Corporate Credit Risk Rating Model Based on Financial Big Data

Mingzhi Tang[1, #, *], Wenhao Zeng[2, #], Runzhou Zhao[3, #]

[1]School of Electronic Engineering BUPT, Beijing University of Posts and Telecommunications, Beijing, China, 100876

[2]School of Economics and Management, Jiangxi University of Chinese Medicine, Jiangxi, China, 330004

[3]School of Finance, Zhongnan University of Economics and Law, Hubei, China, 430070

*Corresponding author: tmztoo@163.com

[#]These authors contributed equally

Abstract. In recent years, leveraging financial big data and machine learning to identify corporate risks has emerged as a crucial approach for financial risk management. This paper proposes a method based on financial big data and the LightGBM model to effectively assess corporate credit risk ratings. Feature engineering is performed on corporate financial datasets, using correlation coefficients, chi-square tests, and machine learning techniques to select essential financial indicators. Subsequently, bayesian optimization is employed for hyperparameter tuning, using the classification accuracy of high risk and highest risk categories as the objective function. This process yields a multi-classification model capable of effectively identifying corporate credit risk ratings through financial data. The results demonstrate that the model exhibits strong identification capabilities for high credit risk corporates. The model achieves the best classification performance for high-risk categories, with an accuracy of 74%. The comprehensive classification accuracy and recall rate for both high-risk and highest-risk categories reach 70%. The overall classification accuracy across all categories is approximately 64%. In summary, through judicious model selection, data preprocessing, feature selection, Bayesian parameter tuning, and the establishment of appropriate objective functions, the LightGBM model demonstrates robust performance in addressing corporate credit risk rating problems.

Keywords: Corporate Credit Rating, LightGBM, Bayesian Tuning, Feature Filtering.

1. Introduction

Corporate credit risk assessment has long been a crucial concern for financial institutions, investors, and policymakers. Historically, many financial institutions relied on traditional statistical methods and expert judgment to evaluate credit risk. However, such approaches are often influenced by subjective factors and inadequate data, potentially resulting in inaccurate predictions of corporate credit risk. In recent years, advancements in financial big data and machine learning technologies have led the industry to increasingly adopt these cutting-edge tools to enhance the performance of credit risk assessment models [1]. Researchers have tried various machine learning algorithms, such as decision trees, random forests, support vector machines, and neural networks, to improve the accuracy and reliability of credit risk predictions.

Among the various models, LightGBM stands out as an efficient gradient boosting decision tree algorithm, widely employed in credit risk assessment within the financial sector. Its primary advantage lies in offering accelerated training speeds and enhanced accuracy when handling large-scale datasets. Implementing LightGBM can streamline the credit risk assessment process, mitigate default risks, and yield benefits like increased profits and reinforced financial stability for financial institutions. Concurrently, investors can capitalize on it to make more informed investment decisions. Furthermore, these models have the potential to generate positive social and economic impacts, such as encouraging lending to SMEs and other underrepresented groups, thereby fostering economic

growth and job creation. They may even help avert financial crises and bolster the stability of the financial system.

2. Data Description

2.1 Feature explanation

The dataset utilized in this paper comprises one target variable and 26 features. These features are continuous numerical variables, representing typical indicators associated with corporate financial risks. The names and specific meanings of these features are presented in Table.1:

Table.1. Feature names and their meanings

Feature	Meaning
currentRatio	Ratio between current assets and current liabilities of an corporate
quickRatio	The ratio of the corporate's current assets that are rapidly realizable to its current liabilities
cashRatio	Ratio of an corporate's current assets such as cash, cash equivalents and marketable securities that are immediately convertible to cash to current liabilities
daysOfSales Outstanding	Used to measure the average number of days it takes for a company to collect its accounts receivable
netProfitMargin	Represents the percentage of net profit per sale, i.e., the percentage of net profit the firm earns from each sale
preTaxProfitMargin	The percentage of pre-tax profit a business earns from each sale
grossProfitMargin	The ratio between the gross profit obtained by the corporate after subtracting direct costs from sales revenue and sales revenue
operatingProfitMargin	The ratio between the operating profit earned by a corporate after deducting the cost of goods sold and operating expenses and the sales revenue
returnOnAssets	The ratio between the profit of the corporate and the assets it owns
returnOnCapitalEmployed	The ratio between the profit of the corporate and its total capital employed
returnOnEquity	Ratio between the net profit of a corporate and its shareholders' equity
assetTurnover	Ratio between sales revenue of a corporate and the assets it uses
fixedAssetTurnover	Ratio between corporate sales revenue and its fixed assets
debtEquityRatio	Ratio between corporate liabilities and shareholders' equity
debtRatio	Total corporate liabilities as a percentage of total assets
effectiveTaxRate	Corporate income tax as a percentage of corporate profit before tax
freeCashFlow	Net cash inflow generated by a corporate from operating activities, net of capital expenditure
operatingCashFlowRatio	Operating cash flow divided by current liabilities
freeCashFlowPerShare	Free cash flow generated by the business per share of stock
cashPerShare	Amount of cash owned by the business per share of stock
companyEquityMultiplier	Ratio between total assets of a corporate and total shareholders' equity
ebitPerRevenue	Ratio between a corporate's operating profit and operating income
corporateValueMultiple	The ratio of a firm's price per share divided by its earnings per share, i.e., the P/E ratio
operatingCashFlowPer Share	Cash flows from operating activities per ordinary share of the business
operatingCashFlowSales Ratio	Ratio of net cash flow from operating activities to operating income of a corporate
payablesTurnover	The frequency with which a business spends a given amount of time paying its accounts payable, i.e., the accounts turnover rate

2.2 Grade distribution map of target variable

Corporate credit risk rating is defined as a multi-classification task, and the target variable of the task is the result of risk rating. In this paper, corporate credit profiles are defined into four categories:

1) Category 1: the risk of default is low, and its category of the target variable is named "Low Risk".

2) Category 2: the default risk is medium, and its category of the target variable is named "Medium Risk".

3) Category 3: the default risk is higher, and its category of the target variable is named "High Risk".

4) Category 4: the default risk is extremely high, and its category of the target variable is named "Highest Risk".

Figure 1 shows the distribution of the target variable in the sample space.

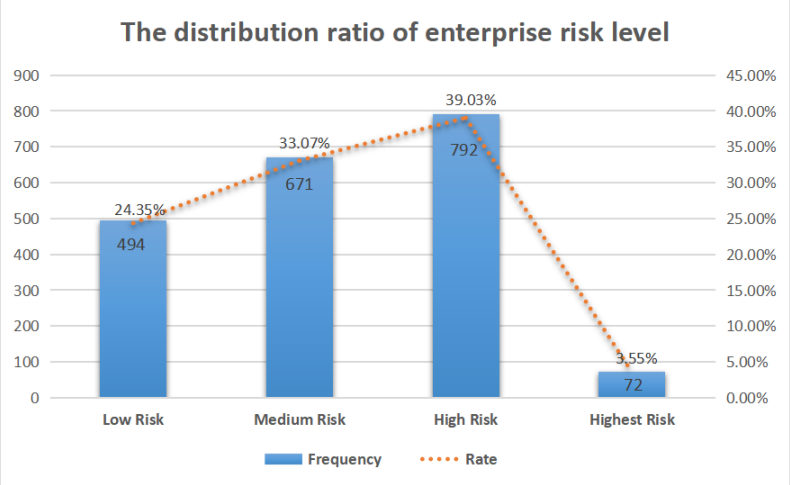

Figure 1. The Distribution Ratio of Corporate Credit Risk Rank

The total number of samples is 2029. Among them, the number of samples in each category of Low Risk, Medium Risk, High Risk, and Highest Risk is 494, 671, 792, and 72, respectively, accounting for 24.35%, 33.07%, 39.03%, and 3.55% respectively. From the distribution characteristics of the target variable, it can be found that the number distribution of Low Risk, Medium Risk, and High Risk is relatively balanced, but the number of samples of the Highest Risk category is small. The dataset has high-class imbalance characteristics.

3. Methods

3.1 Data normalization

Data normalization involves scaling data to fit within a specific, small range, and is commonly used when comparing or evaluating indicators. The objective is to eliminate the constraints imposed by units of measurement, transforming the data into dimensionless values that enable comparisons and weighting across different units or magnitudes. A typical example of data normalization is mapping data to the [0, 1] interval, allowing for feature comparisons across varying dimensions in terms of value [2]. This process can significantly enhance the accuracy of classifiers. To achieve normalization, the following formula is applied to each numeric feature:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Among them, X_{norm} is the data after normalization, and X_{min} and X_{max} are the minimum and maximum values of the data. In this paper, Min-Max Scaling method in the python library sklearn is used for feature normalization, each column of features is normalized separately.

3.2 Feature selection

The feature data employed to train a machine learning model greatly influences the performance that can be attained. However, an excessive number of features can not only increase the model's complexity but also introduce more noise data, heightening the risk of overfitting. As a result, feature selection is often necessary prior to model training.

In this paper, three methods are employed to evaluate feature importance: correlation coefficient, chi-square test, and the feature importance ranking technique integrated within the LightGBM model algorithm.

3.2.1 Correlation coefficient method.

The correlation coefficient typically refers to the Pearson coefficient, which is the ratio of the product of the covariance of two variables and the standard deviation of the two variables. It can measure the linear correlation between variables. The value range of the result lies between [-1, 1], where -1 signifies complete negative correlation, +1 indicates entirely positive correlation, and 0 denotes no linear correlation. In this paper, feature selection is performed by calculating the correlation between the feature and the target variable [3]. The correlation coefficient is presented in the form of a heatmap in Figure 2. The heatmap results reveal that the six features with a strong correlation to the target variable Rating are debtRatio, cashRatio, corporateValueMultiple, assetsTurnover, fixedAssetTurnover, and returnOnEquity.

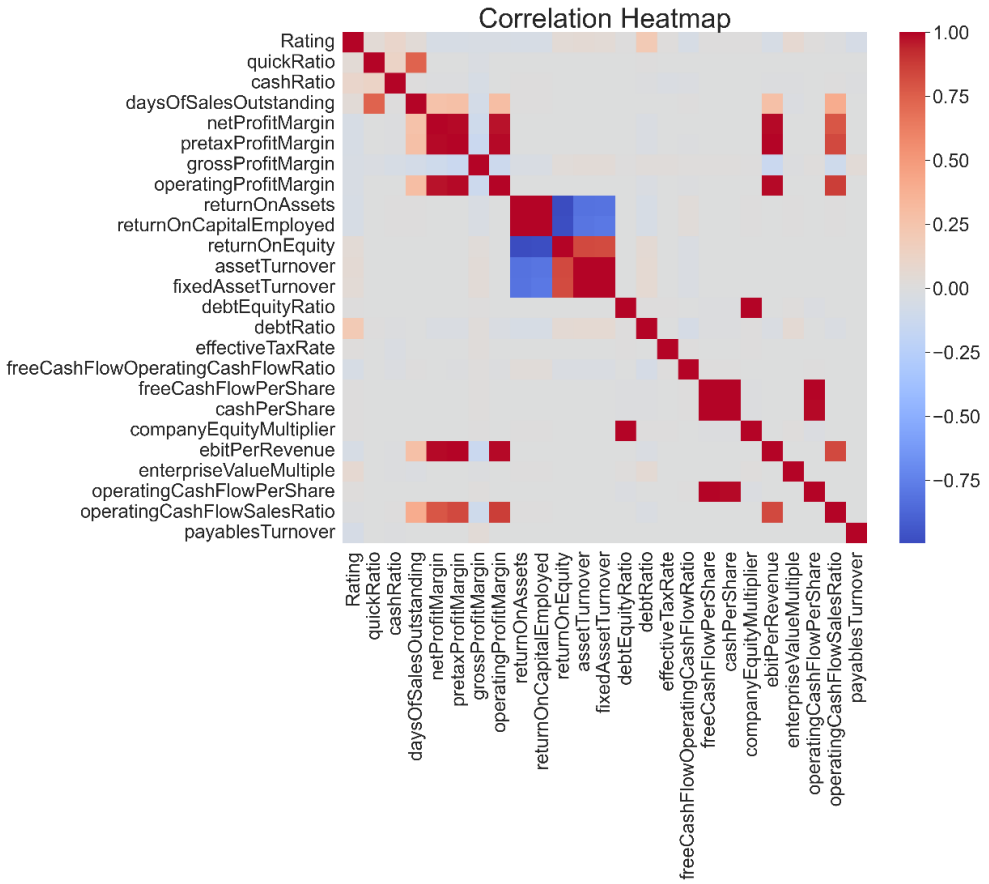

Figure 2. Heat map of correlation coefficient

3.2.2 The chi-square test.

The classical chi-square test examines the correlation between qualitative independent variables and qualitative dependent variables. Assuming that the independent variable has N values and the dependent variable has M values, this test considers the difference between the observed values and the expected frequency of the sample frequency equal to i and the dependent variable equal to j, and constructs a statistic that represents the correlation between the independent variable and the dependent variable [4]. In Python, the Sklearn module is commonly used to implement chi-square testing, as it integrates multiple machine-learning methods. Sklearn provides APIs with the Select_K_Best feature selection methods. This paper presents the top 10 important features screened out by the Select_K_Best method in Figure 3.

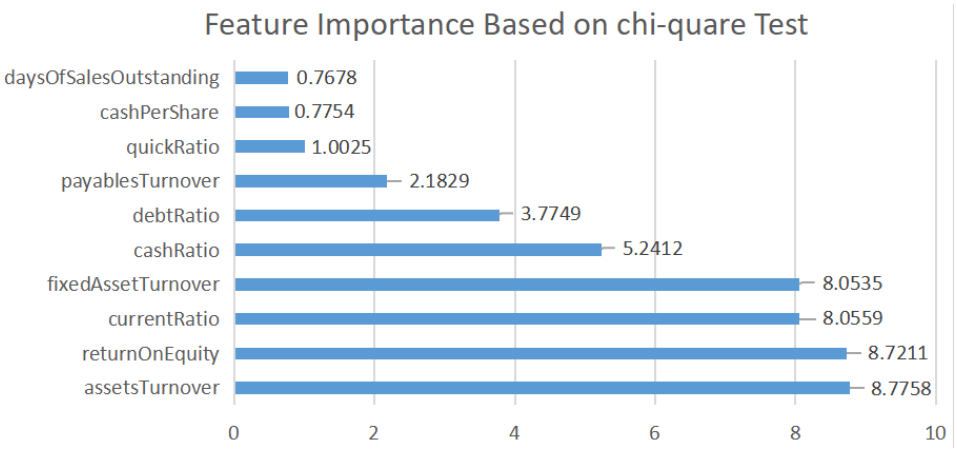

Figure 3. Important features based on chi-square test.

3.2.3 Machine learning methods.

The LightGBM algorithm comes with three methods of feature screening: weight, gain, and cover. The logic behind this approach involves ordering information gain by using embedded feature importance selection. The weight method is more commonly utilized to calculate feature importance, representing the number of times a feature is used as a splitting attribute in all trees. The more frequently a feature is used, the greater the information gain brought by the feature splitting, and the stronger the distinguishing ability of this feature [5]. Utilizing the importance calculation interface provided by the LightGBM model, this paper derives the feature importance ranking as displayed in Figure 4:

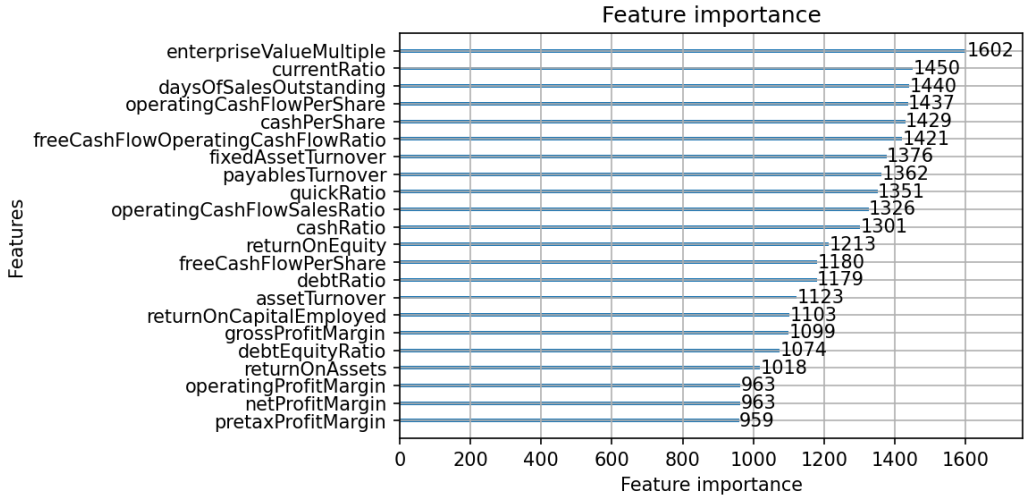

Figure 4. Feature importance ranking based on LightGBM model.

Based on the results of feature filtering, the difference in the importance of each feature in this dataset is not very pronounced. Among them, the top fifteen features identified by this paper include enterpriseValueMultiple, currentRatio, returnOnCapitalEmployed, cashPerShare, freeCashFlowOperatingCashFlowRatio, quickRatio, debtRatio, payablesTurnover, daysOfSalesOutstanding, cashRatio, corporateValueMultiple, assetsTurnover, fixedAssetTurnover, returnOnAssets, and netProfitMargin.

3.3 LightGBM

3.3.1 Introduction to the model and its concepts

LightGBM is a gradient boosting framework. The algorithm is derived from the GBDT (Gradient Boosting Decision Tree) framework. LightGBM utilizes the concept of ensemble learning to enhance the model's overall performance by combining the prediction results of multiple weak classifiers (typically decision trees). Each weak classifier is trained on the residuals of the previous classifier to learn new data features in each iteration round and then obtain the most effective model [6]. LightGBM supports efficient parallel training with faster training speeds, lower memory consumption, better accuracy, distributed computing support, and rapid processing of massive data.

LightGBM adopts a histogram optimization strategy, sorting the features in each dimension of the sample before training, and then dividing the feature histogram. In subsequent training, the algorithm only needs to use the histogram as a "feature" for constructing the decision tree, which significantly reduces the number of traversals of the sample set [7]. The algorithm is shown in Figure 5

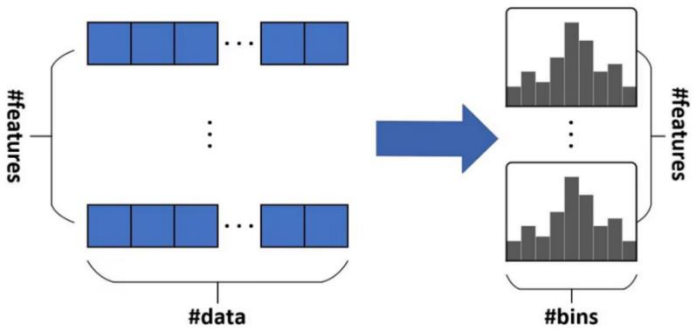

Figure 5. Histogram-based decision tree algorithm

3.3.2 Implementation Algorithm of LightGBM Model

Algorithm of LightGBM

Input: Training dataset:

$$X = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n), x_r \in \boldsymbol{R^n}, \boldsymbol{y_r} \in \{-1,1\} \tag{2}$$

CART weak classifier:$f(x)$; Loss function:$L(y, F(x))$; number of iterations:J

Initialization:$F_0(x) = 0$

1．For $j = 1$ to J {

2．Calculate the negative gradient of the loss function，which is the so-called pseudo-residuals:

$$\widetilde{y_r} = -\left[\frac{\partial L(y_r, \mathrm{F}(x_r))}{\partial F(x_r)}\right]_{F(x)=F_{j-1}(x)}, r = 1,2, \dots, n \tag{3}$$

3．Use CART model to fit the$\{x_r, \widetilde{y_r}\}_{r=1}^{n}$,by using Newton's method to minimize the loss function to obtain the jth regression tree:

$$f_j(x) = \arg\min_{f_j(x)} \sum_{r=1}^{n} L(y_r, F_{j-1}(x_r) + f_j(x_r)) \tag{4}$$

4．Update the model:

$$F_j(x) = F_{j-1}(x) + f_j(x) \tag{5}$$

}

Output: Ensemble model：

$$F_J(x) = \sum_{j=1}^{J} \sum_{r=1}^{n} f_j(x_r) \tag{6}$$

4. Results and Discussion

4.1 Parameter tuning and model training

To facilitate model training, this paper replaces the labels of the multi-category target variables. The high-risk category is replaced by "0", the medium risk category is replaced by "1", the high risk category is replaced by "2", and the highest risk category is replaced by "3". The dataset's completeness is very high, with no missing values and a total of 2028 observations. Through feature screening, 23 features are ultimately retained. To verify the model's accuracy, 20% of the data (405 observations) is used as the validation set, and 80% of the data (1623 observations) is used as the training set. In the LightGBM model, the core parameters that significantly impact the model's performance, their adjustment range, and meanings are explained in Table.2:

Table.2. Core parameters of LightGBM

Parameter	Range	Meaning
max_depth	[3, 10]	Maximum depth of the tree
num_leaves	[10, 300]	Specify the number of leaves, the default value is 31, the value of this parameter should be less than 2max_depth
bagging_fraction	[0.1, 1]	Proportion of data used in each iteration
feature_fraction	[0.1, 1]	The ratio of random sampling of features when building a weak learner, the default value is 1
objective	multiclass	For assigning learning tasks and corresponding learning objectives
num_class	4	Used to set the number of categories for multi-category problems
boosting_type	gbdt	Used to specify the type of the weak learner, the default value is 'gbdt'
min_data_in_leaf	[5, 10]	Minimum number of leaf node samples, default value 20, used to prevent overfitting.
learning_rate	[0.01, 0.1]	LightGBM does not fully trust the residuals learned by each weak learner, so it is necessary to multiply the residuals fitted by each weak learner by learning_rate with values in the range (0, 1]

Among these parameters, *max_depth* and *num_leaves* are the most important for improving accuracy. *max_depth* sets the tree depth. The deeper the tree, the more likely the model is to overfit. *num_leaves* is a conversion result of *max_depth*. Since LightGBM uses leaf-wise algorithms, *num_leaves* can be used instead of *max_depth* when tuning the tree's complexity. The approximate conversion relationship is: $num\_leaves = 2^{max\_depth}$, but its value is often set to less than 2^{max_depth} to avoid overfitting.

To achieve better classification performance, the model's parameters need to be adjusted. Common parameter tuning methods include grid search, random search, and Bayesian parameter tuning. In this paper's parameter tuning process, the Bayesian parameter tuning method is used. The main idea of Bayesian parameter tuning is to constantly update the posterior distribution of the optimized objective

function (a generalized function that only needs to specify inputs and outputs without knowing the internal structure and mathematical properties) by adding sample points [8]. This approach considers the information of the previous parameter to better adjust the current parameter.

In classifying corporate credit risk, this paper focuses more on companies with higher risk. Consequently, the classification accuracy of the LightGBM model for the two categories of High Risk and Highest Risk is selected as the objective function, and 400 rounds of iterative training are conducted to obtain the parameter set with the best classification effect.

During the Bayesian parameter tuning process, the comprehensive classification accuracy of the two categories of High risk and Highest risk, as well as the values of the two parameters *max_depth* and *num_leaves*, change with the number of iterations of the model as shown in the Figure 6.

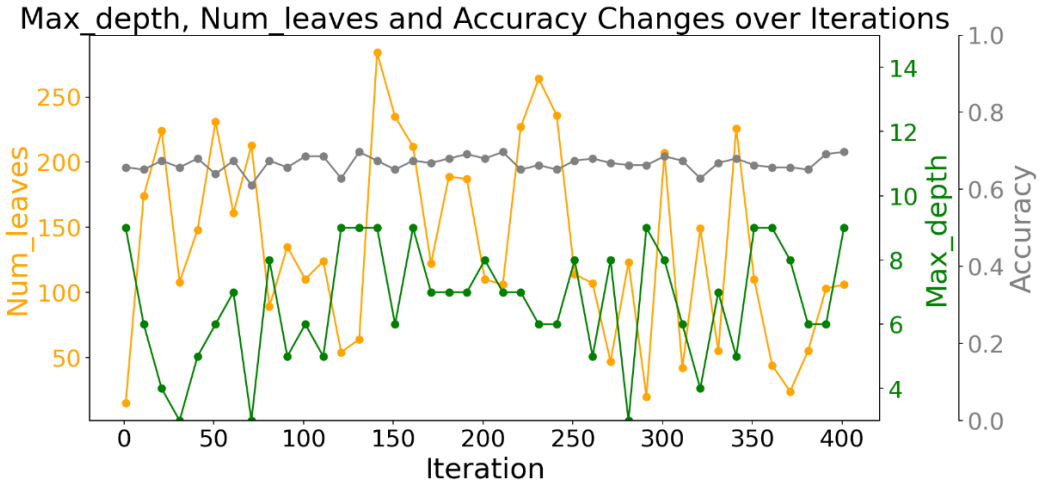

Figure 6. Trend of parameters and accuracy with the number of iteration rounds

During the Bayesian tuning process, the ranges for *max_depth* and *num_leaves* are set between 3-10 and 10-300. Following the Bayesian parameter tuning, the optimal values for these parameters are determined to be 6 and 100. Throughout the entire tuning process, the model's combined classification accuracy for the High Risk and Highest Risk categories peaks at 72%. As observed in the figure above, the value of *max_depth* has a more pronounced impact on the objective function, particularly when the value of *max_depth* is small, which can lead to a considerable decline in classification accuracy. As this paper addresses a multi-classification problem, the complexity and classification difficulty of the model are greater than those of typical binary classification tasks. Consequently, it is essential to set *max_depth* to a larger value to achieve better performance.

4.2 Overall performance evaluation

Upon incorporating the optimal parameter set derived from Bayesian tuning into the model, this paper evaluates the model's performance from various perspectives, including overall classification accuracy (encompassing all four categories), overall precision, overall F1 score, combined classification accuracy for High Risk and Highest Risk categories, and the comprehensive recall for High Risk and Highest Risk categories. The values for each metric are depicted in Figure7:

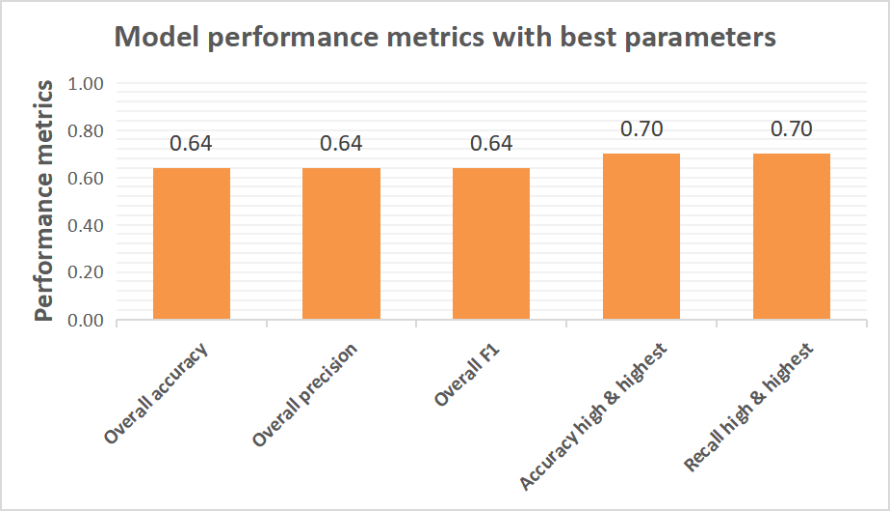

Figure 7. Value of performance metrics of the model

From the performance evaluation results, it is evident that our model excels in classifying the High Risk and Highest Risk categories, with the comprehensive accuracy and recall rate reaching 70%. However, the model's classification performance for Low Risk and Medium Risk categories is suboptimal, leading to less-than-ideal comprehensive accuracy, precision, and overall F1 score across the four classification labels. This can be attributed to two factors.

Firstly, during feature engineering, greater emphasis was placed on the sample distribution of high-risk categories, ensuring that the features used in the training process accurately reflect the distribution differences among high-risk category corporates. Secondly, during the Bayesian parameter tuning process, the objective function (the criterion for determining model classification performance) prioritized the comprehensive classification accuracy of the High Risk and Highest Risk categories. Consequently, the optimal set of parameters focused on enhancing the model's ability to identify high-risk corporates.

Despite these limitations, the LightGBM model demonstrates satisfactory performance in addressing the corporate credit risk rating problem based on financial big data. Through feature screening, Bayesian parameter tuning, and the establishment of suitable objective functions, the model exhibits enhanced recognition ability for samples with higher risk rankings.

5. Conclusions

In conclusion, the aforementioned solutions effectively address the present challenges associated with financial big data and machine learning modeling, thereby enhancing the accuracy and efficiency of credit risk assessment. This, in turn, offers more reliable and precise credit rating services for the financial industry. The insights derived from this paper hold significant value for understanding the application of financial big data and machine learning algorithms in credit risk rating models, paving the way for further advancements in this field.

References

[1] Belhadi A, Kamble S S, Mani V, et al. An ensemble machine learning approach for forecasting credit risk of agricultural SMEs' investments in agriculture 4.0 through supply chain finance [J]. Annals of Operations Research, 2021: 1-29.

[2] Gao G, Wang H, Gao P. Establishing a credit risk evaluation system for smes using the soft voting fusion model [J]. Risks, 2021, 9(11): 202.

[3] Wang D, Li L, Zhao D. Corporate finance risk prediction based on LightGBM [J]. Information Sciences, 2022, 602: 259-268.

[4] Sun J, Li J, Fujita H. Multi-class imbalanced enterprise credit evaluation based on asymmetric bagging combined with light gradient boosting machine [J]. Applied Soft Computing, 2022, 130: 109637.

[5] Gao B, Balyan V. Construction of a financial default risk prediction model based on the LightGBM algorithm [J]. Journal of Intelligent Systems, 2022, 31(1): 767-779.

[6] Ruan S, Zhang J, Li W. CUS-LightGBM-based financial distress prediction for small-and medium-sized enterprises with imbalanced data [J]. 2021.

[7] Ponsam J G, Gracia S V J B, Geetha G, et al. Credit Risk Analysis using LightGBM and a comparative study of popular algorithms [C]//2021 4th International Conference on Computing and Communications Technologies (ICCCT). IEEE, 2021: 634-641.

[8] Qi M. LightGBM: A Highly Efficient Gradient Boosting Decision Tree [C]// Neural Information Processing Systems. Curran Associates Inc. 2017.