Frontiers in Economics and Management Volume 7 Issue 9, 2026
ISSN: 2692-7608 DOI: 10.6981/FEM.202609_7(9).0010

Research on Financial Data Mining and Investment Decision Optimization under the Background of Digital Transformation

Leyi Shi[a]

Lingnan College, Sun Yat-sen University, Guangzhou, Guangdong, China

[a]sly28829911@126.com

Abstract

Under the background of digital transformation, traditional investment decision-making methods are facing challenges due to insufficient information processing ability and time lag. Based on the ecological transformation of financial data, this study constructs an analytical framework integrating data collection, data mining and investment decision optimization. By comparing the performance of three algorithms, XGBoost, Long Short Term Memory (LSTM) and random forest, it is found that XGBoost is significantly superior to other models in terms of mean absolute error (MAE), root mean square error (RMSE) and other indicators, and its gradient lifting framework can effectively capture the nonlinear relationship between high-dimensional features. Based on the signal generated by XGBoost, a two-stage investment strategy is constructed: in the first stage, the income forecast signal is generated by comprehensive feature engineering, and in the second stage, the asset weight is allocated by means of mean-variance optimization model. The empirical results show that this strategy achieves an annualized rate of return of 20.56% and a Sharp ratio of 0.81% in the out-of-sample interval, and the maximum retracement is controlled to -20.36%, which is superior to the benchmark index and the single model strategy. The research verifies the effectiveness of multi-source data mining technology in improving the scientific nature of investment decision-making and risk adjustment income, and provides an intelligent decision-making tool for financial institutions and individual investors.

Keywords

Digital Transformation; Financial Data Mining; Investment Decision Optimization; Machine Learning.

1. Introduction

Traditional investment decision-making methods, whether it is fundamental analysis with value discovery as the core or technical analysis based on price trend, mainly rely on limited and structured data sources. When faced with massive, high-frequency and multi-dimensional financial data, they often expose problems such as insufficient information processing ability, lagging timeliness and limited ability to capture nonlinear relationships [1]. Especially at the moment when market volatility is intensified and the speed of information dissemination is accelerated, the demand of investors for more efficient and accurate decision support tools is increasingly urgent.

The rapid progress of data mining and machine learning technology provides strong technical support for extracting valuable investment signals from massive financial data. Random forest, Long Short Term Memory (LSTM), XGBoost, and other algorithms have demonstrated significant advantages in scenarios such as stock price prediction, risk assessment, and credit scoring [2-3]. In addition, the application of natural language processing (NLP) technology in emotional analysis of financial texts and event-driven investment strategies is becoming more

ISSN: 2692-7608 DOI: 10.6981/FEM.202609_7(9).0010

and more mature [4]. These technical means make it possible to automate and intelligentize "from data to insight, from insight to decision-making". However, most of the existing research focuses on the application of single technology or the empirical analysis of single market, and few literatures systematically integrate data collection, data mining and investment decision optimization into an integrated analysis framework from the macro background of "digital transformation". This research gap just constitutes the core starting point of this paper.

This paper combs the influence mechanism of digital transformation on financial data ecology, compares the performance of various data mining methods in financial forecasting tasks, designs an investment strategy based on data mining signals, and tests its improvement effect on risk-adjusted income through back-testing. The theoretical significance of this paper lies in the interdisciplinary integration of financial technology, data science and investment decision theory, which expands the research perspective of data-driven investment decision. The practical significance lies in providing a set of reference methodology for individual investors, quantitative funds and financial institutions to help them improve the scientific and efficiency of investment decision-making in the digital age.

2. Definition of Core Concepts

2.1. Digital Transformation

Digital transformation refers to the process in which enterprises or industries use digital technology to fundamentally change their business processes, business models, organizational structures and value creation methods [5]. In the financial field, digital transformation is not a simple technological upgrade, but a systematic change covering the following three levels:

1) Data layer. The source of financial data has expanded from traditional structured data to massive unstructured and semi-structured data, forming a new data ecology of "multi-source, heterogeneous and high frequency";

2) technical layer. Cloud computing provides elastic computing support, distributed storage and real-time computing framework make large-scale data processing possible, and artificial intelligence and machine learning algorithms give data the ability of "intelligent analysis";

3) Application layer. Digital transformation has spawned new financial formats such as intelligent investment, algorithmic trading, digital risk control and blockchain clearing, which fundamentally changed the way of information acquisition, analysis and decision-making.

The "digital transformation" referred to in this paper focuses on its transformation effect on the supply pattern and analysis paradigm of financial data, rather than discussing the digitalization of organizational management within financial institutions.

2.2. Financial Data Mining

Financial data mining refers to the process of automatically discovering hidden and potentially valuable knowledge, laws and patterns from large-scale, high-dimensional and noisy financial data by using statistics, machine learning and pattern recognition [6]. Its core goal is to transform raw data into structured information that can be used for analysis and decision-making. In the research context of this paper, financial data mining covers data acquisition and fusion, feature engineering and selection, and pattern discovery and prediction. In particular, this paper regards "financial data mining" as a key technical bridge connecting "data resources" and "investment decision-making". It is not the end point, but the core link that provides input signals for subsequent decision-making.

2.3. Investment Decision Optimization

Investment decision optimization refers to the process of determining the optimal or near-optimal asset allocation and trading strategy through mathematical modeling and algorithm

ISSN: 2692-7608 DOI: 10.6981/FEM.202609_7(9).0010

solution under given risk constraints and objective functions, so as to maximize expected returns, minimize risks or optimize risk-adjusted returns [7].

Traditional investment decision-making optimization mainly relies on the mean-variance model and its variants, and its assumptions include investor rationality, normal distribution of returns, complete information and so on. However, in the digital age, financial markets are characterized by stronger nonlinearity, high volatility and information asymmetry, and the applicability of traditional models is challenged.

The definition of "investment decision optimization" in this paper has two meanings: in a broad sense, it refers to the process of improving the quality of investment decision by using any available information and methods, including asset selection, position management, timing judgment and risk control; In a narrow sense, it refers to constructing or improving investment strategies based on the characteristics and signals extracted from data mining, and verifying its optimization effect on risk-adjusted returns through back testing.

3. Research Methods and Data

3.1. Data Source

Multi-source data acquisition strategy is adopted to capture the multi-dimensional characteristics of financial markets. The main data sources are as follows:

(1) Stock market transaction data. Daily OHLCV (open, high, low, close, volume) data for constituent stocks of a major benchmark index are obtained from the CSMAR database. The CSMAR database is widely used in empirical financial research and provides high-quality, fine-grained financial data.

(2) Fundamental data of the company. Extract quarterly and annual financial statements from CSMAR, including balance sheet, income statement and cash flow statement, to construct fundamental indicators such as profitability, solvency and growth.

(3) Macroeconomic indicators. Key macroeconomic variables, including interest rate, inflation rate and GDP growth rate, are obtained from the National Bureau of Statistics and the People's Bank of China to control the macro-level impact.

The sample interval is from January 2015 to December 2025, covering about 11 years of market data. This interval contains multiple market cycles, which can ensure the robustness of empirical analysis. Eliminate the stocks with incomplete transaction records or delisting in the sample interval to reduce the influence of survivor bias.

3.2. Data Preprocessing

The original financial data usually have the problems of missing value, abnormal value and dimensional heterogeneity. The following pretreatment processes are implemented:

(1) Missing value filling. For the time series data with isolated missing values, the forward filling method is adopted to maintain the time continuity of the series. For the case of continuous missing for more than 5 trading days, cubic spline interpolation method is adopted, which approximates the missing value by fitting smooth cubic polynomial to the existing data points, while maintaining the overall trend.

(2) Abnormal value detection and processing. Identify extreme observations that exceed the standard deviation of the mean by three times, and truncate them to the 1st and 99th percentiles to reduce the impact of data errors and extreme market events.

(3) Normalization. Minimax normalization is used to scale all feature variables to a unified range [0,1]:

ISSN: 2692-7608 DOI: 10.6981/FEM.202609_7(9).0010

$$x_i^{'} = \frac{x_i - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})} \tag{1}$$

Where x_i represents the original value of the feature i, $\min(\mathbf{x}), \max(\mathbf{x})$ is the minimum and maximum value of the feature on the training set, and $x_i^{'}$ is the normalized value. This transformation ensures that the features of different units and dimensions make equal contributions to the model training process.

(4) Division of training-verification-test set. In order to prevent forward-looking bias, the data are divided in time sequence, 60% of which are used for training (January 2015–December 2020), 20% for verification (January 2021–December 2022) and 20% for out-of-sample testing (January 2023–December 2025).

3.3. Selection of Data Mining Methods

Based on the different theoretical basis of each method and its empirical performance in financial forecasting, XGBoost, random forest and LSTM are selected for comparative evaluation.

XGBoost is an integrated learning method. By building decision trees in sequence, each new tree corrects the mistakes of the previous tree through gradient-based optimization. It introduces regularization term to control model complexity and prevent over-fitting. Previous studies have proved that XGBoost has consistently achieved high forecasting accuracy in the stock return forecasting task [8]. The minimum objective function in the training process is:

$$L(\phi) = \sum_i l(y_i - \hat{y}_i) + \sum_i \Omega(f_k) \tag{2}$$

Among them, $l(y_i - \hat{y}_i)$is a differentiable convex loss function, which measures the difference between the actual income y_i of the i observation and the predicted income $\hat{y}_i$; $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$ is the regularization term of tree f_k, T is the number of leaf nodes, ω is the leaf weight, and γ and λ are the penalty coefficients.

Random forest is a bagging integration method, which uses self-help samples and random feature subsets to independently construct multiple decision trees, and the final prediction is obtained by averaging the outputs of all trees. Although random forest is robust and interpretable, previous studies show that its performance may be worse than that of lifting method in financial scenarios because it treats data patterns equally without distinction [9].

LSTM is a recurrent neural network (RNN) architecture specifically designed to capture long-term dependencies in sequential data. Due to its ability to model nonlinear time series patterns, it is particularly suitable for financial time series forecasting. The LSTM core unit consists of forgetting gate, input gate and output gate to regulate information flow.

In order to enhance the prediction ability of the model, a comprehensive feature set is constructed. Technical indicators include the simple moving average (SMA), relative strength index (RSI), MACD (exponential moving average convergence divergence), and Bollinger Bands. Fundamental ratios include return on equity (ROE), debt-to-equity ratio, and year-over-year growth rate of earnings per share (EPS). Lagged returns from 1, 5, 10, and 20 trading days are also included to capture momentum effects. Correlation analysis is used for feature selection, removing redundant variables to reduce computational complexity.

ISSN: 2692-7608 DOI: 10.6981/FEM.202609_7(9).0010

3.4. Investment Decision Optimization Model

As shown in Figure 1, the optimization framework adopts a two-stage structure: signal generation and asset allocation.

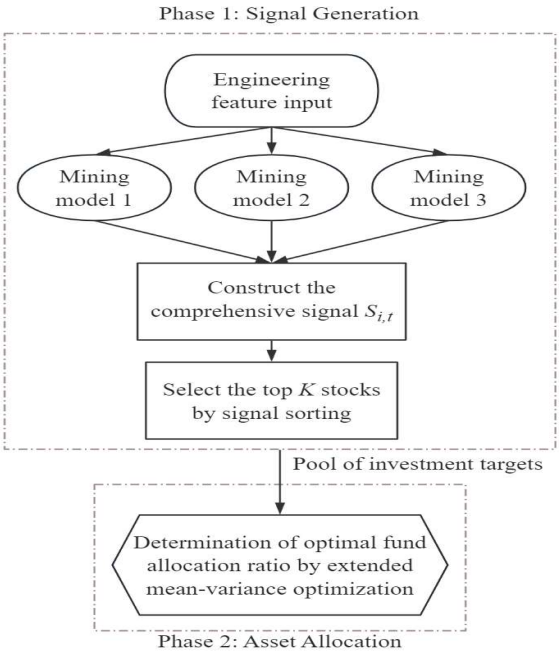


Figure 1. Framework of investment decision-making optimization model

Phase 1 signal generation. On each trading day, three data mining models use engineering features to generate the earnings forecast of each stock. In order to reduce the deviation of a single model, the predicted values of the three models are averaged with equal weight to construct a comprehensive signal $S_{i,t}$. Then the stocks are sorted according to the signal score, and the top K stocks are selected as the investment targets for the next phase of asset allocation.

Phase 2 asset allocation. The extended mean-variance optimization framework is adopted to determine the optimal allocation ratio of funds among the selected stocks. The optimization problem is expressed as follows:

$$\min_{\mathbf{w}} \mathbf{w}^T \sum \mathbf{w} - \lambda \mathbf{w}^T \boldsymbol{\mu} \tag{3}$$

Constraints:

$$\sum_{i=1}^{n} w_i = 1, \quad w_i \geq 0 \ \forall i \tag{4}$$

Among them, $\mathbf{w} = (w_1, w_2, \cdots, w_n)^T$ is the weight vector of the investment portfolio; Σ is the covariance matrix of asset returns for $n \times n$; $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^T$ is the predicted revenue vector generated in stage 1; λ is the risk aversion coefficient, which controls the trade-off between income maximization and risk minimization; n is the number of stocks selected. Non-negative constraint $w_i \geq 0$ prohibits short selling, which is consistent with the realistic constraints faced by many institutional investors. In the out-of-sample interval from January 2023 to December 2025, the above strategies were back-tested, and the positions were adjusted once a month.

ISSN: 2692-7608 DOI: 10.6981/FEM.202609_7(9).0010

4. Empirical Analysis and Results

4.1. Prediction Performance Comparison

To evaluate the performance of three data mining methods in stock return prediction tasks, this paper focuses on daily return predictions using an out-of-sample test set (January 2023 – December 2025), employing four standard evaluation metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R^2). The predictive performance of each model is summarized in Table 1.

Table 1. Comparison of out-of-sample prediction performance of each model

Model	MAE	RMSE	MAPE (%)	R^2
XGBoost	0.0011	0.0027	1.095	0.9999
LSTM	0.0116	0.0189	7.012	0.9995
Random forest	0.0158	0.0243	9.876	0.9991

The results in Table 1 show that XGBoost is the best in all evaluation indexes, and its MAE and RMSE are lower than those of LSTM and random forest. MAPE of XGBoost is only 1.095%, while LSTM and random forest are 7.012% and 9.876% respectively, indicating that XGBoost has obvious advantages in relative prediction accuracy. The prediction performance of LSTM is second, and the random forest is relatively weak. This result may be related to XGBoost's natural advantages in processing high-dimensional table data. Its gradient lifting framework can effectively capture the nonlinear interaction between features, while LSTM is good at capturing time-series dependence, but it may be limited by the expression efficiency of circular structure when facing a large number of cross-sectional features.

4.2. Investment Strategy Backtesting Performance

Based on the signals generated by each model, this paper constructs the corresponding portfolio, and carries out monthly position adjustment and back test in the out-of-sample interval (January 2023–December 2025). In order to control the risk of portfolio concentration, 10 stocks with the highest signal score are selected for equal allocation in each period. At the same time, with the market value weighted benchmark index as the control, the risk-adjusted return performance of each strategy is evaluated. The back test performance is summarized in Table 2.

Table 2. Comparison of the performance of each strategy (January 2023–December 2025)

Tactics	Annualized rate of return (%)	Annualized volatility (%)	Sharpe ratio	Maximum retracement (%)	Winning percentage (%)
XGBoost strategy	20.56	18.42	0.81	-20.36	58.3
LSTM strategy	16.38	19.75	0.64	-23.51	54.2
Random forest strategy	12.10	20.14	0.52	-25.40	50.0
Benchmark index	8.92	17.83	0.38	-28.67	45.8

Table 2 shows that XGBoost strategy is superior to other strategies and benchmark indexes in annualized rate of return, Sharp ratio and maximum retracement. The annualized rate of return of LSTM strategy is 16.38%, and Sharp ratio is 0.64, followed by performance. The annualized rate of return of stochastic forest strategy is 12.10% and the Sharp ratio is 0.52. The annualized rate of return of the benchmark index is only 8.92%, and the Sharp ratio is 0.38. It is noteworthy that XGBoost strategy not only gains the highest income, but also its maximum retreat is less

ISSN: 2692-7608 DOI: 10.6981/FEM.202609_7(9).0010

than the benchmark index, which shows that this strategy not only has advantages in the income side, but also shows outstanding performance in risk control. Although the income of LSTM strategy is lower than XGBoost, its winning rate is still higher than the benchmark, which shows that the time series model still has some effective information in judging the market direction.

4.3. Cumulative Income and Net Value Trend

Figure 2 shows the trend of the cumulative net value of each strategy and benchmark index during the out-of-sample back test. The cumulative rate of return of XGBoost strategy reached about 33.26% at the end of the backtesting period, which was higher than that of LSTM strategy and random forest strategy, while the cumulative rate of return of benchmark index was about 14%. XGBoost strategy is particularly prominent in the upward phase of the market from the second half of 2023 to the beginning of 2024, and the accumulated income rises rapidly; However, in the market callback stage in mid-2024, its net value retracement range is obviously smaller than the benchmark index, which reflects that the signals generated by data mining methods have certain defensive characteristics when the market goes down.

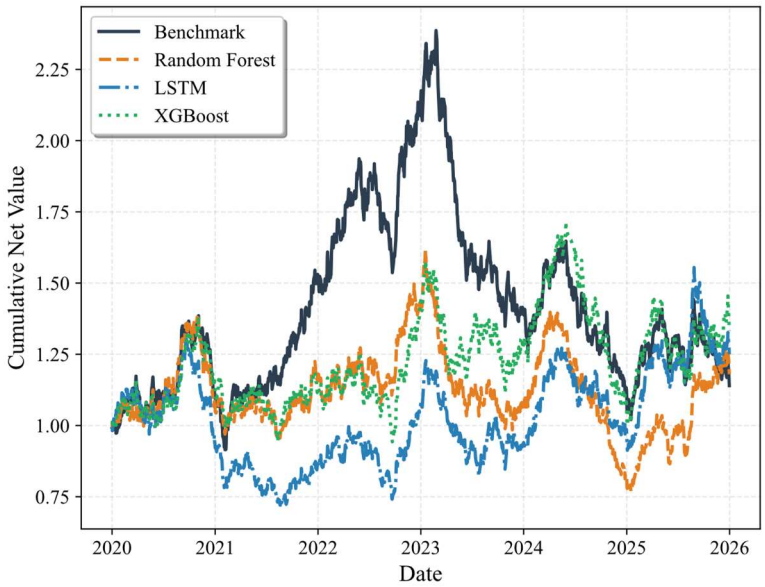


Figure 2. The cumulative net value trend of each strategy and benchmark index during the out-of-sample backtesting period

4.4. Robustness Test

In order to ensure the reliability of the above conclusions, this paper carries out two robustness tests. (1) the strategic performance under different shareholding numbers. In order to test whether the strategic performance is affected by the number of shares set, this paper tests three scenarios of the number of shares K=5, K=8 and K=10 respectively. The results show that when K=5, the annualized rate of return is the highest, about 23%, but the volatility and maximum retracement also increase accordingly. When K=10, the combined annualized rate of return is about 20.56%, the volatility is reduced to 18.42%, and the risk-adjusted return is the best. This result shows that moderate decentralized allocation is helpful to achieve a better balance between income and risk. (2) The performance of different sample subintervals. In this paper, the out-of-sample interval is further divided into two sub-intervals of 2023 and 2024-2025, which are tested separately. The Sharp ratios of XGBoost strategy in the two sub-intervals are 0.92 and 0.73, respectively, which are significantly higher than the benchmark index of 0.45

ISSN: 2692-7608 DOI: 10.6981/FEM.202609_7(9).0010

and 0.31, indicating that the strategy has maintained a relatively stable excess return capability in different market environments.

5. Conclusion

The empirical results verify the effectiveness of the investment decision-making framework integrating multi-source financial data mining signals. In the aspect of forecasting performance, XGBoost is superior to LSTM and random forest in MAE, RMSE, MAPE and R^2, which shows the outstanding advantages of gradient lifting method in financial income forecasting task. In terms of investment performance, the portfolio based on XGBoost signal is superior to the comparison strategy and benchmark index in annualized rate of return, Sharp ratio and maximum retracement, which verifies the effectiveness of data mining method in improving the quality of investment decision. The two-stage investment optimization framework constructed in this study has strong applicability. For individual investors, we can refer to this method to build a lightweight intelligent stock selection strategy based on open market and financial data to replace the traditional trading model that relies entirely on subjective experience and improve the scientific decision-making. For quantitative funds and financial institutions, the framework can be used as a supplementary module of the existing algorithmic trading system, which can optimize the asset allocation weight by integrating multi-model forecasting signals, and steadily improve the risk-adjusted income under strict risk constraints. There is still room for further improvement in this study. At present, the feature set has not been included in the unstructured data dimension. NLP technology can be introduced to analyze the sentiment of research reports, social media public opinions and policy announcements, and unstructured signals can be included in the forecasting system to further enhance the model's ability to capture market information.

References

[1] Shava, H. (2026). Entrepreneurial alertness, managerial networks and innovation culture as catalysts to sustained small and medium-sized enterprise digital transformation. Journal of Entrepreneurship in Emerging Economies, 18(7), 281-317. https://doi.org/10.1108/JEEE-02-2025-0052

[2] Zhang, J., Xu, Z., & Wang, S. (2022). A novel hybrid learning paradigm with feature extraction for carbon price prediction based on Bi-directional long short-term memory network optimized by an improved sparrow search algorithm. Environmental Science and Pollution Research International, 29(43), 65585-65598. https://doi.org/10.1007/s11356-022-20450-4

[3] Zou, Y., Liu, Y., & Chen, J. (2024). Optimal weight random forest ensemble with Fuzzy C-means cluster-based subsampling for carbon price forecasting. Journal of Intelligent & Fuzzy Systems, 46(1), 991-1003. https://doi.org/10.3233/JIFS-233422

[4] Li, H., Jin, X., & Liu, Y. (2026). Credibilistic portfolio optimization with investors' coherent perceptions: a three-way decision approach. Soft Computing, 30(6), 1-40. https://doi.org/10.1007/s00500-025-10933-0

[5] Cao, Q., Zhao, T., Guo, Q., et al. (2026). Impact of Financial Misallocation on Enterprise Digital Transformation. Emerging Markets Finance and Trade, 62(10), 3639-3659. https://doi.org/10.1080/1540496X.2025.2599432

[6] Wang, H. (2025). A feature mining model for financial product purchase behaviour data based on consumer behaviour perspective. International Journal of Business Intelligence and Data Mining, 26(1-2), 205-217. https://doi.org/10.1504/IJBIDM.2025.143935

[7] Găman, Ş., Tak, R., & Jheng, S. L. (2026). ESG-Aware Asset Selection and Portfolio Optimization: A Comparative Study of Multi-Criteria Decision-Making Methods. Proceedings of the International Conference on Business Excellence, 20(1), 133-147. https://doi.org/10.2478/PICBE-2026-0016

ISSN: 2692-7608 DOI: 10.6981/FEM.202609_7(9).0010

[8] Zeng, Z. (2026). From Price to Returns: The Performance Evolution of Machine Learning Models in Stock Prediction Tasks. Journal of Innovation and Development, 15(2), 178-184. https://doi.org/10.54097/PRGT6724

[9] Xu, X., Ye, T., Gao, J., et al. (2024). The effect of green, supply chain factors in predicting China's stock price crash risk: evidence from random forest model. Environment, Development and Sustainability, 27(10), 1-24. https://doi.org/10.1007/s10668-023-04300-Y