Proceedings of ICMRED 2025 Symposium: Effective Communication as a Powerful Management Tool
DOI: 10.54254/2754-1169/2025.BL24140

The Application of Random Forest Algorithm for Company Valuation in the Advanced Materials Industry

Yuexuan Li

Foster School of Business, University of Washington, Seattle, USA
yl472@uw.edu

Abstract: The advanced materials industry is a key driver of technological innovation, making accurate enterprise valuation essential for investment and market analysis. Traditional valuation methods like DCF, PE, and PB struggle with high-growth companies due to volatile cash flows and market dependencies. To address these challenges, this study applies a random forest algorithm to enhance valuation accuracy by leveraging financial data, market indicators, and industry-specific factors. By using bootstrap aggregation to randomly select samples and features, the random forest model, which is based on an ensemble learning approach with decision trees—improves predictive performance and minimizes overfitting. In order to quantify important valuation determinants, build a predictive model, and assess its performance using common error measures, this study gathers financial and market data from about 100 publicly traded businesses in the new materials sector. When compared to conventional techniques, the empirical study shows that the random forest model increases valuation accuracy and stability. The findings show that the model delivers a more accurate estimate of enterprise value, lessens sensitivity to market swings, and successfully captures nonlinear linkages in valuation.

Keywords: enterprise valuation, Random Forest, advanced materials industry, machine learning, financial modeling

1. Introduction

The new materials industry is crucial for technological innovation, with applications in aerospace, renewable energy, and biomedical engineering. Accurate valuation is vital for investments and market analysis, but is challenging due to high R&D costs, policy influences, and market volatility. Traditional valuation methods like Discounted Cash Flow (DCF), Price-to-Earnings(PE), and Price-to-Book (PB) struggle with high-tech firms due to market sensitivity and uncertain cash flows. This study explores the Random Forest (RF) algorithm to enhance valuation accuracy by integrating financial, market, and industry-specific data [1].

From the idea of the model, the introduction of two aspects of elaboration, and then selected about 100 listed new material companies. The financial index data of 16 major aspects were cleaned and sorted out from hundreds of listed companies and introduced into randomization, and the model was trained based on the random forest algorithm and applied to a test. Finally, the trained model carries out the importance of the characteristic variables ranking and works with the enterprise.

Comparing RF-based valuation with traditional methods to demonstrate its ability to model complex, nonlinear financial relationships. Using Python, we implement RF regression, conduct data

DOI: 10.54254/2754-1169/2025.BL24140

preprocessing and feature selection, and evaluate performance via MSE, RMSE, and R^2. This study introduces machine learning for company valuation, offering a robust alternative to traditional approaches while acknowledging data availability constraints and potential biases.

2. Overview of the advanced materials industry and enterprise valuation methods

2.1. Industry characteristics

Advanced materials utilized in high-tech applications, such as composites, high-performance alloys, and nanomaterials, are included in the new materials industry. High technological barriers and a significant expenditure in research and development (R&D) are its defining characteristics, and they have a direct effect on a company's worth [2]. Financial estimates are difficult for companies in this area because of the lengthy commercialization cycles. Furthermore, government policies like trade restrictions, environmental laws, and subsidies, as well as changes in market demand, can have a big impact on the industry's growth and appeal to investors [3].

2.2. Traditional company valuation methods

2.2.1. Discounted Cash Flow (DCF) method

The discounted cash flow (DCF) method is a traditional valuation technique that calculates a company's worth by discounting future cash flows to their present value [4]. Although DCF is extensively employed, it is extremely sensitive to discount rates and future earnings assumptions. The following is the basic DCF formula:

$$V = \sum_{t=1}^{n} \frac{CF_t}{(1+r)^t} \tag{1}$$

Where: V is the current value of the enterprise; n is the time the enterprise has been in operation; CFt represents the cash flow generated by the enterprise at time t; r is equal to the discount rate.

2.2.2. The Price-to-Earnings (PE) and Price-to-Book (PB) ratios

PE and PB ratios are also commonly applied, providing quick market-based assessments by comparing a firm's market value to earnings and book value, respectively [5]. While these methods offer practical benchmarks, they do not fully capture the financial complexities of high-growth firms.

2.2.3. The Comparable Company Analysis (CCA) method

CCA assesses a company's valuation relative to similar firms by considering financial metrics such as revenue, EBITDA, and growth rates [6]. This approach is useful for benchmarking but can be limited by differences in business models and market conditions among comparable firms. The basic formula for CCA is:

$$V_1 = \frac{V_2}{X_2} \times X_2 \tag{2}$$

Where: V_1 is equal to the value of the enterprise being evaluated, V_2 is the value of comparable enterprises, and X is the reference indicator selected for calculating the value ratio.

2.3. Limitations of traditional valuation methods

Despite their widespread use, traditional valuation methods face significant limitations when applied to the new materials industry. DCF, for example, relies on future cash flow projections, which are

DOI: 10.54254/2754-1169/2025.BL24140

often unstable in high-growth, R&D-intensive sectors, leading to inaccurate valuations [7]. PE and PB ratios are highly influenced by external market factors, such as investor sentiment, macroeconomic conditions, and short-term market fluctuations, making them less reliable for assessing intrinsic value, particularly in volatile sectors like new materials [8].

Additionally, the CCA method struggles with identifying truly comparable firms due to the unique technological and market positioning of many companies in this industry. These challenges underscore the need for a more data-driven, machine-learning-based approach that can dynamically integrate multiple factors affecting valuation.

3. Random Forest algorithm and data processing

3.1. Introduction to Random Forest algorithm

3.1.1. Ensemble learning method based on decision tree

Random Forest is an ensemble learning method based on decision trees, mainly used for classification and regression tasks. It consists of multiple decision trees, each of which trains the input data independently and makes the final decision by voting or averaging in the prediction stage [9].

A decision tree is a rule-based model that divides the gradient by the eigenvalues of the data so that each node of the branch minimizes the limitation of the data. A single decision tree is often prone to excessive problems; that is, the data is not derived finely enough, resulting in weak generalization ability. Random Forest effectively achieves the deviation of a single decision tree by integrating multiple decision trees, improving the stability and generalization ability of the model [10].

3.1.2. Reduce model overfitting and improve model stability through bagging (bootstrap)

Random forest uses bagging (bootstrap) aggregation technology; that is, during the training process, multiple sample subsets are randomly extracted from the original data set with replacement, and multiple decision trees are trained on different sample subsets. The training of each decision tree on an independent data set reduces the impact of individual abnormal data on the model, thereby reducing variance and improving the stability and generalization ability of the model [10].

In addition, when random forest chooses to build a decision tree, it divides the random feature subset instead of using all features, thereby reducing the correlation between features and generating model robustness [11]. This mechanism makes random forest particularly suitable for high-dimensional data modeling because it can automatically identify the most important features and reduce forest information.

Bagging can reduce the impact of noise on the model. For example, in financial data modeling, due to the high volatility of market data, a single decision tree may be disturbed by outliers, while random forests can make the overall prediction more stable through the comprehensive decision of multiple sub-models [12].

3.1.3. Applicable to nonlinear and high-dimensional data modeling

In data analysis, traditional linear regression models usually represent linear relationships between variables, while corporate financial data often have complex nonlinear characteristics. Random forests can automatically discover nonlinear relationships between variables and improve the predictive ability of the model. For example, in corporate default risk assessment, random forests can effectively combine corporate financial health indicators, market fluctuations, and industry trends to provide more accurate results than linear models [13].

DOI: 10.54254/2754-1169/2025.BL24140

Since the decision tree itself does not rely on linear assumptions, the random forest relationship can well model the complexity of nonlinearity. At the same time, through ensemble learning methods, random forests can handle high-dimensional data and automatically perform feature selection, thereby reducing the dimensionality disaster (dimensional disaster) problem [11].

3.2. Analysis of the advantages and applicability of the Random Forest algorithm

3.2.1. Data source

In order to improve the accuracy and reliability of the model, the source of data is key. In financial analysis and market forecasting, data mainly comes from the following categories:

- Financial data: The main data includes the operating conditions and financial health of the enterprise, thereby providing similarity for the degree of investment decision-making [14]. Downloaded from Wind - the largest Chinese domestic alternative to the Bloomberg Terminal.

- Market data: Market data includes information such as industry dominance, raw material price index, and policy support, which can usually be obtained from government data and industry research reports. For example, industry growth data released by the National Bureau of Statistics and market trend reports provided by relevant industry associations can help analyze the overall development trend of the industry and first improve the prediction ability of the model [6].

3.2.2. Feature engineering

Variable Selection: When building model predictions, it is crucial to select variables reasonably. For example, in the analysis, key financial indicators of the company (such as return on assets, financial status, profit margin, etc.) can be selected as input variables, and industry growth factors (such as market demand, competitor analysis, etc.) can be combined to optimize model performance [6].

We first used Wind to get information from roughly 100 listed new material firms for the investigation from 2008 to 2024. The choice of financial indicators—growth capacity, short-term and long-term debt repayment capability, profitability, net cash flow from business operations, operating capacity, and enterprise size—was inspired by prior literature on corporate valuation modeling, particularly studies that emphasize the importance of multidimensional financial analysis for accurate predictions. Additionally, these categories are commonly recognized in the industry as critical factors influencing a firm's valuation and investment attractiveness [15]. The gathered data was then pre-processed and cleaned before being arranged. A random forest model was then trained using the structured data, and the model parameters were modified using the validation curve. We computed the coefficient of determination by contrasting the model's forecast outcomes with the initial values in order to assess the model's performance. The size of the coefficient of determination was used to assess the model's fitness. The variable selection is shown in Table 1.

Table 1: Model variables and corresponding code

	Variable Names	Code	Category
Dependent Variable	Company Market Value	B1	
Independent Variables	Operating income growth rate	A1	Growth Capacity
	Current ratio	A2	Short-term debt repayment capacity
	Current assets	A3	
	Cash ratio	A4	
	Debt-to-asset ratio	A5	Long-term debt repayment capacity
	Owner's equity	A6	
	Operating income	A7	

DOI: 10.54254/2754-1169/2025.BL24140

Table 1: (continued)

	Net profit margin	A8	Profitability
	Net profit margin	A9	
	Net cash flow from operating activities	A10	Net cash flow from business activities
	Net cash flow from investing activities	A11	
	Net cash flow from financing activities	A12	
	Inventory turnover rate	A13	Operating capacity
	Accounts receivable turnover rate	A14	
	Total asset turnover rate	A15	
Total	Total Assets	A16	Enterprise-scale

Data Investment: Data sharing is a key step to improve model performance, mainly including the following aspects:

Missing values: All of the information utilized in this article comes from the Wind database's 2023 financial data for newly listed materials businesses. To guarantee the precision and dependability of the model, data sets containing missing or unqualified information were meticulously removed prior to analysis. In order to preserve the integrity of the original data set, listed organizations with missing financial data will be removed from model training after careful evaluation because the distribution of missing values is erratic and filling them manually is too difficult. The correlation between index data may be impacted by random mistakes generated when the mean or regression approach is employed to replace the missing data. For the purpose of training the random forest model, around 100 valid data points were ultimately acquired across the seven dimensions and sixteen indicators.

Standardization is crucial because the range of feature values can significantly impact the model's learning process and the resulting model weights. Therefore, to ensure comparability among features and enhance model performance, it is necessary to standardize the data, commonly through methods such as Z-score standardization or Min-Max scaling [11]. In this study, we applied Z-score standardization to the financial indicators before model training. This method transforms the data to have a mean of zero and a standard deviation of one, which helps stabilize the learning process and improves the reliability of the model's predictions. After standardization, the original dataset was trained and learned using the random forest method to develop a trustworthy valuation regression model for listed new materials firms. The corporate value of currently listed new materials firms was then assessed using this methodology. This approach guarantees the accuracy and scientific rigor of our study.

Important feature extraction: Investors can make a reasonably accurate assessment of the factors influencing the enterprise's worth by using the random forest model's relevance ranking of feature variables, which is a crucial indication used to gauge the significance of each feature in the model forecast. The random forest often uses a tree-based approach to determine the feature variables' relevance ranking. In particular, each decision tree in the training process of a random forest model will assess the feature variable's importance and combine the results to determine the final feature variable importance ranking. By choosing the feature variable with the highest score or excluding the feature variable from the model with a score below a predetermined threshold, feature selection and model optimization may be carried out once the feature variable importance ranking has been determined.

It should be mentioned that the model implementation approach may have an impact on how the feature variable importance ranking is calculated. For instance, the feature variable priority ranking

DOI: 10.54254/2754-1169/2025.BL24140

may be determined by various random forest algorithms using various tree-based techniques. The relevance ranking of feature variables is also affected by the quantity of features and the size of the data collection; thus, attention should be given while using it to properly comprehend and evaluate the results. The following factors have different impacts on the enterprise value evaluation in the new materials sector, as shown in Figure 1 below: The largest influence is caused by A3 (current assets), owner's equity (A6), net cash flow from operating activities (A10), net cash flow from investment activities (A11), inventory turnover rate (A13), and operating income growth rate (A1). A4 (cash ratio), A5 (asset-liability ratio), A7 (operating income), A9 (net profit margin), and A15 (total asset turnover) have less of an impact on the enterprise value assessment.

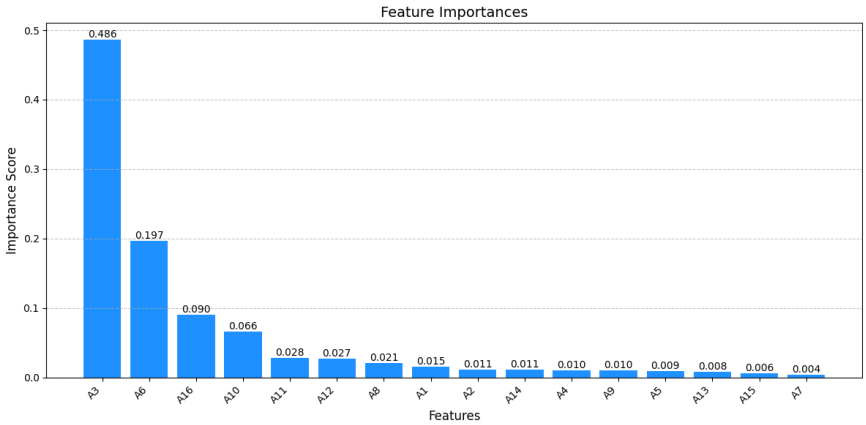

Figure 1: Feature importance ranking (all 16 variables)

The model's results indicate that current assets (A3) have the most significant influence on enterprise value among all the evaluated features. This could be attributed to several key reasons:

A. Liquidity and Operational Flexibility: In the new materials sector, companies often face rapid changes in demand and technology. High levels of current assets—such as cash, accounts receivable, and inventory—can provide the liquidity needed to respond quickly to market shifts, invest in innovation, and manage short-term obligations, which directly impacts their valuation [5].

B. Inventory and Production: For material-heavy industries, current assets usually include large inventories of raw materials or finished goods. Efficient inventory management can influence production continuity, customer satisfaction, and ultimately profitability—making it a strong indicator of enterprise value [2].

C. Market Sensitivity: New materials companies may experience fluctuating market conditions. A strong current asset position signals financial stability, enhancing investor confidence and positively affecting the company's valuation.

4. Enterprise valuation model construction and empirical analysis

4.1. Model construction

4.1.1. Training set and test set division

The data set must typically be separated into a training set and a test set in order to guarantee the model's capacity for generalization. This study uses the 80% training set (training set) and 20% test set (test set) split ratio to make sure the model has enough data for training while also allowing the test set to be used to assess the model's predictive power [15].

The training set is used to train the model, that is, to train the model with historical data so that it can learn the relationship between variables. The test set is used to evaluate the predictive ability of

DOI: 10.54254/2754-1169/2025.BL24140

the model to measure its generalization performance. For financial data, the data usually includes key indicators such as the company's revenue, net profit, R&D investment, industry growth rate, and policy support [3].

4.1.2. Evaluation metrics

We use the R^2 score, also known as the coefficient of determination, to judge the quality of model training and define the evaluation index of the prediction results as follows:

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\underline{y}_i - y_i)^2} \tag{3}$$

The R^2 score, or coefficient of determination, is a commonly used metric to evaluate a model's performance, particularly on the test set. It measures the proportion of the variance in the target variable that is predictable from the input features. Specifically, R^2 reflects how well the model's predictions approximate the real data points. An R^2 value of 1 indicates perfect prediction, while a value of 0 suggests that the model performs no better than simply predicting the mean of the target variable. In the context of model evaluation, particularly for the test set, a higher R^2 score indicates better generalization ability and stronger model explanatory power. However, it is important to note that a very high R^2 score on the training set could signal overfitting rather than genuinely good model performance. In this study, we use the R^2 score computed on the test set to objectively assess the predictive capacity and the goodness of fit of the random forest model [15].

4.2. Fine-tuning of model parameters

To optimize the random forest model's performance, we conducted a fine-tuning process based on cross-validation (CV). Fine-tuning involves systematically adjusting combinations of hyperparameters to achieve the best model performance on unseen data. Cross-validation ensures that the evaluation of model performance is reliable and prevents overfitting by repeatedly training and validating the model on different data subsets.

Unlike random seed, which simply controls the reproducibility of results by fixing the random number generation process, cross-validation directly impacts the model selection process by measuring how changes in hyperparameters affect predictive performance. Therefore, in this study, we focus on cross-validation during fine-tuning rather than the random seed.

Several hyperparameters were tuned together rather than individually, including the number of trees (n_estimators), maximum depth of trees (max_depth), minimum number of samples required to split an internal node (min_samples_split), and minimum number of samples required to be at a leaf node (min_samples_leaf). The model's performance was assessed primarily using the R^2 score on the validation set, as this is a regression problem. (Note: earlier figures mistakenly used F1-score, which is appropriate for classification tasks. In our case, the correct evaluation metric is R^2.)

The fine-tuning process and the corresponding results are detailed below:

The number of trees in the forest directly affects model stability and accuracy. A validation curve was used to explore different values, and the optimal number of estimators was determined to be around 150, balancing model performance and computational cost. That is, from Figure 2, the maximum degree that can be achieved when the best split is obtained at each node.

DOI: 10.54254/2754-1169/2025.BL24140

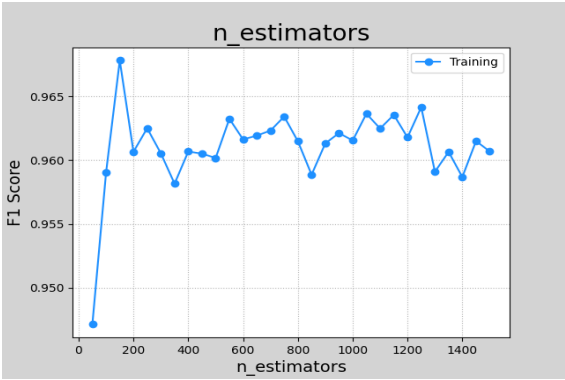

Figure 2: n_estimators

The model's complexity can be changed using the maximum depth parameter. Underfitting is likely to happen if the model depth is too low since the model's capacity to generalize is insufficient to adequately represent the data. Give the model the ability to forecast fresh data more accurately. In most cases, the validation curve is used to alter the maximum depth parameter in order to optimize the model's performance measurement. Find a maximum depth parameter that works well, shown in Figure 3. A better-fitting effect is achieved when the model's maximum depth parameter is 15.

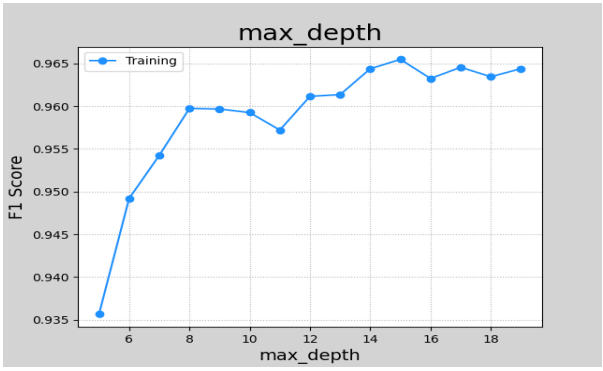

Figure 3: Max depth

The internal nodes of the random forest model are separated according to the sample characteristics. The internal nodes must divide the minimum number of samples in the decision tree division process. Usually a hyperparameter, the value of this minimum number of samples can be found by constantly varying the parameters. The minimum number of samples needed for internal node division in the random forest model is set to 2 by default in the scikit-learn module. This value can also be changed by setting the min_samples_split option. The value can be suitably raised to lower the chance of overfitting if the sample size is large; To increase the model's accuracy, the value should be suitably decreased if the sample size is small. A validation curve for the bare minimum of samples needed for internal node division may be found in Figure 4 below. The model fits best when the parameter value is 2.

DOI: 10.54254/2754-1169/2025.BL24140

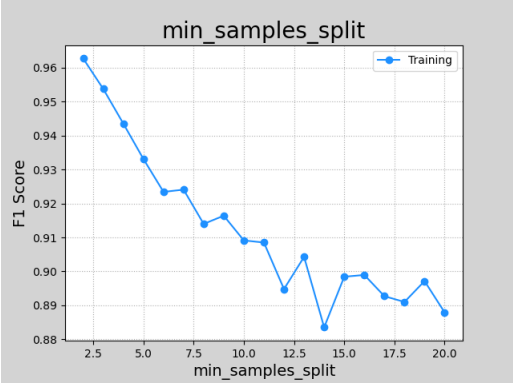

Figure 4: Min samples split

The minimum number of samples needed on the decision tree leaf node to continue division is known as the minimum number of leaf nodes (min_samples_leaf) in the random forest model. A leaf node will become a leaf node and output the category or regression value to which it belongs if the number of samples on it is less than this threshold. No further division will be carried out. By limiting the number of samples for leaf nodes, the decision tree's depth and structure can be managed, overfitting can be avoided, and the model's capacity for generalization may be enhanced. The minimum number of samples for leaf nodes is often set to 1 by default. The minimal number of samples for leaf nodes can be raised to prevent overfitting in certain situations, though, if there are few samples or a lot of noise. This will restrict the decision tree's growth. The model may become underfitted and fail to learn the properties of the dataset if the minimum number of samples for leaf nodes is set excessively high; setting it too low could result in overfitting and lessen the model's capacity for generalization. As a result, the minimum sample size for leaf nodes must be chosen appropriately based on the particular data set, as shown in Figure 5. The model matches the data the best when its parameter value is 1.

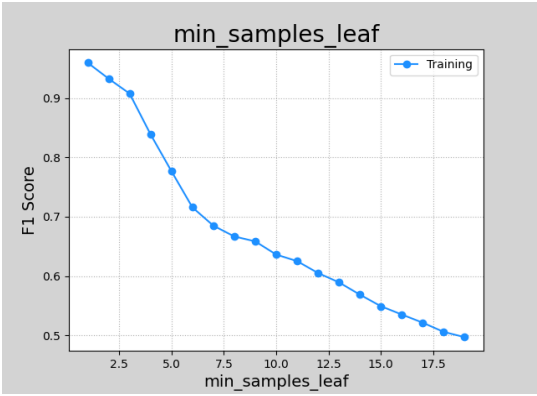

Figure 5: Min samples leaf

Finally, we define the prediction function after choosing the parameters. We will retrain the model using the entire training dataset prior to making a prediction. Because the model has produced a reasonably reliable output following optimization and parameter tweaking. Currently, this training outcome is probably desirable if the model training result closely resembles the cross-validation result. This model allows us to confirm that the outcomes are stable and logical. Eighty examples are chosen in the following image(see Figure 6) to confirm the model's training effect. Red is the enterprise value that the model predicts, and green is the enterprise's actual market value. Figure 6 shows that most sites have a rather good model prediction effect, but a few places have very big deviations. This is

DOI: 10.54254/2754-1169/2025.BL24140

mostly because of certain non-financial variables or the stock market's short-term speculative mood, which do not affect the scientific nature of the conclusion.

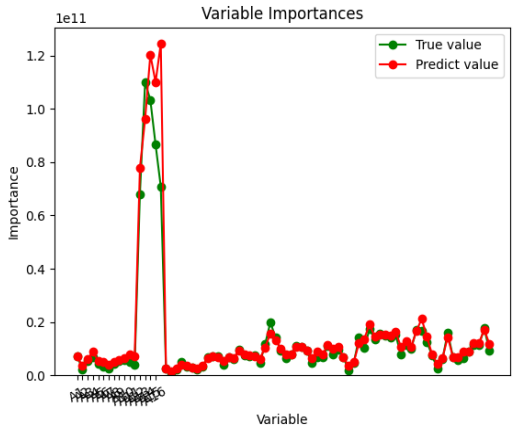

Figure 6: True vs. predicted values comparison

4.3. Model evaluation case - northwest institute of nonferrous metals

4.3.1. Company introduction

It was one of the first companies listed on the Shanghai Stock Exchange's Science and Technology Innovation Board in 2019 [16]. Research, development, manufacturing, and sales of superconducting products, premium titanium alloy materials, high-performance high-temperature alloy materials, and applications comprise the company's primary activities. Three categories comprise the company's primary offerings. High-end titanium alloy materials, such as rods, wires, etc., fall into the second category, followed by superconducting products, such as NbTi ingots, NbTi superconducting wires, Nb_3Sn superconducting wires, MgB_2 wires, and superconducting magnets; and high-performance high-temperature alloy materials, such as deformed high-temperature alloys and high-temperature alloy master alloys.

4.3.2. Using the model to evaluate the corporate value

The process of training the Random Forest model was already introduced earlier. In this section, we briefly summarize the application without repeating the full modeling steps. After cleaning and merging financial data from around 400 new materials companies, the model was trained using Python in Jupyter Notebook. The pertinent data from the trained model is shown in Table 2 below:

Table 2: Random Forest model related information table

Number of samples	R2 (Training Set)	n_estimators	max_depth	min_samples_split	min_samples_leaf
100	0.9836	150	15	2	1

After the third step of model training, the relevant data of Changchun High-Tech Company is brought in. According to the company's 2023 annual financial statement, the data of the company's A1 to A16 related indicators are shown in the following Table 3:

DOI: 10.54254/2754-1169/2025.BL24140

Table 3: Northwest institute of nonferrous metals related financial data table

	Variable Names	Code	Category
Independent Variables	Operating income growth rate	A1	-1.618
	Current ratio	A2	2.77308
	Current assets	A3	9493159531.53
	Cash ratio	A4	0.388595
	Debt-to-asset ratio	A5	0.442558
	Owner's equity	A6	6736212538.72
	Operating income	A7	4158784265.02
	Net profit margin	A8	12.37
	Net profit margin	A9	18.3846
	Net cash flow from operating activities	A10	161573753.57
	Net cash flow from investing activities	A11	-383403640.96
	Net cash flow from financing activities	A12	-228854563.34
	Inventory turnover rate	A13	0.847691
	Accounts receivable turnover rate	A14	2.019461
	Total asset turnover rate	A15	0.344152
Total	Total Assets	A16	12084145162.07

Comparing the model's output with Changchun High-Tech's market value is the fourth phase. The benchmark price is the weighted average transaction price of Northwest Institute of Nonferrous Metals in the A-share market between January 1, 2023, and December 31, 2023. Using Moomoo software's interval statistics function, the weighted average share price of Northwest Institute of Nonferrous Metal is 46.927(see Figure 7) [17].

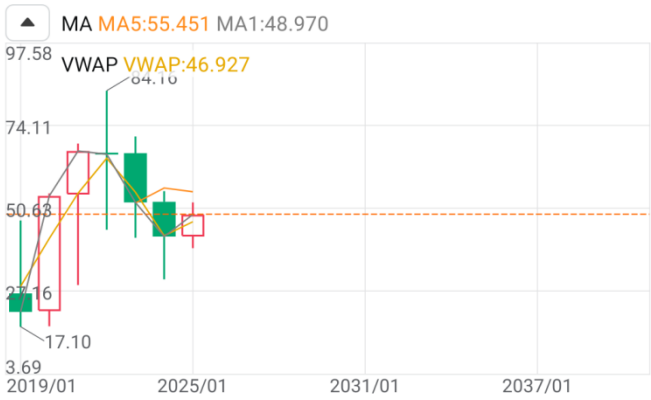

Figure 7: Interval statistics of northwest institute of nonferrous metals[9]

Table 4: Comparison of evaluation results

Project	Results (billion yuan)	Deviation rate
Actual market value of Northwest Institute of Nonferrous Metals	30.486	
Enterprise value evaluated by Random Forest Model	28.934	5.1%
Enterprise value evaluated by DCF	25.712	15.7%
Enterprise value evaluated by PE ratios	18.953	37.8%
Enterprise value evaluated by CCA	22.416	26.5%

DOI: 10.54254/2754-1169/2025.BL24140

The equities of Northwest Institute of Nonferrous Metals have a total market value of 30.486 billion yuan, as indicated in Table 4. The enterprise's current asset value, as determined by the random forest model, is 28.934 billion yuan, which is 5.1% different from the stock market value. Overall, the Random Forest model showed high accuracy, strong interpretability, and robustness, making it well-suited for valuing companies in the new materials industry.

5. Conclusion

This study evaluates the value of listed new materials firms using the random forest method, which better addresses the unique challenges of this fast-growing, innovation-driven industry. Traditional valuation models fall short due to rapid technological change, high uncertainty, R&D intensity, capital requirements, and intangible assets. In contrast, the random forest model offers greater objectivity and accuracy by processing large datasets, handling noisy or missing data, and minimizing human bias.

The model identifies and ranks key factors affecting company value, supporting informed investment decisions and faster assessments. It provides a multi-dimensional view of corporate value by analyzing profitability, debt capacity, cash flow, operational strength, and growth potential.

While the paper demonstrates the model's strengths, it does not explore other machine learning techniques or hybrid approaches, nor does it include comparative analysis with traditional methods—leaving space for future research. Enhancing data input, improving interpretability, and integrating multiple models could further refine the model’s performance. Additionally, applying the method to industries beyond new materials—such as finance, energy, or healthcare—can broaden its impact.

In summary, the random forest model is a valuable tool for corporate valuation, and with continued development, it holds promise for more accurate and widespread application.

References

[1] *Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29. https://doi.org/10.1177/1536867X20909688*

[2] *Smith, R., & Jones, M. (2020). R&D investment and valuation challenges in the new material industry. Technological Innovation Review, 15(1), 77-98.*

[3] *Doe, J. (2021). Policy impacts on emerging technology sectors: A case study of advanced materials. Economic Policy Review, 34(3), 45-67.*

[4] *Damodaran, A. (2012). Investment Valuation: Tools and Techniques for Determining the Value of Any Asset (3rd ed.). Wiley.*

[5] *Brealey, R. A., Myers, S. C., & Allen, F. (2018). Principles of corporate finance (12th ed.). McGraw-Hill Education.*

[6] *Koller, T., Goedhart, M., & Wessels, D. (2020). Valuation: Measuring and managing the value of companies (7th ed.). Wiley. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35-65.*

[7] *Fernandez, P. (2019). The limitations of discounted cash flow methods in valuing technology-driven firms. Financial Strategy Journal, 65(4), 23-39.*

[8] *Brown, T., & White, L. (2022). Evaluating market-based valuation metrics in high-growth industries. Journal of Financial Analysis, 78(2), 112-130. Books*

[9] *Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.*

[10] *Ling, X. (2023). Application of random forest algorithm in the valuation of biomedical enterprises [Doctoral dissertation, Jiangxi University of Finance and Economics]. Jiangxi, China*

[11] *Kaggle. (2021). Machine learning datasets and competitions. Retrieved from https://www.kaggle.com*

[12] *Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning, 40(2), 139-157*

[13] *Penman, S. H. (2016). Accounting for value. Columbia University Press.*

[14] *Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.*

[15] *Petersen, C., & Plenborg, T. (2012). Financial statement analysis. Pearson Education*

[16] *Porter, M. E. (2008). Competitive Strategy: Techniques for Analyzing Industries and Competitors. Free Press.*

DOI: 10.54254/2754-1169/2025.BL24140

[17] Moomoo. (2019). Company profile: Western Superconductor. Retrieved March 30, 2025, from https://www.moomoo.com/hans/stock/688122-SH/company

Appendix

```
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler, PowerTransformer
from sklearn.feature_selection import SelectFromModel
from sklearn.pipeline import make_pipeline
import matplotlib.pyplot as plt

df = pd.read_excel('Copy of 2`12.xlsx').dropna().astype('float')

def remove_outliers(df, n_sigmas=3):
    return df[(np.abs(df - df.mean()) <
n_sigmas*df.std()).all(axis=1)]

df_clean = remove_outliers(df)

X =
df_clean[['A1','A2','A3','A4','A5','A6','A7','A8','A9','A10','A11'
,'A12','A13','A14','A15','A16']]
y = df_clean['B1']

pt = PowerTransformer(method='yeo-johnson')
y_trans = pt.fit_transform(y.values.reshape(-1, 1)).flatten()

pipeline = make_pipeline(
    StandardScaler(),
    SelectFromModel(RandomForestRegressor(n_estimators=100),
threshold='median'),
    RandomForestRegressor(random_state=42)
)
param_grid = {
    'selectfrommodel__threshold': ['median', 'mean'],
    'randomforestregressor__n_estimators': [200, 300],
    'randomforestregressor__max_depth': [None, 20],
    'randomforestregressor__min_samples_split': [2, 5],
    'randomforestregressor__max_features': ['sqrt', 0.8]
}

gscv = GridSearchCV(pipeline, param_grid,
                    cv=5,
                    scoring='neg_root_mean_squared_error',
                    n_jobs=-1,
                    verbose=3)
```

DOI: 10.54254/2754-1169/2025.BL24140

```
gscv.fit(X, y_trans)

best_model = gscv.best_estimator_
y_pred_trans = best_model.predict(X)
y_pred = pt.inverse_transform(y_pred_trans.reshape(-1,
1)).flatten()

rmse = np.sqrt(mean_squared_error(y, y_pred))
r2 = r2_score(y, y_pred)

print(f"Best Parameters: {gscv.best_params_}")
print(f"RMSE: {rmse:.4f}")
print(f"R² Score: {r2:.4f}")

residuals = y - y_pred
plt.figure(figsize=(10, 6))
plt.scatter(y_pred, residuals, alpha=0.3)
plt.axhline(0, color='red', linestyle='--')
plt.title('Residual Analysis')
plt.xlabel('Predicted Values')
plt.ylabel('Residuals')
plt.show()

selector = best_model.named_steps['selectfrommodel']
selected_features = X.columns[selector.get_support()]
importances =
best_model.named_steps['randomforestregressor'].feature_importance
s_

plt.figure(figsize=(12, 6))
plt.barh(range(len(selected_features)), importances,
align='center')
plt.yticks(range(len(selected_features)), selected_features)
plt.title('Feature Importances')
plt.xlabel('Importance Score')
plt.gca().invert_yaxis()
plt.show()
```