Proceedings of the 3rd International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/2025.BL31296

M&A Completion Re-exploration: New Evidence from the LightGBM Model with SMOTE Oversampling

Xuesheng Meng[1*], Rujin Huo[2], Zhichao Hao[3]

[1]*Department of International Business, University of International Business and Economics, Beijing, China*
[2]*Faculty of Science and Technology, Beijing Normal University & Hong Kong Baptist University United International College, Zhuhai, China*
[3]*Ningbo Binhai International Cooperative School, Ningbo, China*
**Corresponding Author. Email: 202310330934@uibe.edu.cn*

Abstract. Predictions about M&A success always suffer from the challenge of unbalanced data, which can easily lead to biased predictions. In addition, previous empirical studies have certain limitations when facing high-dimensional relationships, making it difficult to provide a more global perspective. This study constructs and compares several machine learning models to propose an optimal model. This optimal model is lightGBM, which is constructed from the data after SMOTE oversampling. The results of LightGBM come from the CSMAR database of 3672 M&A transactions, revealing the relative importance and directions of 50 predictors on M&A outcomes, which may have contradictory results or do not appear in previous literature. The findings of this study provide new insights into predicting the success of M&A deals of Chinese listed companies and suggest new directions for future research.

Keywords: M&A, machine learning, lightGBM, SMOTE.

1. Introduction

Although successful M&A transactions can bring a series of positive benefits to the development of the firm, the industry, and even the country [1-5], due to its high degree of complexity, a portion of firms ultimately fail to complete the thrilling leap from planning an M&A to formally signing an agreement, which finally results in varying degrees of losses for companies' stakeholders.

First of all, the failure of M&A transactions usually leads to more sunk costs in terms of capital and time. Both parties to the merger and acquisition need to hire a third-party agency to conduct due diligence in the pre-acquisition period, as well as to hire professional consultants to carry out consulting services, which incurred costs that cannot be recovered [6]. For example, on April 6, 2016, Pfizer officially announced that the merger agreement between itself and Allergan had been terminated, with Pfizer paying Allergan $150 million to reimburse Allergan for the costs associated with the previous transaction [7]. Meanwhile, the management team focuses a lot of attention on the deal, which may trivialize other strategic decisions and ignore new market opportunities [8]. In addition to this, the failure of an M&A transaction may also result in negative market reactions and reputational damage for high-profile M&A events, considering that different companies have different levels of social attention [9,10]. In the case of Allergan, the day before Pfizer announced the termination of the M&A process between these two companies, the share price of Allergan

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

plummeted by 15% [11,12]. At the same time, six third-party consulting firms involved in Pfizer's acquisition of Allergan lost a total of approximately $236 million [11]. Therefore, it is important for both the companies leading the M&A transaction and other stakeholders involved in it to identify the potential factors that could influence M&A completion.

There has been a great deal of discussions in the traditional empirical field on the potential factors of M&A success. However, most of these discussions have been based on specific theories, which has limited the potential factors to some specific key variables. These studies then select appropriate datasets based on these key variables, which may ultimately lead to bias in sample selection and omission of other potential variables [13]. Then it may not be conducive to the generalization of the findings to the general population [14], for example, Dong et al's study excludes samples that are missing relevant information about the deal such as the deal size, the engagement of financial advisors, and the attitude of the target company [15]. At the same time, these studies belong to different subfields, some of them discuss the impact of corporate governance on M&A, while others focus on the role played by language and religious distance between buyers and sellers [16]. These potential factors are not considered simultaneously in an article that uses traditional empirical research methods, which leads to a fragmented relationship between the studies and makes us difficult to determine which part is more decisive in influencing the success of M&A.

Unlike traditional empirical methods, machine learning does not specify the variables under a specific theory, but firstly input all potential variables. And it does not pre specify the form of the function, but determine the shape of the function through the learning of the data in the training set and the validation of the data in the test set [13]. Such a process makes machine learning more predictive and helps to discover more possible causal relationships [17]. On this basis, we were able to generate a ranking of the importance of all variables according to the predicted outcome, linking variables that were originally dispersed across theoretical frameworks, and ultimately providing us with a relatively global perspective for exploring potential factors for M&A success. Therefore, based on logistic regression and machine learning algorithms (neural networks, CART decision trees, and light gradient boosting machine method), this study will address three questions about M&A success prediction modeling (1) Is there a new model building methodology that can enable M&A prediction models to improve their predictive accuracy further on the basis of previous research? (2) When a large number of potential M&A predictors exist, which factors contribute more to predicting M&A success relative to others? (3) Which dimensions of predictors have been missed in previous studies, or which factors that have been explored have been overestimated?

The main contributions of this study are as follows: In the field of traditional M&A research, this study exploratively test some feature variables that have not been discussed in the previous literature, which provides a new perspective for discovering other potential mechanisms. Besides, this study uses the SHAP (SHapley Additive exPlanations) technique to analyze the significance of predictors, which helps to provide a relatively global perspective. In the field of M&A prediction models, the model (lightGBM with SMOTE oversampling) constructed in this study exceeds previous models in prediction precision. It better mitigates the data imbalance and improves the predictive performance of the model more substantially, especially for minor category. Additionally, this study pays extra attention to industry level predictors, which was previously overlooked in predictive M&A studies.

The rest of the paper is organized as follows: Chapter 2 provides a systematic review of previous literature and discusses research progress on M&A failure factors, machine learning for processing classification tasks, and M&A prediction based on machine learning. Chapter 3 introduces sample selection and variable definitions, algorithm selection, algorithm tuning, and model performance evaluation metrics. Chapter 4 conducts a model performance analysis as well as an analysis of the importance of predictors. Chapter 5 summarizes the main findings, theoretical implications, practical implications, limitations and future research directions.

2. Literature review

2.1. Antecedents of M&A failure

This study focuses on the pre-merger stage, rather than the type of failure where the value generated by M&A does not compensate for the original costs incurred resulting in a decline in the acquirer's financial performance after the deal is completed. Specifically, the success of M&A discussed in this study means that the two firms sign a formal transaction contract after agreeing on all the terms and assuming the contractually agreed joint responsibilities. Conversely, if the acquirer announces the termination of the acquisition at the pre-merger stage, the M&A transaction is considered a failure [18].

Based on this definition, previous empirical studies have shown that variables from four dimensions can affect the success rate of M&A transactions: deal level, firm level, industry level, and country level. Specifically, the deal level factors include the payment method [4], whether it involves intellectual property rights [19], and consultant involvement [20,21], etc. The firm level factors can be distinguished into financial and non-financial characteristics, with financial characteristics including financial leverage [22], liquidity [23], and market value [24], etc., and non-financial characteristics including the governance structure [25], previous M&A experience [26], and the degree of industry match [27], etc. The industry level factors include the industry competitive structure [28], the industry risk [29], etc. The country level factors are mainly used for cross-border M&A transactions, including the institutional distance between the two countries [30], cultural distance [15], economic distance [31], etc. In terms of the theories used, these studies are mainly from the perspectives of institutional theory and organizational learning theory. And in terms of the research methods used, the classical discrete models such as logit or probit models [13] are mainly adopted.

However, due to the possible high-dimensional relationship between feature variables and M&A success, some studies have yielded contradictory results. For example, Fuad & Gaur and Zhu et al argue that the larger the deal size the higher the likelihood of a successful M&A transaction [32,33], but Dong et al suggest the opposite [15]. This is partly due to the fact that these studies consider the relationship between the variables and M&A outcomes as a simple linear relationship. On the other hand, different studies are based on specific theoretical contexts, and the mechanisms in the studies work in one context but do not necessarily hold in another. It can be seen that traditional empirical methods do have some limitations when facing complex M&A problems [13], so based on the ability of machine learning to inductively identify causal relationships, there is a need to introduce this powerful technique in the field.

2.2. M&A Prediction based on machine learning

Current research in the field of M&A predictive modeling can be divided into two main parts. One part of the research analyzes the characteristics of an organization and predicts whether the organization will become an acquirer or a target company in the future [34-41]. This type of research usually uses text analytics. For instance, Hajek & Henriques predicts whether a company will be a target for M&A in the future by extracting the news moods and news topics [35].

The other part of the study focuses on the pre-transactional stage, where the identity of the acquirer and the target company has been clarified but the M&A agreement has not yet been formalized, and the prediction target is whether the ongoing transaction will be successful or not. This study summarizes this part of the study in Table 1. The feature variables selected for this type of study can be classified into three levels, including the deal level, the firm level, and the macro level. Specifically, the deal level usually involves the capital size of that M&A transaction, the payment method, whether it is a cross-regional M&A, the termination costs of both parties, and the competitors' bids, etc. The firm level mainly involves the characteristics of the enterprise, which can be differentiated into financial and non-financial ones. The financial characteristics are some basic financial indicators include the share capital, the total assets, the sales growth rate and so on, while the non-financial characteristics include the corporate governance, environmental social responsibility, and technological indicators. Macro level variables are basically used in studies discussing cross-border mergers and acquisitions. In addition to focusing on the economic distance, cultural distance and political distance between the countries in which the M&A parties are located, these studies also discuss indicators related to environmental protection that have been of greater interest in recent years, such as the country's forest area, carbon dioxide emissions, and pollution indices.

In terms of the machine learning algorithms used, most of the studies used two and more algorithms and compared the prediction effectiveness between them, which can be categorized as decision trees and their ensemble ones, neural networks, clustering, and support vector machines. As for the most widely used decision trees and their ensemble ones among them, including AdaBoost, C4.5, LightGBM, GBDT (Gradient Boosting Decision Tree), and Random Forest Model (RFM), C4.5 is the base decision tree model, while the others are based on it. Utilizing the powerful classification performance of the decision tree itself, it is formed by different ensemble learning methods such as Boosting and Bagging, which better enhances the efficiency and accuracy of the model prediction.

While there are a number of machine learning models that can be used to predict M&A deals, it is likely that not all of them are suitable for solving this problem. From a probabilistic point of view, the probability of an M&A deal being successful is much greater than the probability of an M&A failure, which leads to a much smaller sample of M&A failures than successes. Thus, it leads to a data imbalance problem, and models employing this type of data will have biases oriented towards successful outcomes during training and testing [42]. From a practical point of view, incorrectly predicting actual failures as successes leads to even more serious consequences, as it may encourage M&A parties to continue incurring unnecessary sunk costs on a deal that has a high probability of failing. This is similarly explained in other areas. In the course of an audit by a CPA, judging fraud that has occurred as not having occurred will result in lawsuits, whereas judging fraud that has not occurred as having occurred simply results in over-auditing. Obviously, lawsuits result in more serious reputational and monetary losses [43]. Based on the above points, in the subsequent model construction, we need to focus on the predictive performance of both majority and minority classes to minimize the data imbalance problem.

Table 1. Summary of research on M&A prediction

Authors	Journal	Variables for M&A Prediction	Data Sources	Sample Time Period/ Size	Method
Hong et al. [44]	Sustainability	53 variables: **Macro-level country characteristics** (e.g., GDP, GDP per capita); **Macro-level geographic characteristics** (e.g., agriculture, forest area, natural resource endowment, land area, shared borders); **Macro-level climate characteristics (**e.g., PM 2.5 pollution, CO area, natural resource endowment, land area, shared borders); **Macro-level climate characteristics** (e.g., PM 2.5 pollution, CO2 emissions); Deal- level characteristics (transaction value, stake percentage of acquired value, land area, shared borders, etc.) level characteristics (transaction value,	MSCI ESG KLD. World Bank open data. Compustat. CRSP. Thomson Reuters SDC Platinum.	1973-2018 /215,160	AdaBoost. SVM

Table 1. (continued)

		percentage of stake acquired, payment); **Firm-level ESG characteristics** (e.g., environmental, social, governance scores); **Firm-level ESG characteristics** (e.g., environmental, social, governance scores); and **Firm-level ESG characteristics** (e.g., environmental, social, governance scores); **Firm-level financial characteristics** (e.g., size, growth, profitability).			
Zhang et al. [13]	Journal of World Business	59 variables: **Deal-level characteristics** (e.g., deal size, cash payment); **Firm-level characteristics** (e.g., acquirer experience, firm size); **Macro-level country characteristics** (e.g., institutional quality, cultural distance). **Firm-level characteristics** (e.g., acquirer experience, firm size); **Macro-level country characteristics** (e.g., institutional quality, cultural distance).	Thomson Financial Mergers & Acquisitions.	1980-2018 /24,693	LightGBM
S. Zhou et al. [45]	Digital Economy and Sustainable Development	50 variables: **Deal-level characteristics** (e.g., MAsize, MAtype, MApay, MAoffsite; **Firm-level financial characteristics** (e.g., ROA, volatility MAsize, MAtype, MApay, MAoffsite; **Firm-level financial characteristics** (e.g., ROA, volatility, sales, asset, and market value; **Firm non-financial characteristics** (e.g., CEO holdings, percentage of male directors, and MA experience; **Macro-level country characteristics** (e.g., policy uncertainty, CAR, city GDP).	CSMAR	2011-2021 /4,848	GBM. RFM. KNN. SVM
C.-S. Yang et al. [46]	Decision Sciences	43 variables: **Firm non-financial characteristics** (Technological quantity, Technological quality, Technological innovation, Technological diversity, Technological compatibility). diversity, Technological compatibility).	Thomson Reuters SDC Platinum. USPTO patent.	1997-2008 /779	C4.5. AdaBoost
Bi & Zhang [47]	PLOS ONE	35 variables: **Firm-level financial characteristics** (e.g., Assets of last year, Market value, Capital expenditure, Inventory, Total assets, Shareholders' equity, Total dividend, Working capital, Operating income, Net sales, Operating revenue). Shareholders' equity, Total dividend, Working capital, Operating income, Net sales, Operating revenue)	Hithink RoyalFlush Information Network. CSMAR.	2015-2019 /874	PSNN. GBDT

Table 1. (continued)

Lee et al. [42]	Journal of Business Research	17 variables: **Firm-level financial characteristics** (e.g., ROA, DPS, Inventory/Total Assets, Market-to-Book ratio, P/E ratio, Growth in Sales, Capital Expenditures/Operating Revenue, Inventory Turnover, Dividend Payout Ratio, Dividend Yield); **Deal-level characteristics** (e.g., Relative Deal Size, Acquirer Termination Fee, Target Termination Fee, Toe-Hold, Tender Offer, All-Cash Deal, Competing Bid, Hostile, and Termination Fee Clause, Target Price One Day Prior, Target Price One Week Prior, and Target Price Four Weeks Prior).	Securities Data Company's U.S. Mergers and Acquisitions. Compustat	2006-2015 /803	Generalized logit model; Neural network

2.3. Machine learning classification tasks and their unbalanced data

Machine learning has been widely used in classification tasks. For instance, people on a variety of algorithms to input gray scale images of integer matrices varying from 0 to 255 in order to complete the task of image classification. However, when facing binary categorical variables, many algorithms do not perform as well as expected, but this kind of situations often appear in our real life [48]. And the reason why many algorithms do not perform well is the imbalance of data classification. The problem occurs when the number in one category is significantly smaller than the number in others, which leads to a biased estimation of the outcome of the minority category. However, the minority category is often the most important target of interest in many cases, such as bankruptcy [49], rare disease [50] and so on.

According to previous research, the current solutions to the data imbalance problem are mainly categorized into two types: data level and algorithm level methods. Regarding data level methods, they mainly resample the dataset, including undersampling and oversampling. Among the oversampling techniques, Synthetic Minority Over-sampling Technique (SMOTE) is more widely used, which synthesizes the samples through the k-proximity algorithm to be more consistent with the actual situation of the data compared to the random oversampling method that directly copies the existing samples of the minority classes [51]. However, there are different opinions about them. Khushi et al used Random Forest and Random Oversampling (ROS) models on an unbalanced lung cancer dataset, both of which showed high AUC values as well as the lowest standard deviation [52]. In another study, the random forest model with random undersampling operation showed the best performance [53]. Differences in datasets and prediction goals may be responsible for the result [54], so it is necessary to try different approaches to determine the best model.

In contrast to data level methods that improve model performance by altering the data itself, algorithm level methods improve the performance of classifiers on unbalanced datasets by tweaking the mechanisms within the algorithm. These are subdivided into Cost-sensitive Learning and Ensemble Methods, and LightGBM excels in unbalanced data by combining these two methods. Specifically, LightGBM can improve the detection rate of minority classes by adjusting the class weights in the loss function to give higher penalties for misclassification of minority class samples [55]. On the other hand, as an ensemble learning algorithm, LightGBM improves the classification precision by combining multiple weak classifiers, which in itself enhances the ability to handle unbalanced data. Meanwhile, data level and algorithm level hybrid approaches are also recommended by many studies, such as EID-GAN [56]. We will try for these approaches to solve unbalanced data and select the optimal model from them.

3. Research design

3.1. Sample selection and variable definitions

The data of M&A transactions in this study come from CSMAR (China Stock Market & Accounting Research Database), which is widely used by related studies in the field of M&A [57,58]. The database consists of several sub-databases, different sub-databases provide different dimensions and levels of information. Table 2 shows the total database used and the five sub-databases under it. “Merger & Acquisition, Asset Restructuring” provides data on M&A transactions including acquirer's source of funds, value of expenditures, and whether it is a related transaction. “Industry Financial Indicators” provides industry level financial indicators. “China Listed Firm's Basic Information” provides firm level information such as registered capital, location and so on. “Corporate Governance” provides average data on industry regulators' remuneration, the number of executives and their remuneration in a specific firm, as well as other comprehensive information on governance. “Listed Firm's R&D and Innovation” provides information on R&D input of listed firms and other related general financial indicators. This study uses data on M&A transactions from 2003-2024 and merges it with data from four other sub-databases. This study removes samples that satisfy the following conditions: (1) Samples with a missing value for deal status are deleted because the construction of the model for predicting M&A deals in this study requires that the final deal status be used as the dependent variable; (2) Financial firms as well as ST or ST* firms; (3) Samples without the information about the acquirer or the target; and (4) Samples with a missing value for all the other variables used in this study. In addition to this, all unordered category variables were transformed into dummy variables taking the value of 0 or 1, the winsor method was used to reduce the effect of extreme values. Finally, all continuous variables were standardized, and 3672 samples were finally used in this study.

Table 2. Summary of database

	Database
CSMAR	Industry Financial Indicators China Listed Firm's Basic Information Corporate Governance Listed Firm's R&D and Innovation Merger & Acquisition, Asset Restructuring

By combing previous studies on potential factors of M&A success and based on the principle of data availability, this study identifies 50 predictors covering three levels of data namely deal level, firm level and industry level. As this study mainly focuses on China’s M&A deals, variables at the country level are not included, while the variables at the firm level are derived from the information of listed acquirers. The definitions, measurement, and sources of variables are shown in Table 3. Since the main research interest is whether M&A deals are successful or not, the dependent variable is binary, which is labeled 1 if the deal is successful, and 0 if the deal is not. The data of whether the deal is successful or not is provided in the CSMAR Merger & Acquisition database, and the variable is called IsSucceed.

Table 3. Variable definitions

Variables (short names)	**Measurement**	**Sources**
Dependent variable		
IsSucceed	Binary variable, 1 if the acquisition is successful, 0 otherwise.	M&A, Asset Restructuring
Independent variable		

Table 3. (continued)

	UnderlyingTypeID	Represents the type of underlying asset involved in the transaction. (1 = Asset; 2 = Equity; 3 = Asset and Equity).	
	ExpenseValue	Value of the expenditure by the buyer.	
	PayTypeID	Specifies the method used for payment in the transaction. (1 = Asset payment; 2 = Cash payment; 3 = Stock payment; 4 = Bond payment; 5 = Debt assumption; 6 = Cash and asset mixed payment; 7 = Cash and stock mixed payment; 8 = Cash and debt assumption mixed payment; 9 = Other payment methods; 10 = Stock and asset mixed payment). mixed payment).	
	SourceTypeID	Indicates the source of funds for the payment. (1 = Loan; 2 = Own funds; 3 = Bond issuance; 4 = Stock issuance; 5 = Loan and own funds; 6 = Loan and bond issuance; 7 = Loan and stock issuance; 8 = Own funds and bond issuance; 9 = Bond and stock issuance; 10 = Bond and stock issuance; 11 = Directed issuance; 12 = Treasury. = Loan and stock issuance; 8 = Own funds and bond issuance; 9 = Own funds and stock issuance; 10 = Bond and stock issuance; 11 = Directed issuance; 12 = Treasury stock; 13 = Stock held in other stock; 13 = Stock held in other companies; 14 = Other; 15 = Asset).	
Deal level	*MergerTypeID*	Specifies the type of M&A activity. (1 = Horizontal merger; 2 = Vertical merger; 3 = Mixed merger; 4 = Asset adjustment; 10 = Other).	
	IsIntelProMA	Binary variable, 1 if the transaction involves the transfer or purchase of patents, technology, or trademarks, 0 otherwise.	
	RestructuringTypeID	Specifies the type of restructuring activity. (1 = Asset acquisition; S3002 = Asset divestiture; 3 = Asset swap; 4 = Merger by absorption; 5 = Debt restructuring; 6 = Share repurchase; 7 = Tender offer; 8 = Equity transfer). restructuring; 6 = Share repurchase; 7 = Tender offer; 8 = Equity transfer).	
	MARegionTypeID	Indicates the geographic region type of the M&A activity. (1 = Domestic M&A; 2 = Outbound M&A; 3 = Inbound M&A; 4 = Offshore M&A).	
	CrossProvinceSign	Binary variable, 1 if the M&A activity crosses provincial boundaries, 0 otherwise.	
	CrossCitySign	Binary variable, 1 if the M&A activity crosses city boundaries, 0 otherwise.	
	RelevanceSign	Binary variable, 1 if the transaction involves related parties, 0 otherwise.	
	MajorRestructuringSign	Binary variable, 1 if the transaction constitutes a major asset restructuring, 0 otherwise.	
Firm level	*IndustryCode*	Indicates the industry classification of the company.	
	Lng	Specifies the longitude coordinate of the company's office address.	
	Lat	Specifies the latitude coordinate of the company's office address.	
	RegisterCapital	Registered capital.	China Listed Firm's Basic Information
	RegisterLongitude	Specifies the longitude coordinate of the company's registered address.	
	RegisterLatitude	Specifies the latitude coordinate of the company's registered address.	
	PROVINCE	Indicates the province where the listed company's registered address is located.	
	TotalAssets	The sum of all assets owned by a company, including current and non-current assets.	
	Total Liability	The total amount of liabilities a company owes to creditors, including both current and long-term liabilities.	Listed Firm's R&D and Innovation
	Intangible Asset	Non-physical assets that add value to a company, such as patents, trademarks, and goodwill.	
	ProfitParent	The net profit attributable to the parent company's shareholders, excluding minority interests.	

DOI: 10.54254/2754-1169/2025.BL31296

Table 3. (continued)

	NetProfit	The total earnings of a company after all expenses, taxes, and costs have been deducted from total revenue.	
	OperatingEvenue	The income generated from the core business operations of a company, excluding non-operating revenue.	
	OperatingCost	The expenses incurred in the process of generating operating revenue, including cost of goods sold and other operational expenses.	
	OperationProfit	The profit a company makes from its core business operations, calculated as operating revenue minus operating costs.	
	RDSpendSum	The total expenditure on research and development activities within a given period.	
	Y1001b	Whether the Chairman and General Manager are the same person. (1 = Same person; 2 = Different persons)	
	Y1401b	Method of disclosing annual salary in listed companies. (1 = Accurate disclosure of annual salary; 2 = Interval disclosure of annual salary; 3 = Both accurate and interval disclosure). (1 = Accurate disclosure of annual salary; 2 = Interval disclosure of annual salary; 3 = Both accurate and interval disclosure) 4 = Other disclosure methods	
	Y1701a	Total number of committees established during the reporting period.	
	Y1801b	Consistency of work locations between independent directors and the listed company. (1 = Same location; 2 = Different locations; 3 = Cannot determine)	
	TotalNumber	Total number of supervisory staff, including directors, supervisors, and senior management personnel.	
	FemaleNumber	Number of female supervisory staff.	
	DirectorNumber	Number of directors, including the chairman.	
	FemaleDirectorNumber	Number of female directors.	
	ManagerNumber	Number of senior management personnel disclosed in the annual report. Senior management personnel include general managers, presidents, CEOs, deputy general managers, vice presidents, secretaries of the board, and other managerial personnel announced in the annual report (including those who also senior management personnel include general managers, presidents, CEOs, deputy general managers, vice presidents, secretaries of the board, and other managerial personnel announced in the annual report (including those who also The Board of Directors is responsible for the management of the company.)	Corporate Governance
	FemaleManagerNumber	Number of female senior management personnel.	
	Holdshares	Number of shares held by supervisory staff, including directors, supervisors, and senior management personnel. In cases of concurrent positions, the number of shares is not counted multiple times. In cases of concurrent positions, the number of shares is not counted multiple times.	
	DirectorHoldshares	Number of shares held by the board of directors, including concurrent positions.	
	ManageHoldshares	Number of shares held by senior management personnel, including concurrent positions.	
	InduCoAvgSalary	$\frac{\text{Total Salary of Industry Employees}}{\text{Number of Companies in the Industry}}$	
	InduFemaleSumSalaryRatio	$\frac{\text{Total Salary of Female Supervisory Staff (Excluding Allowances)}}{\text{Total Salary of Supervisory Staff (Excluding Allowances)}} \times 100\%$	
	InduMaleSumSalaryRatio	$\frac{\text{Total Salary of Male Supervisory Staff (Excluding Allowances)}}{\text{Total Salary of Supervisory Staff (Excluding Allowances)}} \times 100\%$	
Industry level	*InduFSumSalaryRatioAllow*	$\frac{\text{Total Salary of Female Supervisory Staff (Including Allowances)}}{\text{Total Salary of Supervisory Staff (Including Allowances)}} \times 100\%$	
	InduMaleSumSalaryRAllow	$\frac{\text{Total Salary of Male Supervisory Staff (Including Allowances)}}{\text{Total Salary of Supervisory Staff (Including Allowances)}} \times 100\%$	
	SampleNumber	Total number of companies in the same industry.	Industry Financial Indicators
	HHI_A	Herfindahl-Hirschman Index for measuring market concentration, $HHI = \sum_{i=1}^{N} s_i^2$.	

Table 3. (continued)

HHI_B	Herfindahl-Hirschman Index for measuring market concentration, . $HHI = \left(\sum_{i=1}^{N} s_i\right)^2 = 1$.
HHI_C	Herfindahl-Hirschman Index measuring market concentration, . $HHI = \left(\sum_{i=1}^{N} s_i^2\right) \times 10{,}000$.
HHI_D	Herfindahl-Hirschman Index measuring market concentration, . $HHI = \frac{\sum_{i=1}^{N} s_i^2 - \frac{1}{N}}{1 - \frac{1}{N}}$.

3.2. Algorithm selection

We refer to the process of Zhang et al screening the optimal algorithm [13] : (1) Divide all the data into training set and test set, this step is also to avoid overfitting [59] ; (2) Use different algorithms on the training set, and the results of these algorithms can be cross-validated; (3) Use different algorithms on the test set, plot the corresponding confusion matrix as well as compute the recall, accuracy, precision, and F1 score of different algorithms, and determine the optimal one by comparing the results.

According to the process, we firstly need to determine the proportion of the training set and the test set, according to Nguyen et al, when the proportion of the training set is increased from 30% to 80%, the performance of the model shows a large improvement, while it shows a downward trend when the proportion continues to be increased to 90% [60]. Moreover, Gholamy et al pointed out that using 20%-30% of the data for the test set and the rest for the training set will give optimal results [59]. Therefore, in our study, two ratios of 7:3 and 8:2 were tried, and according to the final results, the 8:2 ratio can get the best performance.

Regarding the choice of specific algorithms, previous literature has found that decision trees and their ensemble algorithms, neural networks, and logistic regression all exhibit better performance in facing classification tasks [61,62]. In addition, considering the data imbalance, it has been pointed out in the literature that neural networks and logistic regression significantly underperform decision trees and their ensemble algorithms [63]. But the bias of results brought about by such algorithms in the face of data imbalance can be effectively mitigated by oversampling or under sampling [64,65]. Therefore, instead of abandoning the use of neural networks and logistic regression, we use the data with oversampling and under sampling to train and test the models, and finally compares all the results to determine the best one.

3.3. Algorithm tuning

Firstly, this study suffers from a more serious data imbalance problem, with approximately only 15% of the data being M&A failures and the other 85% being successes. We deal with this problem through data level, algorithm level, and hybrid approaches. Regarding data level methods, they are mainly categorized into oversampling and under sampling, where oversampling reduces the data imbalance by adding samples and under sampling balances the dataset by removing samples. In this study, we tested the effect of random oversampling, SMOTE oversampling, Tomek Links under sampling, random under sampling, cluster-centered undersampling and Edited Nearest Neighbours (ENN) under sampling applied on different algorithms. The results show that after SMOTE oversampling or Edited Nearest Neighbours (ENN) under sampling, the models perform better than the original ones. This may be due to the fact that SMOTE oversampling, compared to random oversampling, generates samples based on the features of the existing samples, which can avoid replicating noisy data to some extent. While Edited Nearest Neighbours (ENN) under sampling is better able to retain representative samples compared to random under sampling and avoids blind deletion, and also ensures that minority class samples are not overly deleted compared to the Tomek Links method. We also adopt a hybrid algorithm level and data level approach to solve the problem, which yields the best results.

DOI: 10.54254/2754-1169/2025.BL31296

Specifically, in this study, the data were firstly processed with SMOTE oversampling, and then trained and predicted by lightGBM, an integrated algorithm based on multiple CART decision trees, which has been found to have better results in previous studies involving unbalanced data.

Secondly, this study suffers from overfitting problem in the process of constructing the neural network model. The underlying neural network gradually decreases the prediction correctness of the test data and increases the prediction correctness of the training data as the number of iterations increases, which reminds that the overfitting may arise and ultimately leads to a decrease in the accuracy of the results. Therefore, we used the following six main strategies to solve the above problem: (1) Use the L2 function to make the curve during training smoother and reduce the slope of the training curve rising, thus reducing the gap between the two; (2) Use dropout in the training process, which means that the training needs to be given a dropout rate. In this study, the dropout rate varies from 0.2 to 0.5, which is used to remove part of the data in training. The occurrence of overfitting may indicate that the dropout rate is not high enough; (3) Use the early stopping, which is equivalent to give a limit to the fitting of neural networks. If the loss in the validation set does not improve after 10 consecutive epochs, the training will stop.10 is the result of multiple tests, if this number is low, then the number of iterations will be low and underfitting, and vice versa for overfitting; (4) Simplify the model structure to reduce the number of layers or the number of neurons in each layer; (5) Use Batch Normalization regularization as it introduces noise to some extent,; (6) Dynamically adjust the learning rate by reducing the learning rate by half every 5 epochs when the loss in the validation set is no longer decreasing.

Thirdly, we used both automatic parameter optimization and manual parameter tuning. Regarding neural networks, it includes four main parts: input layer, hidden layer, output layer and compiled model. In the process of model optimization, we need to try different numbers of hidden layer layers, different numbers of neurons and activation functions for each layer, as well as the optimizer learning rate for the compiled model part. There are a large number of possible combinations. For the sake of tuning efficiency and the concern of missing the potentially optimal combinations, we use the Keras Tuner automated parameterization tool to automatically search the combinations as many as possible and we set the criterion that the highest precision is determined to be the optimal model. Regarding the decision tree, we use the grid search to find the optimal combination of maximum depth, minimum number of samples required for node splitting, and minimum number of samples for leaf nodes through cross-validation in the pre-pruning session. While in the post pruning model, the same technique is used to find the optimal cost-complexity pruning path.

3.4. Assessment of indicators

Referred to previous research, we evaluate model performance using four indicators, which are precision, recall, F1 score, and accuracy. Since the actual failure of an M&A deal that is incorrectly predicted as a success results in the largest loss, we pay the most attention to precision among the four indicators, which calculates the proportion of all predicted successes that are actually successes as well. In the subsequent model comparisons, precision will be the main basis of judgment.

In addition to this, indicators provided by the program by default represent the prediction performance based on the majority class. However, we also focus on the minority class, it may be not sufficient if only the default indicators are reported. Therefore, this study also reports the weighted indicators in the subsequent model evaluation, and unlike the treatment of macro-ones, the majority and minority categories are given different weights based on the proportion of the total sample size that they occupy, which enables to focus on two categories at the same time. The reason why macro indicators are not considered in this study is that they give the same weight to the majority and minority categories, which may focus excessively on the minority one. Although a failed M&A can cause significant losses to the firm, a successful M&A is a more probable event, and the weighted indicators are obviously closer to the reality. Table 4 demonstrates the specific formula for the indicators.

Table 4. Details of indicators

Indicators	Measurement
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1 score	$2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Weighted Precision	$\sum_{i=1}^{N} w_i \cdot P_i$
Weighted Recall	$\sum_{i=1}^{N} w_i \cdot R_i$
Weighted F1 Score	$\sum_{i=1}^{N} w_i \cdot F1_i$

4. Results

4.1. Model performance analysis

Firstly, the results of logistic regression models are analyzed in this study. Basic model is constructed without data resampling. Each model exhibits two types of indicators we have introduced before, with original being the default output of the model and also representing the prediction performance for the positive class (also the majority class). This study also focuses on the predictive ability for the minority class, while the weighted metrics provide a convenient way to look at the predictive ability for both the positive and negative classes at the same time. If the weighted indicators decrease compared to the original ones, the predictive performance for the minority class can be considered as insufficient. At the same time, this study focuses on the precision since the incorrect prediction of failed M&A is the worst situation.

As can be seen from Table 5, the original precision of all three models is greater than the weighted precision (0.88>0.84, 0.93>0.84, 0.89>0.83), which indicates that even with the use of resampling techniques to alleviate the data imbalance, the logistic regression still shows an insufficient predictive performance of the minority class. In particular, for the model treated with oversampling, the prediction performance for the positive class is improved by 5% (0.93-0.88), but the weighted precision is instead equal to the weighted indicator of the basic model (0.84), which suggests that the prediction performance for the minority class is decreasing, further widening the gap between the original and the weighted indicators. Therefore, using weighted precision as a criterion, this study concludes that the best model for logistic regression is the one obtained from training based on raw data.

Table 5. Result of logistic regression

		Precision	Recall	F1-score	Accuracy
Basic	*Original*	0.88	0.98	0.93	0.87
	Weighted	0.84	0.87	0.84	
Oversampling	*Original*	0.93	0.74	0.82	0.72
	Weighted	0.84	0.72	0.76	
Undersampling	*Original*	0.89	0.97	0.93	0.87

Table 5. (continued)

	Weighted	0.83	0.87	0.84	

The results of neural networks are shown in Table 6. Comparing with the basic model of logistic regression. The neural network basic model improves its weighted precision by 2% (0.86-0.84) while keeping original precision constant, indicating a slight improvement in the overall predictive performance. However, after the data resampling, neural networks show a similar problem with logistic regression. Although the predictive ability for majority class goes up, the weighted precision does not change or even decreases. Therefore, the best model in neural networks is the basic model.

Table 6. Result of neural network

		Precision	**Recall**	**F1-score**	**Accuracy**
Basic	*Original*	0.88	1.00	0.93	0.87
	Weighted	0.86	0.87	0.83	
Oversampling	*Original*	0.90	0.90	0.90	0.82
	Weighted	0.82	0.82	0.82	
Undersampling	*Original*	0.92	0.82	0.87	0.78
	Weighted	0.84	0.78	0.80	

The decision tree used in this study is the CART algorithm and its results are shown in Table 7. Comparing within the decision tree models, the best performing model is basic post pruning one. While comparing to the neural networks, the decision tree has limited improvement in prediction performance, both for the majority class and the minority class. Nonetheless, decision trees are fewer complex models than neural networks, which means they can take up less computational resources and provide faster operation in the practical application. With this in mind, the decision tree can still maintain over 90% precision in predicting majority class and over 80% weighted precision, except for the post-sampling pruning model.

Table 7. Result of decision tree

		Precision	**Recall**	**F1-score**	**Accuracy**
Basic pre pruning	*Original*	0.92	0.73	0.82	0.72
	Weighted	0.83	0.72	0.76	
Basic post pruning	*Original*	0.94	0.71	0.81	0.71
	Weighted	0.85	0.71	0.75	
Oversampling pre pruning	*Original*	0.90	0.87	0.88	0.80
	Weighted	0.82	0.80	0.81	
Oversampling post pruning	*Original*	0.89	0.87	0.88	0.79
	Weighted	0.81	0.79	0.80	
Undersampling pre pruning	*Original*	0.92	0.67	0.77	0.66
	Weighted	0.83	0.66	0.71	
Undersampling post pruning	*Original*	0.94	0.59	0.72	0.61
	Weighted	0.84	0.61	0.67	

Finally, in order to further improve the model performance, we use an integrated CART-based algorithm, namely the LightGBM model. Table 8 demonstrates the obtained results. After SMOTE oversampling of the data, the lightGBM model exhibits a leap in predictive performance, with all the original and weighted indicators reaching more than 90%. Specifically, about the precision, the weighted one exceeds the ORIGINAL one (0.95 > 0.93), which indicates that the model outperforms both for prediction of the majority class and the minority class. Other than that, there is no significant difference in performance between the other models of lightGBM and the previously constructed models.

DOI: 10.54254/2754-1169/2025.BL31296

Table 8. Result of lightGBM

		Precision	Recall	F1-score	Accuracy
Basic	*Original*	0.89	0.97	0.93	0.87
	Weighted	0.85	0.87	0.85	
Oversampling	*Original*	**0.93**	**0.98**	**0.95**	**0.95**
	Weighted	**0.95**	**0.95**	**0.95**	
Undersampling	*Original*	0.90	0.98	0.94	0.89
	Weighted	0.87	0.89	0.87	

4.2. Model selection

Table 9 summarizes the results of the best models. Among these four models, the lightGBM with SMOTE oversampling was selected as the optimal model using weighted precision as the judgment criterion.

Table 9. Total result

		Precision	Recall	F1-score	Accuracy
LightGBM	*Original*	**0.93**	**0.98**	**0.95**	**0.95**
(Oversampling)	*Weighted*	**0.95**	**0.95**	**0.95**	
Logistic Regression	*Original*	0.88	0.98	0.93	0.87
(Basic)	*Weighted*	0.84	0.87	0.84	
Neural Network	*Original*	0.88	1.00	0.93	0.87
(Basic)	*Weighted*	0.86	0.87	0.83	
Decision Tree	*Original*	0.94	0.71	0.81	0.71
(Basic post pruning)	*Weighted*	0.85	0.71	0.75	

4.3. Predictor importance analysis

Based on the optimal model constructed in this study (SMOTE oversampled lightGBM), we use SHAP (SHapley Additive exPlanations) to further analyze the specific contribution and influence direction of each predictor to the model. Figure 1 illustrates mean SHAP values of predictors in the test set in reverse order from largest to smallest, with higher values demonstrating that the feature variables contribute more to the model. Among the top twenty feature variables, all levels are involved except for the industry level variables. The variables at the deal level are in the relative top position, including the payment method, the value of the buyer's expenditure, the source of funds of the acquirer, the type of M&A, and whether it is cross-province. After that, there are corporate governance-related variables at the company level, including the number of female directors, the number of female executives, the consistency of the independent director's workplace with that of the listed company, the total number of shareholders, the number of senior executives, the net profit attributable to the listed company's shareholders, whether the top ten shareholders of the acquirer are related, and the number of directors. Finally, we can see the company-level financial variables, including total assets, and the amount of R&D investment.

DOI: 10.54254/2754-1169/2025.BL31296

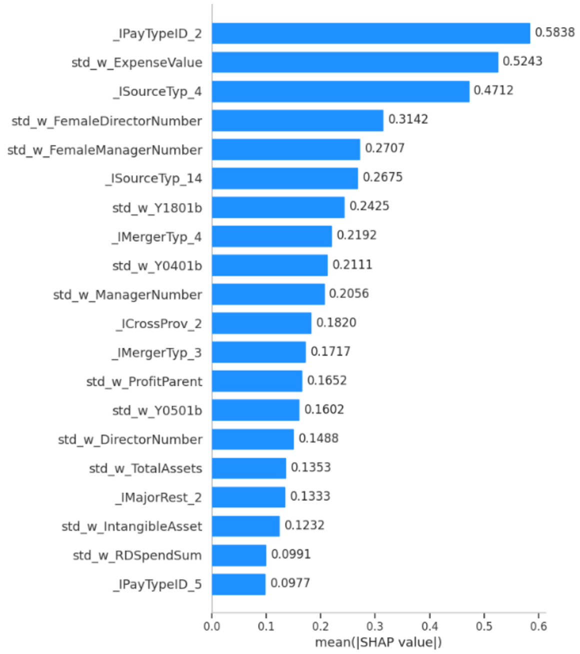

Figure 1. Mean SHAP values of predictors in the test set

Meanwhile, in order to further gain the influence direction of each predictor on the dependent variable, Figure 2 is plotted and it is ordered in the same way as Figure 1. Each point in the Figure 2 represents a sample, and points may be stacked together to form larger clumps. With the vertical axis in the center as zero, the right side represents the influence of the variable in the positive direction and the left side represents it in the negative direction. Colors represent predicted values, with red representing high eigenvalues and blue representing low eigenvalues. Taking _IPayType_2 as an example, the red point on the right side indicates that the samples with high eigenvalue have a positive influence in the model, while the blue point on the left side indicates that the samples with low eigenvalue have a negative influence on the model, which means that cash payment positively influences the success of M&A. Besides, std_w_ExpenseValue, whose blue and red scatters are not more cleanly distributed around the zero point, may imply that it has a non-linear relationship with M&A success or an interaction effect with other feature variables.

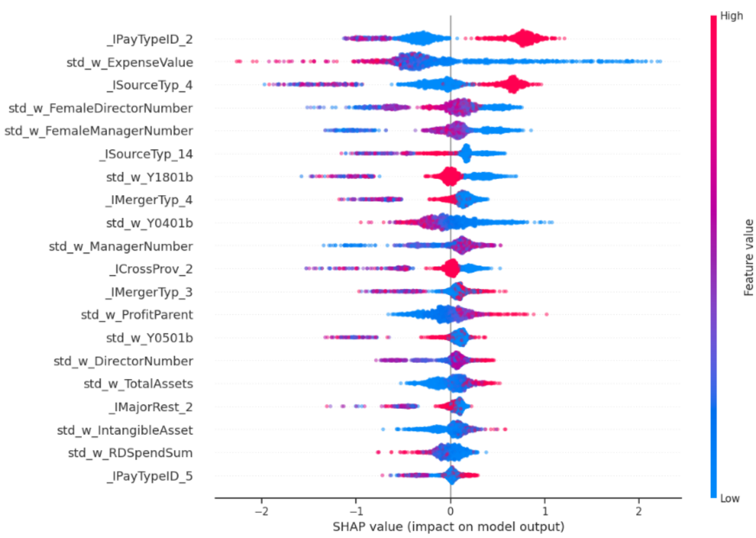

Figure 2. SHAP values of the predictors

DOI: 10.54254/2754-1169/2025.BL31296

Since we transformed some of the categorical variables into multiple dummy variables in the previous period, we recombined some dummy variables into one multi-category variable for interpretation in the subsequent analysis. Based on the results of the SHAP values and previous empirical studies, the top twenty contributing predictors can be devided in two categories: high contributing and discussed by previous studies, and high contributing but not discussed by previous studies.

Regarding the feature variables with high contribution and discussed by previous studies, they include whether cash payments are used, the acquirer's expense value, whether the deal is cross-regional, the number of female directors, the number of female executives, and net profit attributable to shareholders of listed companies.

Some of the studies are consistent with the conclusions drawn from previous empirical studies. Specifically, the use of cash payments tends to signal that the acquirer has stronger cash flow and financial capability [66], and the acquirer also has confidence in the value of the target company, which makes it easier to facilitate the M&A deal. Another possibility is that the use of cash payment avoids dilution of existing shareholders' interests by the acquirer and thus is less likely to be opposed by existing shareholders during the M&A process [67]. Regarding whether the deal is cross-regional, it is certain that non-cross-regional M&A is more likely to be successful compared to the cross-regional M&A. This is because even if both firms are located in China, cross-regional M&A faces the obstacles of inconsistencies in policies, economic development, etc. And that’s why the center government of China has introduced a number of policies to help M&A to clear cross-regional obstacles. Regarding the relationship between acquirer's net profit attributable to shareholders of listed companies and M&A success, although previous studies do not directly use net profit as an independent variable, it is considered as an indicator reflecting a firm's financial status, indirectly affects the success rate of M&A. For example, Huang's study points out that better-profitable firms are more inclined to execute M&A deals [68].

In addition, the relationship between some of the variables and M&A success is complex. According to the SHAP results, the acquirer's expense value and M&A success may be non-linear. Within a certain limit, they are positively correlated, and above this threshold, a higher expense value leads to a higher probability of failure. Previous research presents a contradictory view, with Bessler and Schneck arguing that an "excessive" expense value significantly increases the probability of a deal's success, even though it exceeds the expected level [69], while other research suggests that too much expense value may lead to M&A failure because the market may perceive that the acquirer is overvaluing the target company, which in turn leads to shareholder opposition to the deal[70]. The impact of the number of female directors and the number of female executives on the M&A success is similar, probably because both relate to the theme of making strategic decisions by women. According to previous studies, female directors are more cautious in M&A decision making and are more concerned about potential risks, which leads to a negative correlation between the percentage of women on the board of directors and the likelihood that the company will undertake a M&A [71,72]. This finding also coincides with the blue portion on the right side of this variable in Figure 2. However, we also observe the presence of clusters of blue and red mixed together on the right side of the vertical axis closest to the middle, implying that the effect of the number of female directors and the number of female executives on the M&A success is not simply a linear relationship either. We hypothesize that this is because there are also different leadership styles and personality traits in the group of female executives, with the presence of female leaders who prefer aggressive decision-making as well as risk averse male leaders. This conclusion inspired us to verify for the interaction effect subsequently.

Regarding the feature variables with high contribution but not discussed in previous studies, they include the source of the acquirer’s funds, the type of M&A, the consistency of the independent

DOI: 10.54254/2754-1169/2025.BL31296

director's workplace with that of the listed company, the total number of shareholders, the number of senior executives, the existence of affiliation of the top ten shareholders, the number of board of directors, the total assets, and the amount of research and development (R&D) investment.

The source of the acquirer's funds in the top twenty include whether the acquirer raises funds by issuing shares and other non-mainstream fundraising methods. The use of non-mainstream methods of fundraising roughly presents an inverse relationship with M&A success, since it may have payment risks and are not easily recognized by the target company to some extent. Types of M&A can be divided into four situations, namely horizontal, vertical, hybrid and asset adjusted. The results show that asset-adjusted M&A is more likely to be fruitful, probably because it tends to involve only the transfer of specific assets or businesses, and it usually does not involve significant antitrust or other regulatory issues. About the independent directors' workplace congruence with listed companies, previous studies have shown that it leads to greater difficulty for the independent director to obtain information and then block the directors to effectively monitor the company if they are in different locations [73]. Although based on this finding we can hypothesize that independent directors can provide more constructive opinions on M&A activities by receiving information in a timelier manner. However, from another perspective, independent directors work in the same place as listed companies may make them more susceptible to local social, political, or economic factors, resulting in their decisions being interfered with by external pressures or personal interests, which reduces their ability to objectively and impartially evaluate M&A deals and ultimately reduces the success rate of it.

The number of shareholders in this study shows an inverse relationship with M&A success. That might because when the number of shareholders is large and dispersed, it is difficult for shareholders to reach a consensus among themselves. Although both the shareholders and are senior managers involved in decision-making on strategic activities, the scatterplot of the number of senior managers shows mixed clusters, which means further interaction effect validation is needed to determine the direction of the effect.

There is uncertainty about the impact of the existence of affiliation of the top ten shareholders within the acquirer on M&A success. This may be due to the fact that when there is a strong relationship between the shareholders, these shareholders might have aligned goals and interests in decision making. However, on the other hand, these relationships may lead to poorer decision quality and lower shareholder value creation, which may affect the success of the M&A. The number of board members usually reflects the size of the board, and firms with larger boards perform better in cross-border M&As, and these firms create higher value for shareholders after the deal, especially in cross-border M&As in the manufacturing industry [74]. Although post-merger success is not the same thing as the successful signing of M&A agreements discussed in this study, we can refer to the former mechanism that larger boards can provide more oversight and advice, which can help to improve the quality and execution of M&A decisions.

About the financial variables. The total assets of the acquirer may indirectly increase the likelihood of M&A success through enhanced financial performance and increased market confidence. For acquirer's R&D investment, although intuitively we may think that acquirer's R&D investment can enhance its negotiating position in M&A, especially if the target firm has important technological assets, high R&D investment implies the acquirer's advantage in terms of technological integration and innovation capability, which can help to increase the success rate of M&A. However, according to the results of this study, although there are fewer red scatters, they roughly show an inverse effect. We hypothesize that this may be due to uncertainty about the compatibility of the acquirer's and target's technologies.

DOI: 10.54254/2754-1169/2025.BL31296

5. Discussion

This study focuses on the domestic M&A deals of Chinese listed companies from 2003 to 2024, and constructs three levels of feature variables at the deal level, firm level, and industry level. Meanwhile, this study utilizes machine learning to explore the predictive ability of 50 feature variables on the M&A success, and finally obtains the following results: (1) In terms of the predicted performance, the SMOTE oversampling-treated lightGBM model performed the best. Various resampling methods are applied into neural network, logistic regression and CART models, but the improvements are limited. The unprocessed lightGBM model also performs in average. These results suggest that a hybrid of data level and algorithmic approaches is more effective in dealing with the classification problem of data imbalance. (2) The importance of predictors was analyzed using the SHAP technique, and among the top twenty predictors, there were no predictors at the industry level, which implies that the influence of the industry environment on firms is relatively minor compared to the influence of other level factors. (3) This study finds that some of the feature variables have not been discussed in previous studies, including the acquirers' source of funds, the type of M&A, the consistency of the independent director's workplace with the listed company, the total number of shareholders, the number of senior managers, the existence of the top ten shareholders' affiliation, the number of the board of directors, the total assets, and the amount of R&D input, and we makes inference about the mechanisms involved through some indirect evidences. (4) For the feature variables that have been studied whether cash payment is used, whether cross-regional mergers and acquisitions are made, and the acquirer's net profit attributable to shareholders of the listed company are consistent with the findings of previous studies. Besides, this study also finds a more complex and high-dimensional relationship between some feature variables (buyer's expense value, the number of female directors, and the number of female executives) and M&A success, which has only been roughly defined as a simple linear relationship in the past.

5.1. Theoretical implications

Previous empirical studies have explained the potential factors of M&A outcomes based on different theoretical backgrounds, advancing our understanding of M&A success in different settings from different perspectives. However, this heterogeneity prevents us from obtaining a comprehensive understanding of the potential factors of M&A success, which is provided by the ability of machine learning to handle high-dimensional relationships. The results of the best model in this study show that deal level information and firm level corporate governance related variables are relatively the most important for the successful completion of an M&A deal. Firm level financial related variables, although also appearing in the top twenty list of contributions, are relatively low position on the list. Besides, industry related variables such as the degree of industry competition are not as important as intuitively perceived and have the most minimal role to play in it. The reason why the degree of industry competition was considered to be an important factor for M&A success may be due to the fact that in the theoretical context at that time, only the degree of industry competition was considered, while the truly important factors were masked.

And some of the results of this study are inconsistent with previous studies, which suggests that the theoretical perspectives of existing studies may be insufficient. For example, the buyer's expense value is positively correlated with the success of M&A in the initial stage but is reversed to a negative correlation when it reaches the critical value, and finally shows a U-shaped curve. But the existing studies only discuss the first or second half, which needs a theory to be harmonized, especially this theory can clarify when this critical value will be reached.

5.2. Practical implications

Constructing a machine learning prediction model for M&A success has significant practical implications for both the enterprise itself and its multiple stakeholders. For enterprises, first of all, traditional decision-making methods often rely on expert experience and limited financial indicators, while machine learning models can comprehensively consider all kinds of factors and make the prediction results more reliable by analyzing a large amount of historical data and multidimensional variables. Thus, the lightGBM model with SMOTE oversampling constructed in this study can help enterprises make more accurate decisions in the process of M&A. In particular, the model constructed in this study improves the prediction of a small number of categories, and enterprises can focus their resources on deals with higher success rates and reduce their investment in projects with low probability of success, thus improving the overall return on investment.

For third-party service organizations, such as consulting firms, investment banks and law firms, through the method of constructing the model in this study, combined with their own data, they can provide their clients with more accurate due diligence, risk assessment and decision-making advice, which can improve their market competitiveness. In addition, the automation feature of the model can also help these organizations to improve their work efficiency and reduce the time and cost of manual operations.

For other stakeholders, such as investors, suppliers and customers, the forecasting model can also be used to alleviate the information inequality between them and the focused enterprises. Investors can assess the potential rewards and risks of M&A deals and make more rational investment decisions, while suppliers and customers can understand in advance the possible changes in the market brought about by M&A and adjust their own strategies to adapt to the new market landscape.

5.3. Limitations and future research

Although this study uses machine learning methods to alleviate the problems that may arise from traditional empirical methods and to mine some of the missed predictors, there are still some limitations. First, while machine learning methods help explore high-dimensional, nonlinear causal relationships, we do not know the exact mechanisms involved. Neural networks, in particular, require only a large number of input variables from which it autonomously searches for combinations of feature variables that are related, which reduces the interpretability of the theory. Although we can use the SHAP technique to clarify the direction and extent of the effect of some variables on the M&A results, there are some other variables whose scatters show a more confusing shape, which may be due to the high-dimensional relationship or interactive effect. What exactly is the reason for this and through which variables the interactions take place, this study is limited by data availability and does not provide a clear conclusion.

Second, although we have tried our best to include enough variables, due to data availability, we have censored some variables with more than half of the missing values such as the information about the third-party organizations in the deal, whether the acquirer is a state-owned enterprise, etc. In addition, there are some variables that are hard to obtain through the secondary data, such as executives' attitudes towards the M&A, the executives' personality traits and so on. These omitted variables are likely to be important predictors of M&A success.

Third, due to the availability of data, this study only focuses on firms listed in China. Due to the disparity in culture, institutions, and economy between countries, the model constructed from Chinese M&A data alone may be biased when used to predict the M&A status of firms in other countries.

Based on the limitations of the above research, this study proposes directions for future research. First, we can use the machine learning as an exploratory study, based on which we can add traditional empirical models to discuss the specific mechanism of action between feature variables and M&A

results. At the same time, we can apply more powerful algorithms with better performance in the field of machine learning, such as Generalized Linear Mixed-Effects Modeling Tree (GLMM Tree) [75], Mixed-Effects Random Forest (MERF) [76] and Derived Mixed-Effects Machine Learning (MEml) [77]. Second, we can consider more potential variables from other dimensions, such as policy implementation, digital economy level of a country, acquirers' M&A experience. Third, incorporate M&A deal data from other countries and consider variables that characterize cross-country contexts.

Acknowledgments

Xuesheng Meng: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft, Writing review & editing. **Rujin Huo:** Formal analysis, Methodology, Writing – original draft. **Zhichao Hao:** Formal analysis, Visualization.

We thank Cosimo Arnesano, Chenye Zhong for invaluable guidance and comments throughout this project.

References

[1] Christofi, M., Vrontis, D., Thrassou, A., & Shams, S. M. R. (2019). Triggering technological innovation through cross-border mergers and acquisitions: a micro-foundational perspective. technological Forecasting and Social Change, 146, 148-166. https://doi.org/10.1016/j.techfore.2019.05.026

[2] Hsu, P.-H., Huang, P., Humphery-Jenner, M., & Powell, R. (2021). Cross-border mergers and acquisitions for innovation. Journal of International Money and Finance, 112, 102320. https://doi.org/10.1016/j. jimonfin.2020.102320

[3] Idowu, S. O., Capaldi, N., Zu, L., & Das Gupta, A. (2013). Encyclopedia of corporate social responsibility. springer. https://doi.org/10.1007/978-3-642-28036-8_36

[4] Moschieri, C., & Campa, J. M. (2014). New trends in mergers and acquisitions: Idiosyncrasies of the European market. Journal of Business Research, 67(7), 1478-1485. https:// doi.org/10.1016/j.jbusres.2013.07.018

[5] Xue, H., Zeng, S., Sun, D., & Shi, J. J. (2024). Impacts of Peers' Mergers and Acquisitions on Firm Innovation. IEEE Transactions on Engineering Management, 71, 598-613. https://doi.org/10.1109/TEM.2021.3134698

[6] cole, r. a., ferris, k. r., & melnik, a. l. (2010). The Cost of Advice in Merger & Acquisition Transactions. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1458465

[7] pfizer. (2016). Pfizer announces termination of proposed combination with allergan | pfizer. https://www.pfizer.com/news/press-release/press-release-detail/ pfizer_announces_termination_of_proposed_combination_with_allergan

[8] Burlew, L. D., Pederson, J. E., & Bradley, B. (1994). The reaction of managers to the pre-acquisition stage of a corporate merger: A qualitative study. Journal of Career Development, 21(1), 11 -22. https://doi.org/10.1007/BF02107100

[9] Hawn, O. (2021). How media coverage of corporate social responsibility and irresponsibility influences cross-border acquisitions. strategic Management Journal, 42(1), 58-83. https://doi.org/10.1002/smj.3212

[10] Neuhauser, K. L., Davidson, W. N., & Glascock, J. L. (2011). An analysis of failed takeover attempts and merger cancellations. International Journal of Managerial Finance, 7(4), 347-376. https:// International Journal of Managerial Finance, 7(4), 347-376. doi.org/10.1108/17439131111166375

[11] bajpai, p. (2016). Fallout from the collapsed allergan-pfizer deal | nasdaq. https://www.nasdaq.com/articles/fallout-collapsed-allergan-pfizer-deal-2016-04- 07

[12] InvestorPlace.(2016). PFE stock: why the pfizer inc.-allergan plc mega-merger got murked (AGN) | nasdaq. https://www.nasdaq.com/articles/pfe-stock%3A-why-the-pfizer- inc.-allergan-plc-mega-merger-got-murked-agn-2016-04-06

[13] Zhang, J., Van Witteloostuijn, A., Zhou, C., & Zhou, S. (2024). Cross-border acquisition completion by emerging market MNEs revisited: Inductive evidence from a machine learning analysis. Journal of World Business, 59(2), 101517. https://doi.org/10.1016/j.jwb.2024.101517

[14] bettis, r. a., helfat, c. e., & shaver, j. m. (2016). The necessity, logic, and forms of replication. Strategic Management Journal, 37(11), 2193-2203. https://doi.org/10.1002/smj.2580

[15] dong, L., Li, X., McDonald, F., & Xie, J. (2019). Distance and the completion of Chinese cross-border mergers and acquisitions. Baltic Journal of Management, 14(3), 500-519. https:// doi.org/10.1108/BJM-06-2018-0223

[16] Li, Y., & Sai, Q. (2020). The effects of language and religion on cross-border acquisition completion. Research in International Business and Finance, 54, 101294. https://doi .org/10.1016/j.ribaf.2020.101294

[17] mullainathan, s., & spiess, j. (2017). Machine Learning: an Applied Econometric Approach. Journal of Economic Perspectives, 31(2), 87-106. https://doi.org/10.1257/jep.31 .2.87
[18] Welch, X., Pavićević, S., Keil, T., & Laamanen, T. (2020). The Pre-Deal Phase of Mergers and Acquisitions: a Review and Research Agenda. Journal of Management, 46(6), 843-878. https://doi.org/ 10.1177/0149206319886908
[19] Alimov, A., & Officer, M. S. (2017). Intellectual Property Rights and Cross-Border Mergers and Acquisitions. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2959703
[20] Becher, D. A., Cohn, J. B., & Juergens, J. L. (2015). Do Stock Analysts Influence Merger Completion? An Examination of Postmerger Announcement Recommendations. management science, 61(10), 2430 -2448. https://doi.org/10.1287/mnsc.2014.2065
[21] Zhang, J., Zhou, C., & Ebbers, H. (2011). Completion of Chinese overseas acquisitions: Institutional perspectives and evidence. International Business Review, 20(2), 226-238 . https://doi.org/10.1016/j.ibusrev.2010.07.003
[22] Martynova, M., & Renneboog, L. (2008). A century of corporate takeovers: what have we learned and where do we stand? Journal of Banking & Finance, 32(10), 2148-2177. https:// doi.org/10.1016/j.jbankfin.2007.12.038
[23] Harford, J. (1999). Corporate cash reserves and acquisitions. Journal of Finance, 54(6), 1969-1997. https://doi.org/10.1111/0022-1082.00179
[24] Bouwman, C. H. S., Fuller, K. P., & Nain, A. (2003). Stock market valuation and mergers. Mit Sloan Management Review, 45, 9-11.
[25] Uysal, V. B. (2011). Deviation from the target capital structure and acquisition choices. Journal of Financial Economics, 102(3), 602-620. https://doi.org /10.1016/j.jfineco.2010.11.007
[26] Hayward, M. L. A. (2002). When do firms learn from their acquisition experience? Evidence from 1990 to 1995. Strategic Management Journal, 23(1), 21-39. https:// Strategic Management Journal, 23(1), 21-39. doi.org/10.1002/smj.207
[27] Kennedy, K. H., Payne, G. T., & Whitehead, C. J. (2002). Matching Industries between Target and Acquirer in High-Tech Mergers and Acquisitions. Technology Analysis & Strategic Management, 14(2), 149 -162. https://doi.org/10.1080/09537320220133839
[28] Gorton, G., Kahl, M., & Rosen, R. J. (2009). Eat or Be Eaten: A Theory of Mergers and Firm Size. The Journal of Finance, 64(3), 1291-1344. https://doi.org/10.1111/j.1540-6261.2009. 01465.x
[29] Gracheva, M. (2023). Empirical analysis of risks' influence on the large mergers and acquisitions projects' success in modern companies. Economics and the Mathematical Methods, 59(1), 79. https://doi.org/10.31857/S042473880024877-3
[30] Zhou, C., Xie, J., & Wang, Q. (2016). Failure to Complete Cross-Border M&As: "To" vs. "From" Emerging Markets. journal of International Business Studies, 47(9), 1077-1105. https://doi.org/10.1057/s41267-016-0027-y
[31] lim, M.-H., & Lee, J.-H. (2017). National economic disparity and cross-border acquisition resolution. International Business Review, 26(2), 354-364. https://doi. org/10.1016/j.ibusrev.2016.09.004
[32] fuad, m., & gaur, a. s. (2019). Merger waves, entry-timing, and cross-border acquisition completion: a frictional lens perspective. Journal of World Business, 54(2), 107 -118. https://doi.org/10.1016/j.jwb.2018.12.001
[33] Zhu, P., Jog, V., & Otchere, I. (2014). Idiosyncratic volatility and mergers and acquisitions in emerging markets. Emerging Markets Review, 19, 18-48. https://doi.org/ 10.1016/j.ememar.2014.04.001
[34] aaldering, L. J., Leker, J., & Song, C. H. (2019). Recommending untapped M&A opportunities: a combined approach using principal component analysis and collaborative filtering. expert Systems with Applications, 125, 221-232. https://doi.org/10.1016/j.eswa.2019.02.004
[35] Hajek, P., & Henriques, R. (2024). Predicting M&A targets using news sentiment and topic detection. Technological Forecasting and Social Change, 201, 123270. https://doi.org/ 10.1016/j.techfore.2024.123270
[36] Jiang, T. (2021). Using Machine Learning to Analyze Merger Activity. Frontiers in Applied Mathematics and Statistics, 7, 649501. https://doi.org/10.3389/fams.2021. 649501
[37] Katsafados, A. G., Leledakis, G. N., Pyrgiotakis, E. G., Androutsopoulos, I., & Fergadiotis, M. (2024). Machine learning in bank merger prediction: a text-based approach. European Journal of Operational Research, 312(2), 783-797. https:// doi.org/10.1016/j.ejor.2023.07.039
[38] Li, Y., Shou, J., Treleaven, P., & Wang, J. (2021). Graph neural network for merger and acquisition prediction. Proceedings of the Second ACM International Conference on AI in Finance, 1-8. https://doi.org/10.1145/3490354.3494368
[39] Moriarty, R., Ly, H., Lan, E., & McIntosh, S. K. (2019). Deal or No Deal: Predicting Mergers and Acquisitions at Scale. 2019 IEEE International Conference on Big Data (Big Data), 5552-5558. https ://doi.org/10.1109/BigData47090.2019.9006015
[40] venuti, k. (2021). Predicting mergers and acquisitions using graph-based deep learning (No. arXiv:2104.01757). arXiv. https://doi.org/10.48550/arXiv.2104.01757

[41] Yang, D. (2024). Predicting mergers and acquisitions: Temporal dynamic industry networks (No. arXiv:2404.07298). arXiv. http://arxiv.org/abs/2404.07298
[42] Lee, K., Joo, S., Baik, H., Han, S., & In, J. (2020). Unbalanced data, type II error, and nonlinearity in predicting M&A failure. Journal of Business Research, 109, 271-287. https://doi .org/10.1016/j.jbusres.2019.11.083
[43] Hansen, J. V., McDonald, J. B., Messier, W. F., & Bell, T. B. (1996). A Generalized Qualitative-Response Model and the Analysis of Management Fraud. Management Science, 42(7), 1022-1032. https://doi.org /10.1287/mnsc.42.7.1022
[44] Hong, X., Lin, X., Fang, L., Gao, Y., & Li, R. (2022). Application of Machine Learning Models for Predictions on Cross-Border Merger and Acquisition Decisions with ESG Characteristics from an Ecosystem and Sustainable Development Perspective. and Sustainable Development Perspective. Sustainability, 14(5), 2838. https://doi.org/10.3390/su14052838
[45] Zhou, S., Lan, F., Hu, Z., & Liu, Y. (2024). Determinants of successful mergers and acquisitions in China: Evidence from machine learning. Digital Economy and Sustainable Development, 2(1), 4. https://doi.org/10.1007/s44265-024-00028-3
[46] Yang, C., Wei, C., & Chiang, Y. (2014). Exploiting Technological Indicators for Effective Technology Merger and Acquisition (M&A) Predictions. Decision Sciences, 45(1), 147 -174. https://doi.org/10.1111/deci.12062
[47] Bi, W., & Zhang, Q. (2021). Forecasting mergers and acquisitions failure based on partial-sigmoid neural network and feature selection. PLOS ONE, 16(11), e0259575. https://doi .org/10.1371/journal.pone.0259575
[48] Mahmudah, K. R., Indriani, F., Takemori-Sakai, Y., Iwata, Y., Wada, T., & Satou, K. (2021). Classification of Imbalanced Data Represented as Binary Features. Applied Sciences, 11(17), 7825. https://doi.org/10.3390/app11177825
[49] Amirshahi, B., & Lahmiri, S. (2024). Bankruptcy prediction using optimal ensemble models under balanced and imbalanced data. Expert Systems, 41(8), e13599. https://doi.org/10.1111/ exsy.13599
[50] Mahmudah, K., Purnama, B., Indriani, F., & Satou, K. (2021). Machine learning algorithms for predicting chronic obstructive pulmonary disease from gene expression data with class imbalance: proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, 148-153. https://doi.org/10.5220/ 0010316501480153
[51] Villar, A., & De Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. Discover Artificial Intelligence,. 4(1), 2. https://doi.org/10.1007/s44163-023-00079-z
[52] Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. iEEE Access, 9, 109960-109975. iEEE Access. IEEE Access, 9, 109960-109975. https://doi.org/10.1109/ACCESS.2021.3102399
[53] Malek, N. H. A., Yaacob, W. F. W., Wah, Y. B., Md Nasir, S. A., Shaadan, N., & Indratno, S. W. (2022). Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data. Indonesian Journal of Electrical Engineering and Computer Science, 29(1), Article 1. https://doi.org/10.11591/ijeecs.v29.i1.pp598-608
[54] Albisua, I., Arbelaitz, O., Gurrutxaga, I., Lasarguren, A., Muguerza, J., & Pérez, J. M. (2013). The quest for the optimal class distribution: an approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets. Progress in Artificial Intelligence, 2(1), 45-63. https://doi.org/10.1007/s13748-012-0034-6
[55] lin, L., Zhang, J., Zhang, N., Shi, J., & Chen, C. (2022). Optimized LightGBM power fingerprint identification based on entropy features. Entropy, 24(11), Article 11. https://doi.org/10.3390/e24111558
[56] Li, W., Chen, J., Cao, J., Ma, C., Wang, J., Cui, X., & Chen, P. (2023). EID-GAN: Generative adversarial nets for extremely imbalanced data augmentation. IEEE Transactions on Industrial Informatics, 19(3), 3208-3218. -3218. IEEE Transactions on Industrial Informatics. https://doi.org/10.1109/TII.2022.3182781
[57] Huang, C., Jin, Z., Tian, S., & Wu, E. (2023). The real effects of corruption on M&A flows: Evidence from China's anti-corruption campaign. Journal of Banking & Finance, 150, 106815. 106815. https://doi.org/10.1016/j.jbankfin.2023.106815
[58] Su, Z., & Xue, Y. (2023). Takeover deterrence with state ownership: Evidence from China. Journal of Banking & Finance, 146, 106689. https://doi.org/10.1016/j. jbankfin.2022.106689
[59] Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. International Journal of Intelligent Technologies and Applied Statistics, 11(2), 105-111. https://doi.org/10.6148/IJITAS.201806_11(2).0003
[60] Nguyen, Q. H., Ly, H.-B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I., & Pham, B. T. (2021). Influence of data splitting on the performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering, 2021, 1-15. https://doi.org/10.1155/2021/4832864
[61] Aytekin, C. (2022). Neural Networks are Decision Trees. arXiv.Org, abs/2210.05189. https://doi.org/10.48550/arXiv.2210.05189

DOI: 10.54254/2754-1169/2025.BL31296

[62] fayed, M. S. (2021). Classification of the Chess Endgame problem using Logistic Regression, Decision Trees, and Neural Networks. arXiv.Org, abs/2111.05976. https://doi .org/10.48550/arXiv.2111.05976
[63] brown, i., & mues, c. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446 -3453. https://doi.org/10.1016/j.eswa.2011.09.033
[64] he, X., & Zhang, J. (2018). Emerging market MNCs' cross-border acquisition completion: institutional image and strategies. Journal of Business Research, 93, 139 -150. https://doi.org/10.1016/j.jbusres.2018.04.014
[65] Zhang, J. (2022). Liability of emergingness and EMNEs' cross-border acquisition completion: a legitimacy perspective. International Business Review, 31(2), 101951. . 31(2), 101951. https://doi.org/10.1016/j.ibusrev.2021.101951
[66] Karampatsas, N., Petmezas, D., & Travlos, N. G. (2013). Credit Ratings and the Choice of Payment Method in Mergers and Acquisitions. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2203536
[67] Swieringa, J., & Schauten, M. (2007). The Payment Method Choice in Dutch Mergers and Acquisitions. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1018899
[68] Huang, L. (2010). Who Engages in Mergers and Acquisitions: Evidence from Chinese Listed Companies. 2010 2nd International Workshop on Database Technology and Applications, 1-4. Applications, 1-4. https://doi.org/10.1109/DBTA.2010.5658972
[69] bessler, w., & schneck, c. (2015). Excess premium offers and bidder success in European takeovers. Eurasian Economic Review, 5(1), 23-62. https://doi.org/10.1007/ s40822-015-0017-6
[70] Becht, M., Polo, A., & Rossi, S. (2019). Does Mandatory Shareholder Voting Prevent Bad Acquisitions? the Case of the United Kingdom. journal of Applied Corporate Finance, 31(1), 42 -61. https://doi.org/10.1111/jacf.12329
[71] Ahmed, Y., Song, Y., & Elsayed, M. (2022). On the likelihood and type of merger and acquisition in the US listed companies: the role of females on the board. international Journal of Accounting &amp ; Information Management, 30(3), 391-407. https://doi.org/10.1108/IJAIM-10-2021-0205
[72] Ossorio, M. (2020). Board Gender Diversity and Cross-border Mergers and Acquisitions: Evidence from European Deals. International Journal of Business and Social Science , 11(3). https://doi.org/10.30845/ijbss.v11n3a27
[73] Quan, Y., & Zhang, W. (2021). Geographic distance and board monitoring: Evidence from the relocation of independent directors. Journal of Corporate Finance, 66, 101802. https:// doi.org/10.1016/j.jcorpfin.2020.101802
[74] Datta, D. K., Basuil, D. A., & Agarwal, A. (2020). Effects of board characteristics on post-acquisition performance: a study of cross-border acquisitions by firms in the manufacturing sector. International Business Review, 29(3), 101674. https://doi.org/10.1016/j.ibusrev.2020.101674
[75] fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. Behavior Research Methods, 50(5),. 2016-2034. https://doi.org/10.3758/s13428-017-0971-x
[76] Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation, 84(6), 1313-1328. https://doi.org /10.1080/00949655.2012.741599
[77] Ngufor, C., Van Houten, H., Caffo, B. S., Shah, N. D., & McCoy, R. G. (2019). Mixed effect machine learning: a framework for predicting longitudinal change in hemoglobin A1c. Journal of Biomedical Informatics, 89, 56 -67. https://doi.org/10.1016/j.jbi.2018.09.001