a series of weak base models and

ISSN: 2688-9323

combine them in an additive form to build a strong model. Different from AdaBoost, GBDT uses the gradient information of the loss function residual in each iteration to train the base model.

2) XGBoost

Extreme Gradient Boosting (XGBoost) model proposed by Chen and Guestrin (2016) is an improvement method based on the GBDT model. The XGBoost model is an additive model implemented based on a forward distributed algorithm. By approximating the second-order Taylor expansion of the negative gradient of the loss function and using it as the residual of the previous model, multiple models are iterated in series so that the bias is gradually corrected until the loss satisfies the convergence condition. The commonly used meta-model of XGBoost model is in the form of tree models. Tree models can implement cross-combination of features, and through the serial results of XGBoost, higher-order cross of features can be further realized. Therefore, due to its excellent performance, XGBoost is the winning solution in many machine learning competitions. The XGBoost model improves the prediction accuracy based on the traditional GBDT model, and its goal is to optimize the objective function composed of the loss function and the regularization term, which is different from the loss function used in the GBDT model.

3) LightGBM

LightGBM (Light Gradient Boosting Machine) developed by Ke et al. (2017) is another advanced algorithm based on GBDT. Experiments show that LightGBM is superior to the original GBDT with less computational cost. Ke et al. proved that LightGBM provides predictive results that are even better than XGBoost on some datasets. Although the basic ideas of GBDT and LightGBM are similar, the relatively high performance of LightGBM can be explained by two essential differences, namely the use of "best-first" trees and histogram-based greedy search algorithms. Most GBDT-based methods train base models in a depth manner, while LightGBM trains trees in a leaf manner (i.e., best-first trees). Best-first trees tend to quickly reduce losses but may lead to overfitting. To prevent overfitting, LightGBM can control the depth and split of trees. The histogram-based greedy search algorithm converts continuous variables into discrete bins. This method promises to accelerate the training process and reduce memory usage. LightGBM has done some engineering optimizations in parallel computing and GPU support.

4) CatBoost

CatBoost developed by Prokhorenkova et al. (2018) is a powerful open-source technology based on GBDT that also achieves excellent performance in various machine learning tasks. The model authors claim that the CatBoost model outperforms existing GBDT techniques and sets new records on several benchmarks. Compared to leading GBDT-based algorithms (e.g., XGBoost and LightGBM), CatBoost introduces two major algorithmic improvements, namely ordered boosting and special considerations for categorical features. For ordered boosting, CatBoost proposes a new improved ordered gradient boosting algorithm that is expected to eliminate the effect of biased gradient estimates (i.e., prediction drift problem) while maintaining an acceptable complexity. Categorical features typically occur in credit datasets and GBDT-based commonly used methods, however, these features are converted into gradient statistics in each iteration. This method provides important information for establishing new base models, so it is criticized for consuming a large amount of computational resources. In order to overcome this problem, CatBoost focuses on new computations of target statistics to evaluate the structure of trees.

3.2. Intelligent Financial Risk Identification Model based on Decision Trees

Based on the predictive results of the XGBoost, LightGBM and CatBoost models, the models are weighted combined using formula (1) as shown below.

ISSN: 2688-9323

$$f_{Ensemble,t} = \omega_1 f_{XGBoo\ \ ,t} + \omega_2 f_{LightGBM,t} + \omega_3 f_{CatBoost,t}. \quad (1)$$

Where $f_{Ensemble,t}$, $f_{XGBoost,t}$, $f_{LightGBM,t}$ and $f_{CatBoost,t}$ represent the prediction values of the ensemble model and the corresponding three benchmark models respectively; ω_i $(i \in \{1, 2, 3\})$ represents the weighting of the benchmark model in the ensemble model.

3.3. Data Sources and Sample Selection

The data samples used in this paper come from A-share listed companies that have been publicly issued on the Shanghai and Shenzhen stock exchanges as research samples, from 2000 to 2021. Financial data of listed companies comes from the CSMAR Economic and Financial Database. In this paper, listed company stocks marked as ST (Special Treatment) will be regarded as signals of financial risks. Here, ST stocks refer to listed company stocks that have suffered continuous losses for two years and have been given special treatment, especially if listed company stocks will be given delisting warnings if they suffer losses for three consecutive years and are marked as *ST (Huo Yuanyuan et al., 2019). Listed companies marked as ST or *ST have all encountered financial problems in operation, such as difficulties in funds turnover, inability to repay debts, restricted enterprise investment, leading to default situations.

The descriptive statistics of financial risks occurring from 2000 to 2021 are shown in Table 1 by year. Before 2007, although the number of listed companies increased year by year, the number of enterprises with financial risks existed fluctuations; from 2008 to 2014, the number of enterprises with financial risks showed a slight upward trend; from 2015 to 2021, the number of enterprises with financial risks had an overall stable change and decreased slightly.

Table 1. Distribution of Sample Companies

Year	Companies with Financial Risks	Companies without Financial Risks	Total
2000	111	1079	1190
2001	124	1237	1361
2002	118	1251	1369
2003	122	1327	1449
2004	118	1434	1552
2005	106	1485	1591
2006	111	1628	1739
2007	119	1970	2089
2008	130	2195	2325
2009	140	2303	2443
2010	140	2350	2490
2011	142	2471	2613
2012	145	2778	2792
2013	149	2928	3077
2014	151	3294	3445
2015	150	3379	3529
2016	150	3587	3737
2017	151	4062	4213
2018	150	4486	4636
2019	149	4576	4725
2020	145	4553	4698
2021	142	4540	4682
Total	2963	58782	61745

ISSN: 2688-9323

In terms of feature selection, this paper mainly chooses listed companies' financial ratio indicators as input variables for prediction models. On the one hand, quantitative financial indicators of listed companies are more objective and have been obtained; on the other hand, financial ratio indicators have higher universality and better comparability between enterprises (Liu Yunjing et al., 2022). In order to comprehensively consider the financial risk factors of enterprises, this paper selects 10 first-level indicators with 204 financial ratio indicators from the CSMAR database as feature screening, including repayment ability, structural ratio, operating ability, profitability, cash flow analysis, risk level, development ability, per share indicators, relative value indicators and dividend distribution. Some important financial ratio indicators are shown in Table 2.

Table 2. Definitions of Financial Ratio Indicators

First-level Indicators	Second-level Indicators	Variable Codes	Variable Definitions
Solvency	Current Ratio	F010101A	Current Assets/Current Liabilities
	Cash Ratio	F010401A	Cash and Cash Equivalents at the End of Period/Current Liabilities
	Net Cash Flows from Operating Activities/Current Liabilities	F010801B	Net Cash Flows from Operating Activities/Total Current Liabilities
Structural Ratios	Liquidity Ratio	F030101A	Total Current Assets/Total Assets
	Equity Ratio	F031101A	Total Equity/Total Assets
	Equity-Fixed Asset Ratio	F031401A	Total Equity/Net Fixed Assets
Operating Ability	Accounts Receivable to Revenue Ratio	F040101B	Accounts Receivable/Revenue
	Inventory to Revenue Ratio	F040401B	Inventories/Revenue
	Working Capital Turnover	F040905C	Revenue/Average Working Capital
Profitability	Return on Assets	F050104C	(Profit Before Tax + Finance Costs)/Average Total Assets
	Earnings Before Interest and Taxes	F050601C	Net Profit + Income Tax Expense + Finance Costs
	Profit Before Tax to Earnings Before Interest and Taxes Ratio	F051001B	Gross Profit/Earnings Before Interest and Taxes
Cash Flow Analysis	Net Profit Cash Content	F060101C	Net Cash Flows from Operating Activities/ Net Profit
	Operating Profit Cash Content	F060401C	Net Cash Flows from Operating Activities/Operating Profit
	Cash Recovery Rate	F061701B	(Net Cash Flows from Operating Activities)/(Total Assets) at the End of Period
Risk Level	Financial Leverage	F070101B	(Net Profit + Income Tax Expense + Finance Costs)/(Net Profit + Income Tax Expense)
	Operating Leverage	F070201B	(Net Profit + Income Tax Expense + Finance Costs + Depreciation and Amortization)/(Net Profit + Income Tax Expense +Finance Costs)

ISSN: 2688-9323

	Comprehensive Leverage	F070301B	(Net Profit + Income Tax Expense + Finance Costs + Depreciation and Amortization)/(Net Profit + Income Tax Expense)s
Development Capacity	Capital Preservation and Appreciation Rate	F080102A	(Total Equity at the End of Current Period)/(Total Equity at the End of Same Period of Previous Year)
	Total Assets Growth Rate	F080602A	(Total Assets at the End of Current Period - Total Assets at the End of Same Period of Previous Year)/(Total Assets at the End of Same Period of Previous Year)
	Net Profit Growth Rate	F081002B	(Net Profit of Current Period - Net Profit of Same Period of Previous Year)/(Net Profit of Same Period of Previous Year)
Per Share Indicators	Earnings Per Share	F090102C	Net Profit/Latest Share Capital
	Revenue Per Share	F090602C	Revenue/Latest Share Capital
	Earnings Per Share Before Interest and Tax	F090702C	Net Profit + Income Tax Expense + Finance Costs/Latest Share Capital
Relative Valuation Indicators	Price to Earnings Ratio	F100103C	Closing Price at the End of Current Period/(Net Profit/Paid-in Capital at the End of Current Period)
	Price to Book Ratio	F100401A	Closing Price at the End of Current Period/(Total Equity at the End of Current Period/Paid-in Capital at the End of Current Period)
	Tobin's Q	F100901A	Market Value/Total Assets
Dividend Distribution	Cash Dividend Per Share Before Tax	F110101B	Cash Dividend Per Share Before Tax
	Cash Dividend Per Share After Tax	F110201B	Cash Dividend Per Share After Tax
	Dividend Payout Ratio	F110302B	Cash Dividend Per Share Before Tax/(Net Profit/Latest Share Capital)

3.4. Data Preprocessing

First, listed financial companies such as banks and insurance companies were excluded from the samples; Secondly, in order to avoid the impact of too many missing values on the final model prediction effect, this paper will delete feature variables with missing values more than 25% (the actual input of the model is 180 feature variables); Then, according to whether listed companies have suffered losses for two consecutive years or three consecutive years, each sample is assigned a class label, where samples with financial risks are marked as 1 and samples without financial risks are marked as 0; Finally, all missing values are filled in according to the mean value of the feature variable to which they belong.

3.5. Model Evaluation Criteria

Since the identification of financial risks in this paper can essentially be classified as a binary classification problem (i.e. occurrences of financial risks and no occurrences) in prediction, the evaluation metrics commonly used for classification problems can be used to measure the performance of financial risk identification. A standard classification performance metric is accuracy, defined as: $(TP + TN)/(TP + FN + FP + TN)$, where TP (true positive) is the number of normally operating listed companies that are correctly identified; TN (true negative) is the number of listed companies with financial risks that are correctly identified; FP (false positive) is the number of normally operating listed companies that are misidentified; FN (false

ISSN: 2688-9323

negative) is the number of listed companies with financial risks that are not correctly identified. However, due to the fact that the dataset has a serious imbalance in nature between the number of listed companies with financial risks and those operating normally, accuracy as a so-called standard classification performance metric is not suitable in this paper's scenario (i.e. the number of listed companies with financial risks accounts for about 10% per year. If the accuracy metric is used, the accuracy of the prediction model will remain at around 90%. But such accuracy indicator is meaningless.).

To sum up, the seemingly high performance of financial risk identification metrics in the real application scenario has almost no value. The main goal for investors and financial institutions is to accurately identify listed companies with financial risks as much as possible without confusing normally operating listed companies. That is to say, investment institutions pay more attention to the true positive rate (TPR, also known as sensitivity) and true negative rate (TNR, also known as specificity). Sensitivity is defined as: $SEN = TP/(TP + FN)$, whereas specificity is defined as: $SPE = TN/(TN + FP)$.

According to the above analysis, after the GBDT and XGBoost prediction and classification models are trained, each sample will get 2 corresponding probability values ranging from 0 to 1, representing the probability of being a positive sample or negative sample respectively. Then, the probabilities of positive samples are sorted in descending order and compared with the pre-selected threshold. If the predicted probability is greater than the threshold, it is judged as a positive sample, otherwise it is determined to be a negative sample. In the above predictions, positive samples may be predicted as negative samples, or negative samples may be predicted as positive samples. Thus we can derive two definitions: one is TPR and the other is FPR. The definition of sensitivity (true positive rate) is: $TPR = TP/(TP + FN)$; the definition of false positive rate is: $FPR = FP/(TN + FP)$.The classifier produces a real-valued or probabilistic prediction for each test sample, then compares this prediction with a classification threshold. If it is greater than the threshold, it is classified as positive, otherwise negative. By repeating the process, we can obtain a set of TPR curve and FPR curve. Taking TPR as the y-axis and FPR as the x-axis, we can get the ROC (Receiver Operating Characteristic Curve). Comparing the area under the ROC curve (AUC) is an effective solution to the inadequacy that ROC curve cannot be compared. Therefore, the larger the area under the curve (AUC), the better the performance of the model.

The K-S curve measures the model's discrimination ability. The larger the value, the greater the ability of the model to distinguish between positive and negative customers. When calculating the KS value, TPR and FPR are required. The KS curve consists of two lines. The abscissa is the threshold and the ordinate is the value of TPR and FPR, ranging from 0 to 1. The larger the value, the better the model distinguishes between good customers and bad customers. The definition of K-S value is: $KS = max|TPR - FPR|$.

4. Model Building and Experimental Results Analysis

The experiments in this paper mainly include the following steps: First, the dataset is divided into a training set and a test set. Second, since the samples are imbalanced data, i.e. there is a large gap between the number of samples with financial risks and those without financial risks, the imbalanced training set needs to be balanced. Third, the training set is used to tune the model hyperparameters to determine the optimal model. Fourth, the test set is used for prediction and analysis of the prediction results.

4.1. Dataset Splitting and Imbalanced Data Processing

In the data samples, since the number of samples with financial risks account for a small proportion, if the dataset is divided into a training set, validation set and test set, then the

ISSN: 2688-9323

number of samples with financial risks used for training will be reduced. Therefore, in order to avoid reducing the scale of the training set, this paper divides the dataset into 75% for the training set and 25% for the test set. The training set contains 46308 samples, including 44086 samples without financial risks and 2222 samples with financial risks. Due to the huge gap between the number of samples of the two categories of enterprises, the training set is a typical imbalanced dataset. If this training set is trained directly, the model may tend to classify all samples as the category without financial risks, thus failing to achieve the purpose of identifying enterprises with financial risks. Therefore, this paper will train the model based on balancing the training set.

For imbalanced data processing, this paper uses Synthetic Minority Over-Sampling Technique (SMOTE), which is an oversampling method for minority class oversampling. This balancing processing method calculates the Euclidean distance between each minority sample and other minority samples to determine the k nearest neighbors. Then it randomly selects N neighboring samples points to construct new sample points with the minority sample point according to the following formula:

$$x_{new} = x + (\tilde{x} - x) \cdot random(0,1). \quad (2)$$

The new sample point constructed is x_{new}, N is a randomly selected minority sample point, that is, for each minority class sample point based on k neighboring sample points, N neighboring points are randomly selected and the difference between the original sample point is multiplied by a factor of 0-1, thereby achieving the purpose of artificially synthesizing data. Its advantage lies in the fact that SMOTE does not sample in data space dimensions, but samples in feature space dimensions, so its accuracy is higher than traditional sampling methods.

4.2. Parameter Tuning

Parameter tuning aims to maximize the AUC score in order to examine the performance gap between the ensemble model and the single benchmark models. To comprehensively examine the identification effect of the models, we will consider the TPR, Accuracy, Precision, AUC and KS metrics to evaluate the identification effect of the models. For the benchmark models, we use Grid Search to automatically tune the parameters of the training set to determine the optimal model parameters. This method sets candidate parameters and exhaustively searches through the model identification effect of each parameter combination, and finally outputs the identification result of the optimal parameter model.

The candidate parameters of the three benchmark models are set as follows:1) XGBoost, we set the parameters as follows: $colsample\_bytree$ represents the subsample ratio of columns when building each tree, and subsampling is performed when building each tree. $colsample\_bytree$ is set to {0.6, 0.7, 0.8, 0.9, 1.0}; $subsample$ is used to determine the sampling proportion of subsamples in the training set. $subsample$ is set to{0.7, 0.8, 0.9}. max_depth is used to control the maximum depth and complexity of the subtree. max_depth is set to {3, 4, 5, 6, 7, 8, 9}. Other parameters are set to default values. 2) LightGBM, we set the parameters as follows: $max\_depth \in \{3,4,5,6,7\}$ and $colsample\_bytree \in \{0.5, 0.6, 0.7\}$ are set respectively. $n\_estimators$ is the maximum number of iterations of the model, set $n\_estimators \in \{400, 450, 500, 550, 600\}$. Other parameters are set to default values. 3) CatBoost, we set the parameters as follows: $depth \in \{7, 8, 9\}$, $learning\ rate \in \{0.021, 0.022, 0.023\}$ and $subsample \in \{0.5, 0.6, 0.7, 0.8\}$ are set respectively. Other parameters are set to default values.

4.3. Model Prediction

By trying out the parameter ranges one by one as mentioned above, this paper integrates the three baseline models according to formula (1), respectively setting w1 = 0.3, w2 = 0.1 and w3

ISSN: 2688-9323

= 0.6. According to the respective weights, XGBoost, LightGBM and CatBoost models are integrated and predictions are made on the test set. The prediction performance is shown in Table 3.

Table 3. Ensemble Model Prediction Results

	Precision	Accuracy	Sensitivity	AUC	KS
Ensemble	0.8837	0.9541	0.0513	0.9149	0.6703

Table 3 shows the model's predicted performance on various evaluation metrics. As shown in Table 3, the integrated model has an accuracy of 88.37%, precision of 95.41%, recall also known as TPR of 5.13%, AUC of 0.9149 and KS of 0.6703, with generally high scores. Among them, since AUC as a metric to measure the performance of classifiers is particularly important for this paper, the results show that the integrated model classifier performs well.

4.4. Model Comparison

Table 4 shows the performance of evaluation indicators of the tree ensemble-based financial risk identification model, including three models: XGBoost, LightGBM and CatBoost. In order to evaluate and compare different models, 10-fold cross-validation is used in this paper's models. The financial situation of each listed company will be predicted and compared with the actual results. In machine learning, AUC value is often used to evaluate the training effect of a binary classification model, that is, a model with only two output categories, such as whether or not financial risks occur. Sensitivity refers to the sensitivity to positive samples, for example, if only 5 out of 6 listed companies with financial risks are detected as positive (with risks), while the other 1 is misjudged as negative (without risks), the sensitivity will decrease. Specificity can be understood as the statement of sensitivity on negative samples. In a task with fewer negative samples, the specificity of a classifier shows its ability to make exclusive and special judgments when the negative samples are generally less. The maximum value of the K-S curve is called the K-S value, which ranges from 0 to 1. If it is a random sampling, the Lorenz curve of good customers overlaps with that of bad customers, then the K-S value equals 0. Therefore, for the ideal financial risk identification model, good companies and bad companies are completely separated, then the K-S value equals 1.

Table 4. Evaluation Metrics of Intelligent Financial Risk Identification Model

Model	Precision	Accuracy	Sensitivity	AUC	KS
XGBoost	0.7101	0.9539	0.0661	0.9017	0.6462
LightGBM	0.8276	0.9569	0.1296	0.8960	0.6320
CatBoost	0.9091	0.9532	0.0270	0.9098	0.6547

From Tables 3 and 4, we can observe that the 4 financial risk identification models based on decision trees proposed in this paper, including XGBoost, LightGBM, CatBoost and Ensemble, have achieved optimal predictive performance (the AUC values are close to 1, reaching the optimal identification state). The AUC values are 0.9017, 0.8960, 0.9098 and 0.9149 respectively, with a difference of no more than 2.1%, and the difference in identification performance between them can be ignored. Therefore, the above 4 risk identification models are all robust. Next, the K-S values in the evaluation indicators are all stable above 0.6, indicating that the model has a good degree of identification and can distinguish positive and negative samples, and the higher the value, the higher the model's prediction accuracy. The sensitivity and specificity, the two indicators, are also used to describe the performance of the

ISSN: 2688-9323

classifier. The sensitivity indicator reflects the proportion of listed companies with financial risks that are accurately identified; the specificity indicator shows the probability of correctly judging normally operating listed companies. The two indicators play an important role in risk control and risk identification. The above analysis further proves the excellent predictive performance of machine learning models.

5. Conclusion

This paper introduces machine learning methods and studies samples of A-share listed companies in China from 2000 to 2021. By comparing XGBoost, LightGBM, CatBoost and Ensemble models, the performance difference between the ensemble model and the other three models is examined and analyzed. The study finds that compared with the other three benchmark models, the ensemble model performs better in classifying the performance of whether the finances have risks.In addition, the financial risks of listed companies are extremely difficult to detect, so an important area of accounting research is to develop effective methods to identify financial risks of enterprises in a timely manner, so that investors can avoid unnecessary losses and the financial system can operate stably. This paper finds through experiments that the following 7 indicators play an important role: net profit rate of common stock, retained earnings per share, quick ratio, capital reserve per share, retained earnings per share, turnover tax rate and sales cost rate. Thus, the intelligent financial risk identification system gives the highest score.

The insights of this study are twofold: First, although benchmark machine learning models can also have good classification performance in financial risk samples, by integrating the benchmark models, we find that the integrated model performs better than any single benchmark model; Second, the financial ratios of enterprises to some extent reflect the probability information of enterprises encountering financial risks, and enterprises cannot adjust financial indicators to reduce the possibility of financial risks overnight. Therefore, we need to examine multiple comprehensive indicators such as enterprise development capability, profitability, operational capability, solvency and ratio structure to fully play the supervisory role of enterprise management.

Acknowledgments

Anhui University of Finance and Economics Graduate Student Research and Innovation Fund: Research on Financial Fraud Identification Method based on Ensemble Learning (ACYC2022512).

References

[1] Antunes F, Ribeiro B, Pereira F. Probabilistic Modeling and Visualization for Bankruptcy Prediction. Appl Soft Comput, Vol.60 (2017), p.831-843.

[2] Barboza F, Kimura H, Altman E. Machine Learning Models and Bankruptcy Prediction. Expert Syst Appl, Vol.83 (2017), p.405-417.

[3] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2016), p.785-794.

[4] Gan L R, Wang H M, Yang Z J. Machine Learning Solutions to Challenges in Finance: An Application to the Pricing of Financial Products. Technol Forecast Soc Chang, Vol.153 (2020), p.11.

[5] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Proceedings of the 31st International Conference on Neural Information Processing Systems, (2017), p.3149–3157.

ISSN: 2688-9323

[6] Kirkos E, Spathis C, Manolopoulos Y. Data Mining Techniques for the Detection of Fraudulent Financial Statements. Expert Syst Appl, Vol.32 (2007) No.4, p.995-1003.

[7] Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences. Artif Intell, Vol.267 (2019), p.1-38.

[8] Mselmi N, Lahiani A, Hamza T. Financial Distress Prediction: The Case of French Small and Medium-sized Firms. Int Rev Financ Anal, Vol.50 (2017), p.67-80.

[9] Persons O S. Using Financial Statement Data to Identify Factors Associated With Fraudulent Financial Reporting. Journal of Applied Business Research (JABR), Vol.11 (2011) No.3, p.38.

[10] Prokhorenkova L, Gusev G, Vorobev A, Dorogush A V, Gulin A. CatBoost: Unbiased Boosting with Categorical Features. Proceedings of the 32nd International Conference on Neural Information Processing Systems, (2018), p.6639–6649.

[11] Shen F, Liu Y Y, Wang R, Zhou W. A Dynamic Financial Distress Forecast Model with Multiple Forecast Results under Unbalanced Data Environment. Knowledge-Based Syst, Vol.192 (2020), p.16.

[12] Tasoulis S, Pavlidis N G, Roos T. Nonlinear Dimensionality Reduction for Clustering. Pattern Recognit, Vol.107 (2020), p.11.

[13] Veganzones D, Severin E. An Investigation of Bankruptcy Prediction in Imbalanced Datasets. Decis Support Syst, Vol.112 (2018), p.111-124.

[14] Wang G, Chen G, Chu Y. A New Random Subspace Method Incorporating Sentiment and Textual Information for Financial Distress Prediction. Electron Commer Res Appl, Vol.29 (2018), p.30-49.

[15] Wang G, Ma J L, Chen G, Yang Y. Financial Distress Prediction: Regularized Sparse-based Random Subspace with ER Aggregation Rule Incorporating Textual Disclosures. Appl Soft Comput, Vol.90 (2020), p.18.

[16] Fang Kuannan, Yang Yang. Research on SGL-SVM method and its application in financial distress prediction [J]. Statistical Research, 2018, 35 (08): 104-15.