Profitability Analysis of Listed Companies in the Era of Big Data: Based on the Decision Tree Model

Zhiming Wu[1], Yong Xiong[2,*]

[1]Department of Accounting, AnHui Business and Technology College, Hefei, China
[2]Department of Accounting, Guangzhou College of Technology and Business, Guangzhou, China

*Corresponding author's email: xiongyong@gzgs.edu.cn

Abstract. Profitability is a critical indicator highly valued by potential investors. However, this indicator is often constrained by human bounded rationality and management's manipulative incentives, which compromises the relevance and reliability of financial reporting. This paper selects financial data of Chinese listed companies from the CSMAR database covering the period from 2001 to 2022 as the sample and employs machine learning algorithms to predict their future profitability. The empirical results indicate that machine learning can effectively improve the accuracy of financial forecasting. Specifically, by utilizing the decision tree model, this study provides investors with theoretical support and practical modeling tools for analyzing corporate profitability. Compared with traditional financial forecasting, machine learning algorithms can automatically learn patterns from historical data and capture more complex influencing factors. By analyzing massive multi-source data—including financial and non-financial data, industry information, and real-time market changes—these algorithms facilitate a more comprehensive and dynamic prediction of future financial performance.

Keywords: Machine Learning, Decision Tree Model, Profitability

1 Introduction

Profitability prediction essentially falls within the scope of pattern recognition in complex systems. Currently, forecasting methods commonly used in practice include the weighted average method, moving weighted average method, and linear regression. While these methods are logically simple, they often rely on strict linear assumptions and struggle to capture the non-linear characteristics inherent in financial data. With the rise of big data technology, machine learning methods are being increasingly applied to practical scenarios. Accordingly, this study selects the Decision Tree Model to investigate the profitability prediction of Chinese listed companies.

The Decision Tree Model is capable of automatically learning complex non-linear patterns from historical data. It imposes no strict assumptions on data distribution and possesses strong interpretability, intuitively demonstrating the key decision paths that

P. S. Borah et al. (eds.), *Proceedings of the 2026 6th International Conference on Enterprise Management and Economic Development (ICEMED 2026)*, Advances in Economics, Business and Management Research 394,
https://doi.org/10.2991/978-94-6239-719-4_16

influence profitability. Given that this paper utilizes data from the CSMAR database covering a long period from 2001 to 2022 with a substantial sample size, exhibiting distinct big data attributes, the Decision Tree Model is particularly suitable for analysing the profitability of listed companies characterized by such high dimensionality and large scale. This paper applies this model to financial forecasting across all industries, aiming to leverage its advantage in mining deep data patterns to provide a more transparent analytical tool for investment decision-making in the big data environment.

1.1 Literature Review

Early research on corporate profitability prediction primarily focused on statistical analysis methods based on financial ratios. Classic Logit regression models laid the foundation for financial forecasting. However, traditional methods are mostly based on linear relationship assumptions and impose strict requirements on statistical data distribution (such as normal distribution). Consequently, it is difficult for them to effectively capture the non-linear characteristics and complex interactions prevalent in financial data within the big data environment (Balcaen & Ooghe, 2006)[1].

With the development of big data and artificial intelligence technologies, academia has begun to widely adopt machine learning algorithms to overcome the limitations of traditional statistical models. Extensive empirical research indicates that machine learning is significantly superior to traditional statistical methods in processing high-dimensional data and non-linear patterns. For instance, Olson et al. (2012) compared the performance of decision trees, artificial neural networks (ANN), and support vector machines (SVM) in financial prediction, and the results demonstrated that machine learning algorithms possess significant advantages in prediction accuracy[2].

Among numerous machine learning algorithms, the Decision Tree Model has garnered significant attention due to its unique advantages. Although neural networks offer high prediction accuracy, their "black-box" nature results in a lack of interpretability, making it difficult for investors to understand the specific decision logic. In contrast, the Decision Tree Model can not only handle non-linear relationships but also possesses strong interpretability. When constructing business failure prediction models, Gepp et al. (2010) pointed out that decision tree classifiers can intuitively demonstrate the critical paths affecting corporate survival while maintaining high prediction accuracy, which has important guiding significance for investment decisions[3]. Furthermore, regarding the characteristics of the Chinese capital market, some scholars have noted that decision trees can effectively handle missing data and outliers, making them better adapted to the data environment of emerging markets where information disclosure may be imperfect (Li et al., 2016)[4]. Therefore, in the era of big data, utilizing the Decision Tree Model to mine profitability patterns from the massive financial and non-financial data of listed companies has become an important approach to enhancing the relevance and reliability of predictions.

1.2 Contributions of this Paper

The main contributions of this paper lie in: on the one hand, exploring the construction

of profitability prediction models in the context of big data, which not only enriches the theoretical research in related fields but also provides a powerful tool for stakeholders such as analysts, investors, shareholders, and creditors to accurately predict a company's profitability, thereby promoting the high-quality development of the financing market. On the other hand, introducing decision tree models from machine learning into the profitability prediction of Chinese listed companies has effectively improved the prediction accuracy while also expanding the application scenarios of machine learning algorithms in the financial field.

2 Methodology

2.1 Overview of Decision Tree Models

A Decision Tree is a non-parametric supervised learning method based on tree structures, which can be used for both classification tasks and regression analysis. In this study on the prediction of profitability of listed companies, Decision Tree Regression is primarily adopted to predict continuous financial indicators (such as Return on Assets, ROA)[5]. Unlike traditional Linear Regression models that rely on linear assumptions between features, decision trees use a "Divide and Conquer" strategy to recursively partition the feature space, thereby generating a series of "If-Then" decision rules. This non-parametric characteristic gives it significant advantages in handling the non-linear relationships, interactions, and high-dimensional complex data commonly found in financial data. The construction process of a decision tree is essentially a recursive partitioning process. The algorithm starts from the root node, selects the optimal feature and split point, and partitions the dataset into two or more subsets, such that the purity of the target variable in each subset is higher (for classification) or the variance is smaller (for regression). This process is repeated until a predefined stopping condition is met (such as the tree reaching its maximum depth, or the number of samples in a leaf node being below a threshold). Finally, each leaf node represents a prediction value, where in regression tasks, this value is typically the mean of the target variable of all training samples within the node.

2.2 Node Partition Criteria

The core of decision trees lies in how to choose the "optimal partition attribute." Different decision tree algorithms adopt different partition criteria, mainly including mean squared error for regression and information gain rate and Gini coefficient for classification.

Mean Squared Error (MSE). For regression tasks when constructing a regression tree for profitability prediction, the commonly used partition criterion is minimizing mean squared error. MSE measures the sum of squared deviations between data points and their predicted values (usually the mean). The model attempts to find a split point that

minimizes the sum of MSEs of the left and right subsets after partitioning, thereby making the data within the subsets as concentrated as possible.

Gain Ratio (Information Gain Ratio). When dealing with discrete financial features, the algorithm introduces Gain Ratio to select the splitting attribute. Traditional Information Gain tends to choose attributes with more values, which may lead to overfitting. To overcome this flaw, Gain Ratio corrects this by introducing "Intrinsic Value" (IV) as a penalty term. The calculation formula is as follows:

$$\text{Gain\_ratio}(D, a) = \frac{\text{Gain}(D, a)}{\text{IV}(a)}$$

Among them, the intrinsic value IV(a) of attribute a is defined as:

$$\text{IV}(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$$

The formula indicates that the more possible values attribute a can take, the higher the value of IV(a) is usually, thus penalizing information gain in the denominator. When making specific selections, the algorithm typically employs heuristic strategies: first identifying attributes with information gain above the average from the candidate attributes, and then selecting the one with the highest gain rate as the splitting attribute.

Gini Coefficient. The CART (Classification and Regression Tree) algorithm uses the Gini coefficient to measure the purity of a dataset when dealing with classification problems. The Gini coefficient reflects the probability that two randomly selected samples from the dataset D have inconsistent class labels. The smaller Gini(D), the higher the purity of the dataset. Its calculation formula is:

$$\text{Gini}(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k \, p_{k'} = 1 - \sum_{k=1}^{|y|} p_k^2$$

When selecting attribute a for partitioning, choose the attribute that results in the minimum Gini index after partitioning. The Gini index of attribute a is calculated as follows:

$$\text{Gin\_index}(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} \text{Gini}(D^v)$$

2.3 Analysis of the Advantages and Disadvantages of the Model

Based on the above principles, the decision tree model exhibits the following characteristics in financial data analysis:

Main Advantages of the Decision tree Model.

- Decision trees are non-parametric models, which do not need to assume that the data obey normal or linear distributions, and can naturally simulate the nonlinear relationship between financial indicators.
- Compared with the "black box" characteristics of neural networks, the tree structure generated by decision trees is intuitive and clear, making it easy to show investors the logical path of profit forecasting.
- Equations should fit into a two-column print format and be single spaced.
- It can effectively handle missing and outlier values, relying only on valid feature data on that node when segmenting.

Main Limitations of the Decision tree Model.

- If left unrestricted, decision trees tend to grow too lushly, capturing noise in the training data rather than universal patterns. This is usually solved by "pre-pruning" or "post-pruning" to control the depth of the tree.
- Decision trees are more sensitive to small changes in data. A small amount of variation in the training set can lead to a completely different tree structure, which is often improved by ensemble learning (such as random forests).
- The decision tree adopts the Greedy Approach, selecting the local optimal solution at each step, which cannot guarantee the global optimum.

3 Research Design

3.1 Sample Selection

This experimental data originates from the CSMAR database, using the financial dataset of Chinese listed companies from 2001 to 2022 as the research sample. The original data includes a total of 70,845 valid records, covering 38 types of financial indicators such as net profit margin on total assets, total asset turnover, and debt-to-asset ratio. The distribution of variable data is shown in Table 1.

Table 1. Descriptive statistics of the sample

Variable Name	Sanple Count	Mean	Standard Deviation	Mini Value	25th Percentile	Median	75th Percentile	Max Value
Return on Assets	70845	0.04983	0.044634	-0.084413	0.020955	0.04625	0.073363	0.185831
Total Asset Turnover	70845	0.577653	0.315163	-0.12829	0.348522	0.54624	0.750293	1.565852

3.2 Predicted Variable

Corporate profitability refers to a company's ability to achieve profits within a certain

period, which is an important reflection of its operational and management level. The quality of profitability directly affects the survival and development of the enterprise. Predicting a company's future profitability is of great significance, which is crucial for stakeholders such as investors, management, suppliers, and customers. The study adopts the return on assets (ROA) as the predicted variable. It is calculated as net profit divided by average total assets, reflecting the efficiency of a company in using its assets to generate profits. The higher this indicator, the stronger the profitability of the company's assets.

3.3 Predicted Variables

Based on existing research on corporate profitability and the availability of data, the study selected 38 variables as input values (predicted variables) for the model, mainly from four dimensions: profitability, solvency, operational efficiency, and growth potential.

3.4 Code Implementation

This experiment uses Python code, and the core code of the experiment is shown in Figure 1:

```
import pandas as pd
import numpy as np
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import r2_score, mean_absolute_error

# Core model training and evaluation
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)
grid_search = GridSearchCV(DecisionTreeRegressor(random_state=42),
                           {'max_depth': [2,3,5,7], 'min_samples_leaf': [1,2,4]},
                           cv=5)
grid_search.fit(X_train, y_train)
print(f"Test R²: {r2_score(y_test, grid_search.predict(X_test)):.4f},
MAE: {mean_absolute_error(y_test, grid_search.predict(X_test)):.6f}")
```

Fig. 1. The core procedure of decision tree

3.5 Empirical Results

The training and parameter tuning process of the decision tree regression model for predicting the profitability of listed companies is implemented based on grid search (GridSearchCV), traversing 3 sets of max_depth, 2 sets of min_samples_split, and 2 sets of min_samples_leaf parameter combinations. This parameter combination effectively avoids overfitting while ensuring the model's fitting effect, balancing the interpretability and generalization ability of the model. The decision tree structure for predicting corporation ROA is shown in Figure 2.

In addition, the decision tree model also presents the top ten most important indicators affecting profitability prediction, see Figure 3. These indicators provide a reference

for profitability prediction and also serve as a basis for enterprises to improve their operational conditions.

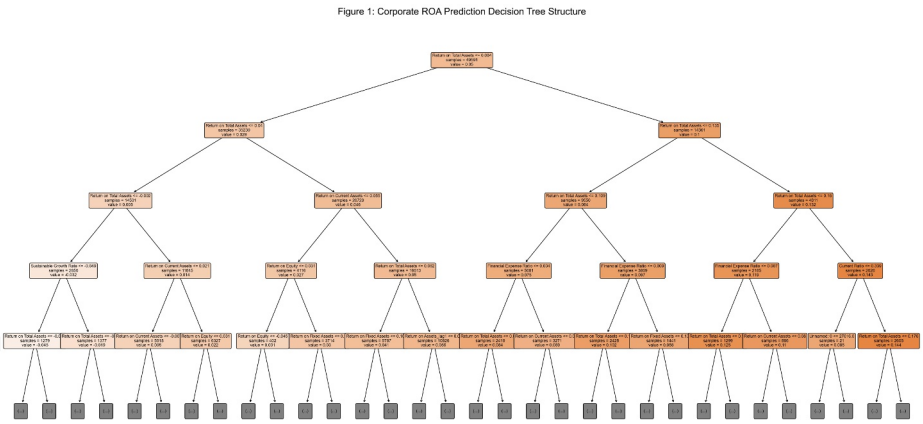

Fig. 2. Corporation ROA Prediction Decision Tree Structure

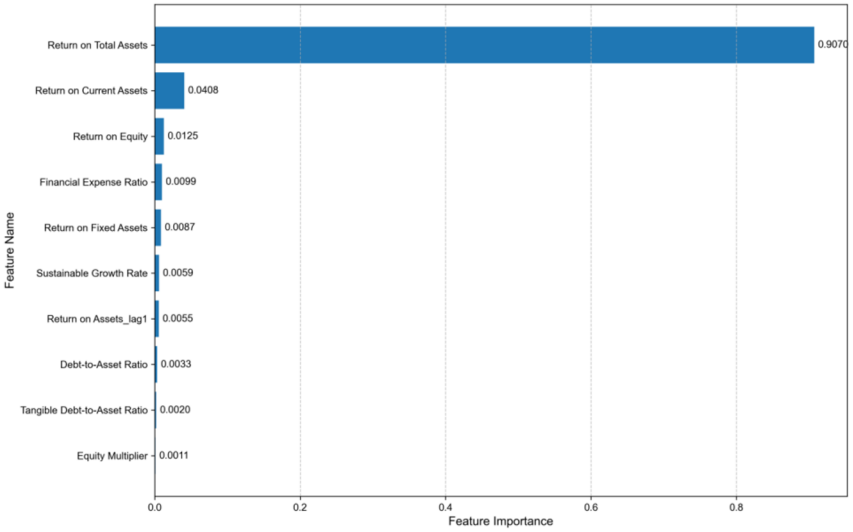

Fig. 3. Decision Tree Model Feature Importance Ranking (Top 10)

A decision tree regression model built using Python ecosystem libraries achieved high-precision prediction of ROA on over 70,000 financial data points of listed companies. The model's test set R^2 reached 0.8652, with stable error metrics across various indicators, confirming the effectiveness and practicality of the decision tree algorithm in financial indicator prediction scenarios. This provides reliable technical support for financial risk assessment of listed companies. Specific data fitting results are referenced in Figure 4.

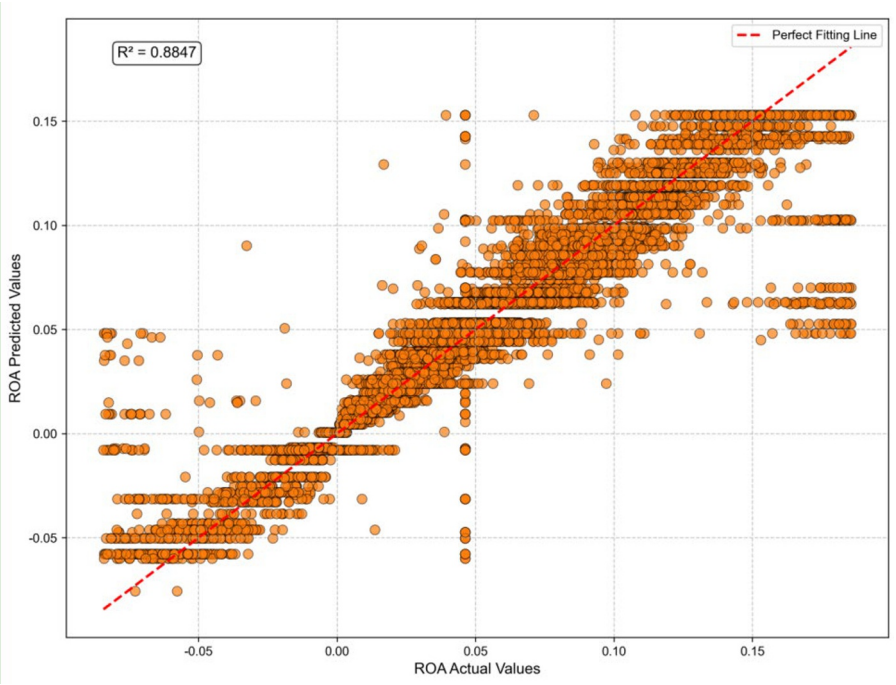

Fig. 4. Test Set ROA Actual vs Predicted Values Scatter Plot

4 Review of Corporate Financial Condition Forecasting

4.1 Traditional Corporate Financial Condition Forecasting Methods

Traditional corporate financial condition forecasting methods primarily rely on the analysis of a company's historical financial data, using past financial trends to predict future financial performance. These methods include financial ratio analysis, trend analysis, regression analysis, etc. Financial ratio analysis calculates and compares financial indicators such as profitability, solvency, and operational efficiency to help businesses assess changes in their financial condition. Trend analysis observes the historical changes in indicators like revenue, costs, and profits to predict future financial trends. Regression analysis establishes models relating financial indicators to other influencing factors to forecast future financial outcomes.

Taking Vanke, a real estate company listed in China with the stock code 00002, as an example, the company's ROA data over the past 10 years are shown in Table 2.

Table 2. Vanke Financial Information

YEAR	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
ROA	4.13%	3.82%	3.79%	4.25%	3.41%	3.19%	3.22%	3.19%	3.17%	1.96%

This article will adopt the Moving Average, 5-Year Moving Average, and Linear Regression Method, and the results of the profitability forecast are shown in Table 3.

Table 3. Traditional forecasting methods

Traditional Predictive Methods	Return on Assets (True)	Return on Assets (Pred)	Prediction Error
Simple Average Method	0.021371	0.0445	0.023129
5-Year Moving Average	0.021371	0.0301	0.008729
Linear Regression Method	0.021371	0.0285	0.007129

The characteristics of traditional financial forecasting methods are relatively simple calculations, easy to understand and implement. They primarily rely on historical data and experience assumptions, assuming that past trends will continue in the future. Therefore, they are widely used in financial management, especially in relatively stable market and economic environments. However, traditional methods are typically based on linear relationships, making it difficult to capture complex nonlinear changes and interactions among multiple factors. Their predictive accuracy depends on the stability of the data and the continuity of past experience, and their effectiveness may be limited when facing rapidly changing markets or extreme situations. Additionally, traditional corporate financial condition forecasting methods rely less on external information, usually only using internal financial data of the enterprise, ignoring the impact of industry changes, macroeconomic fluctuations, and market competition. The limitations of these methods become more apparent in dynamic and uncertain business environments, where managers may be unable to respond promptly to sudden market changes or financial crises. This makes traditional methods more suitable for enterprises or market environments with stable historical data and controllable future changes.

4.2 Intelligent Enterprise Financial Condition Forecasting Methods

Methods for predicting the financial condition of digital-intelligent enterprises are modern financial forecasting tools based on digital and intelligent technologies. Through means such as big data, artificial intelligence, machine learning, and automation technology, they help businesses predict their future financial performance more comprehensively and dynamically. These methods not only rely on traditional financial data but also combine a large amount of external data, such as market trends, competitor dynamics, and the macroeconomic environment, enabling them to capture more complex influencing factors.

The greatest feature of digital-intelligent methods lies in their ability to process large amounts of data. Unlike traditional methods, digital-intelligent methods can analyze massive amounts of data from multiple sources, including financial data, non-financial data, industry information, and real-time market changes. This makes the predictions more comprehensive and able to identify hidden trends and anomalies,

Taking Vanke Company as an example, the ROA predicted using the decision tree model was 2.52%, with an error of 0.00384 compared to the actual ROA, as shown in Table 4. This indicates that the use of intelligent technologies such as machine learning offers higher accuracy and stronger foresight than traditional methods.

Table 4. Profitability Prediction Based on Decision Tree Model

Stock Code	Reporting Year	Stock abbreviation	Return on Assets (True)	Return on Assets (Pred)	Prediction Error
000002	2022	Vanke Company	0.021371	0.024071	0.0027

Through machine learning and deep learning models, digital-intelligent financial prediction systems can automatically learn patterns from historical data and adjust when encountering new data. This dynamic adjustment capability makes predictions more flexible, especially when facing uncertainty or market changes. Digital financial prediction systems can update predictions in real-time, providing more timely decision support. Additionally, digital financial prediction can achieve real-time data collection, processing, and analysis through automated systems, allowing managers to obtain the latest prediction results at any time without waiting for periodic financial reports or data updates. This enables businesses to identify potential financial risks more quickly, optimize resource allocation, and enhance decision-making efficiency.

5 Conclusion

The study found that while machine learning and other algorithmic models have good judgment in predicting future trends, their explanatory power for results is still lacking, and they lack strong persuasiveness. Therefore, to enhance the persuasiveness of the results, it is necessary to incorporate micro-level considerations of real-world business scenarios. This means that when providing consulting and predictive services to businesses, it is essential to combine big data macro-level predictions with micro-level business scenarios. Machine learning algorithm models can be used to enhance traditional financial analysis. Traditional business and finance analysis data can reinforce the conclusions of machine learning algorithms, creating a complementary relationship between the two. Additionally, this study primarily uses decision tree models, which may lead to overfitting of training data, resulting in poor generalization ability, reduced model stability, and susceptibility to noise data. Future research will focus on adopting more advanced algorithm models to further improve the accuracy and stability of predictive results.

Acknowledgements

This work was funded by the research project, "Research on the Cultivation of New Quality Productivity of Strategic Emerging Industries in Anhui Province from the Perspective of Resource Allocation Efficiency" (2025AHGXSK30388); "Anhui Province

University Outstanding Talents Support Program Key Project" (JNFX2025178); Anhui Vocational and Adult Education Association Education Research Planning Project "Research on Inquiry-based Learning Mode of Artificial Intelligence in Higher Vocational Business Education" (AZCJ2025123); Anhui Provincial Humanities and Social Sciences Research Project "Research on How Fiscal and Taxation Policies Drive the High-quality Development of New Quality Productive Forces (SK2024A011); Guangzhou Business School Federation of Social Sciences Humanities and Social Sciences "Data Element-driven Financial Sharing Transformation and Business Environment Optimization of Small and Medium-sized Enterprises - Guangzhou Sample" (SKKYYB202515); Anhui Province University Humanities and Social Sciences Key Research Project "Research on the Digital Empowerment of Anhui New Energy Vehicle Enterprises to Enhance Green International Competitiveness under the Background of "Double Carbon" (2025AHGXSK30221); Anhui Province University Humanities and Social Sciences Key Research Project "Research on Talent Gathering and Innovation Capacity Improvement in Anhui Province in the Process of Yangtze River Delta Integration" (2025AHGXSK30160); Anhui Province University Humanities and Social Sciences Key Research Project "Research on Efficiency Audit and Governance Optimization of Big Data Empowering the "Four Good Rural Roads" Project from the Perspective of Public Value" (2025AHGXSK30454);AnHui Business and Technology College Teaching and Research Project "Smart Inquiry Learning: Research on a New Model of Conversational Learning for Taxation Based on Deepseek" (2024xjjy10); China Institute of Business Accounting Research Project: Research on a New Model of Conversational Learning for Accounting Based on Generative Artificial Intelligence (2025ZJ001);Anhui Provincial Humanities and Social Sciences Research Project “A Study on the Talent Agglomeration Mechanism of Anhui Province's Automotive Industry - An Empirical Analysis Based on Regional Development and the Automotive Industry (SK2024A007)”.

References

1. Balcaen, S., Ooghe, H.: '35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems', The British Accounting Review, 2006, 38, (1), pp. 63–93.
2. Olson, D.L., Delen, D., Meng, Y.: 'Comparative analysis of data mining methods for bankruptcy prediction', Decision Support Systems, 2012, 52, (2), pp. 464–473.
3. Gepp, A., Kumar, K., Bhattacharya, S.: 'Business failure prediction using decision trees', Journal of Forecasting, 2010, 29, (6), pp. 536–555.
4. Li, Z., Crook, J., Andreeva, G., Tang, Y.: 'Predicting the risk of financial distress using corporate governance measures', Pacific-Basin Finance Journal, 2016, 51, pp. 1–32.
5. Du, X., & Tan, Y. (2024). Forecasting ROA using machine learning: Evidence from listed firms. Emerging Markets Finance & Trade, 60(3), 786-804.