European Journal of Business, Economics & Management

Vol. 1 No. 3 2025

Article **Open Access**

Exploration of Optimization Paths Based on Data Modeling in Financial Investment Decision-Making

Chuhan Wang [1,*]

[1] Carey Business School, Johns Hopkins University, Washington, D.C., 20001, USA
* Correspondence: Chuhan Wang, Carey Business School, Johns Hopkins University, Washington, D.C., 20001, USA

Abstract: With the continuous development of the financial industry, the data analysis in the investment decision process has become more and more complex. The application of data modeling technology in financial investment decisions has become a key tool to improve the accuracy and efficiency of decision making. This paper delves into conventional financial investment model regression analysis and machine learning, analyzing challenges such as insufficient data quality, market volatility, hardware resource constraints, and model overfitting. In order to improve the stability and generalization ability of the model, some optimization paths are proposed, such as improving the quality of data preprocessing, introducing robust models and optimizing distributed computing.

Keywords: financial investment decision; data modeling; data preprocessing; market volatility; hardware resource optimization

Received: 18 June 2025
Revised: 01 July 2025
Accepted: 16 July 2025
Published: 20 July 2025

1. Introduction

With the increasing complexity of financial markets, investment decisions based on traditional data have found it difficult to meet the requirements for high speed and precision. Regression analysis, machine learning, and other methods in data modeling technology play an important role. However, data quality, market fluctuation, hardware resource limitation and model overfitting are still the thorny problems that affect the accuracy and stability of model results. The purpose of this paper is to discuss the data modeling methods in financial investment decision making and the challenges it faces, and put forward the optimization path, so as to make the investment decision model achieve the goal of higher efficiency and better effect.

2. Data Modeling Methods Commonly Used in Investment Decision-Making

2.1. Regression Analysis Method

As a common statistical modeling method, regression analysis is widely used in investment decision-making, its purpose is to predict the value of the target variable by establishing the mathematical relationship between the independent variable and the dependent variable. This method is commonly used in financial investment to predict stock prices, asset returns, and market trends. At the same time, according to the number of independent variables involved, it can be divided into simple linear regression, multiple linear regression and nonlinear regression. Among these types of regression, simple linear regression shows the relationship between the independent and dependent variables in a straight line. Multiple linear regression includes all the independent variables that can

affect the dependent variables.[1] Because this method directly expresses the linear relationship between various variables, and the calculation process is simpler, regression analysis is chosen by many investors. In addition, regression analysis can also be used in the process of stock price prediction, asset return analysis and market trend judgment and identification, and provides more effective guidance information for the above process. Figure 1 summarizes the basic flow and model building process of regression analysis:

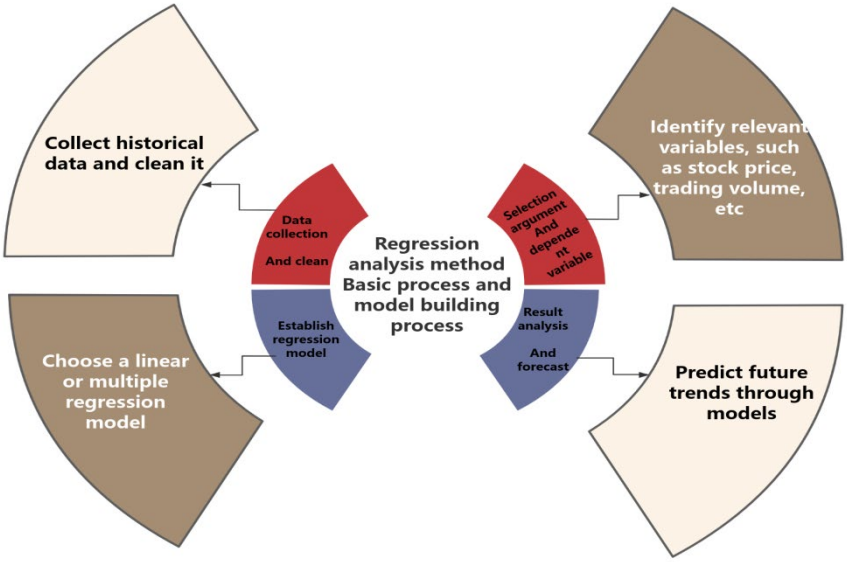

Figure 1. Basic flow and model building process of regression analysis.

2.2. Machine Learning Algorithms

Machine learning algorithms are widely used in financial investment decisions. Using machine learning algorithms, features can be extracted from previous historical data to build predictive models to predict future stock prices, market trends and asset returns. Compared to traditional regression analysis, machine learning algorithms are better at handling complex, non-linear associations and can handle more multi-dimensional information. Machine learning includes supervised learning (linear regression, support vector machines (SVM), decision trees and random forests), unsupervised cluster analysis, and deep learning (artificial neural networks).[2] The main advantage of machine learning is its ability to handle large-scale data. At the same time, the model can continue to self-learn to improve its prediction accuracy. Figure 2 below summarizes the workflow and main links of the machine learning algorithm:

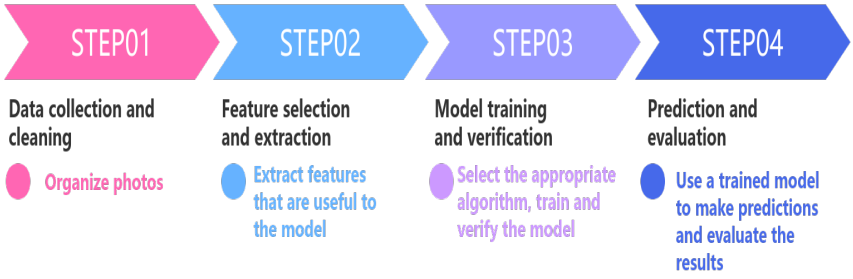

Figure 2. Workflow and main links of machine learning algorithm.

3. Challenges of Data Modeling in Financial Investment Decision-Making

3.1. Insufficient Data Quality Affects Accuracy

Data quality is very important for the construction of effective models. If the data quality is insufficient, especially if the data is missing, wrong, duplicated or inconsistent, the accuracy of the model is likely to be affected. Insufficient data quality is likely to cause the model to make incorrect predictions that do not accurately reflect the actual situation. In addition, missing data can cause modelers to be forced to make inaccurate fillings, further degrading the model. Table 1 below summarizes the common types of data quality problems:

Table 1. Common types of data quality problems.

Problem type	Description	influence
Data missing	Some data items are missing or unavailable	This results in incomplete data and affects the comprehensiveness and accuracy of the model
outlier	Extreme values or incorrect data exist in the data	It affects the fitting effect of the model, resulting in inaccurate prediction results
Data inconsistency	Different data sources have different data formats or standards	Makes data difficult to combine and process, and may introduce inconsistent results
Duplicate data	The same data item is repeated in multiple places	Increasing the amount of computation will affect the training and prediction results of the model

As can be seen from Table 1, insufficient data quality will lead to distortion of the model's prediction of market behavior, thus affecting investment decisions.

3.2. Market Volatility Weakens Stability

Market volatility is mainly manifested by large changes in price or trading volume. This uncertainty and volatility are also major factors that hinder investors from making correct decisions, and it affects the stability of data modeling. Because it is difficult to predict the return in the highly volatile market environment, the model loses its accurate or reliable prediction effect due to the recent market fluctuations. Especially when using traditional regression analysis or machine learning, volatility can cause models to overfit short-term data and ignore long-term trends. In addition, market volatility increases investment risk, potentially invalidating the model's predictions.[3] In the face of extreme market conditions such as financial storms or sudden events, current forecasting models often fail to adapt to these drastic changes, leading to decision-making errors. Table 2 below summarizes the impact of market volatility on model stability:

Table 2. Influences of market volatility on model stability.

Influencing factor	Description	result
Market shock	Sudden economic events or news reports cause the market to fluctuate wildly	Model forecasting failure, can not reflect the market changes in a timely manner
Long-term trends and short-term fluctuations	Market volatility causes models to focus too much on short-term changes and ignore long-term trends	Resulting in a decline in the accuracy of long-term investment decisions

Risk management	High volatility markets increase investment risk	It makes the model difficult to cope with the complex market environment stably
The model is overfitted	High volatility can cause models to overfit short-term volatility data	The model loses the adaptability to the long-term stable trend and reduces the reliability

Table 2 shows that market volatility poses a major challenge to the stability of financial investment models. Sudden events and short-term fluctuations can easily lead to model prediction failure or overfitting, which affects the accuracy of long-term decision-making.

3.3. Hardware Resources Limit Computing Efficiency

Data modeling usually requires processing large amounts of historical and real-time data. As the amount of data increases and models become more complex, hardware devices such as computing power, storage space, and memory are key determinants of computing performance. The lack of hardware equipment can lead to data processing disruptions, which in turn reduce the real-time and accuracy of the model, especially in a market environment where rapid decision making is required, if the required computing power is not available, researchers may need to use simplified models or operate in a batch manner, which can lead to certain errors and reduced expected effectiveness. Table 3 below summarizes the impact of hardware resource constraints on computing efficiency:

Table 3. Effects of hardware resource constraints on computing efficiency.

Influencing factor	Description	result
undercomputing	The computing and processing power of the hardware is weak, and it cannot process complex models quickly	The model training time is long and the analysis cycle is delayed, which affects the decision-making efficiency
Storage capacity limitation	When there is a large amount of data, storage resources are insufficient, resulting in slow data processing	It limits the real-time processing and efficient analysis of large-scale data
Memory bottleneck	The memory resources are insufficient to process multiple computing tasks simultaneously	As a result, the calculation process stalls or memory overflows, reducing efficiency
Parallel computing difficulty	Lack of hardware to support parallel computing or misallocation of resources	The process of model training and data analysis cannot be accelerated, and the calculation time can be increased

As can be seen from Table 3, the limitations of hardware resources significantly affect computing efficiency, especially when dealing with big data and complex models. Insufficient computing power, storage, memory, and parallel computing lead to delays that affect the timeliness and accuracy of decisions.

3.4. Model Overfitting Reduces the Effect

When a model performs better than expected on training samples but poorly on new samples, this is called overfitting. Overfitting usually occurs when the model is too complex, when the model pays too much attention to the noise and randomness in the training sample, and can not find the real data law.[4] Although the model may have very high

accuracy on past samples, it lacks the ability to accurately predict future real-world situations, especially in coping with market volatility. The appearance of overfitting will reduce the generalization ability of the model, leading to failure to correctly predict future trends. Especially in the financial market, the price trend will be affected by many factors. If the model pays too much attention to some special patterns in the past samples, it may not be able to cope with future uncertainties. Table 4 below summarizes the impact of overfitting on the model effect:

Table 4. Influence of overfitting on model effect.

Influencing factor	Description	result
Complex model	The complexity of the model is too high, the parameters are too many, and it is easy to remember the noise in the training data	Results in good performance of training data, but poor ability to predict new data
Insufficient data	The amount of training data is too small, and the model is easy to overfit the limited data	The generalization ability of the model is reduced, and the prediction results are unstable
Feature misselection	Too many features or irrelevant features are involved in training, which increases the complexity of the model	As a result, the model captures meaningless patterns and reduces the prediction accuracy
The training time is too long	Training for a long time can cause the model to overfit the details of the training set	The generalization of the model is affected and the prediction error is caused

As can be seen from Table 4, overfitting will result in excellent performance of the model in training data, but poor effect in actual prediction.

4. Optimization Path Based on Data Modeling in Financial Investment Decision-Making

4.1. Improve Data Preprocessing Quality and Optimize Modeling Results

Data preprocessing is the basis to ensure the modeling effect. High-quality preprocessing can significantly improve model accuracy and stability, including data cleaning, missing value processing, outlier detection, data standardization, and feature engineering. Data cleaning mainly includes eliminating redundant data, data correction, filling missing values, etc. The selection of missing value filling method is very important. The commonly used filling methods include using average value filling, filling interpolation, etc. The detection and processing of outliers is to filter out the key data that violates the law of the market. If these outliers are not processed, the model may cause the wrong learning of the data. In addition, it is also very important to standardize and normalize data, especially in algorithms used to calculate distance, such as KNN algorithm or SVM algorithm, which can help eliminate dimensional differences between features and ensure that the model can fairly deal with the influence of various features. Feature engineering, by selecting and transforming the most relevant features, enhances the model's ability to identify market rules, and improves the model's explanatory and forecasting ability. Path formula for optimized data preprocessing:

$$ModelPerformance = \sum_{i=1}^{n}(w_i \cdot P_i) \quad (1)$$

Among P_i Represents the effect of each preprocessing step (such as data cleaning, missing value filling, outlier processing, etc.), and w_i is the weight of each step. By adjusting the weights of each step, the overall performance of the model can be optimized.

4.2. Introduce a Robust Model to Deal with Market Volatility

In financial investment, market fluctuations are generally manifested by changes in price and quantity. Robust models have strong robustness against noise and disturbances

and can make reliable predictions under uncertainty and abnormal conditions, effectively dealing with market fluctuations. By reducing data noise and response sensitivity of extreme events, robust models can effectively improve the forecasting ability under market instability conditions. Robust regression (such as Huber regression), robust support vector machines (SVM), and integrated learning algorithms (such as random forests and XGBoost, etc.) can reduce the impact of volatility and avoid overfitting short-term fluctuations. The main advantage of the robust model is that it can constrain or assign higher weights to outliers and extreme values, so that they can still make stable judgments in the face of large market fluctuations. To achieve robustness, the model is usually properly regularized. Methods such as cross-validation are used to evaluate its performance in volatile markets, aiming to improve the model's generalizability and ensure its suitability across different markets. Robust model path formula:

$$RodustPerformance = \alpha \cdot Noise\ Re\ s\ is\ tan\ c\ e + \beta \cdot VolatilityAdaptation \quad (2)$$

Where, α and β are weight coefficients, representing noise resistance and adaptability to market volatility respectively. By optimizing these two factors, we can improve the performance of robust models in volatile markets.

4.3. Optimize Distributed Computing to Improve Hardware Resource Utilization

Distributed computing has become an important way to improve the utilization efficiency of hardware resources. Distributed computing is to decompose complicated data analysis tasks into several computing nodes, which are completed by different machines. It is a working mode that can give full play to the efficient capability of the entire hardware resources, greatly improving the computing speed and saving the time of data processing and model construction. Distributed computing is not only suitable for big data, but can also allocate resources based on the workload of each computing unit to prevent overload or redundancy in any single machine node. At the same time, the system structure of distributed computing also has high availability and fault tolerance, even if some components have problems, it will not affect the normal operation of the whole system. Especially in financial market data analysis, real-time requirements are high, and distributed computing can effectively improve the response speed and processing capacity of the model. In addition, two strategies, data parallelism and task parallelism, are often used to optimize the performance of distributed computing. Data parallelism refers to dividing big data into smaller subsets that are processed separately, while task parallelism involves executing multiple computing tasks simultaneously. Through these two strategies, computing resources can be used efficiently and the efficiency of model training and data processing can be improved. Distributed computing optimization formula:

$$TotalEfficiency = \sum_{i=1}^{n} w_i \cdot R_i \quad (3)$$

Among them, R_i represents the contribution of each optimization step (such as parallel processing, load balancing, fault tolerance mechanisms, etc.) to the overall efficiency, w_i is the weight of each link. By optimizing the weights of these links, the overall efficiency of distributed computing and hardware resource utilization can be improved.[5]

4.4. Regularization Technology Was Applied to Alleviate Overfitting of the Model

Regularization alleviates overfitting by introducing penalty terms that constrain model complexity, thereby enhancing its generalization performance. The core idea of regularization is to limit overly complex models and reduce the excessive dependence of models on training data. L1 regularization (also known as Lasso) and L2 regularization (also known as Ridge) are two common regularization methods. L1 regularization uses an absolute value penalty coefficient. When the absolute value of a feature coefficient is small, L1 regularization can reduce it to zero, thus achieving feature selection. L2 regularization uses a square penalty coefficient with which the model coefficients decrease, thereby reducing overfitting. Using these two methods can effectively reduce the sensitivity of the model caused by training samples and improve the prediction ability of the model against

the test samples. In practice, regularization parameters—such as the coefficients in L1 and L2—are optimized via cross-validation to ensure both model stability and strong generalization across different datasets. Regularization formula:

$$CostFunction = LossFunction + \lambda \cdot \sum_{i=1}^{n} ||\theta_i||^2 \quad (4)$$

Among them, λ is a parameter of the regularization strength, θ_i is the parameter of the model, $||\theta_i||$ represents the L2 norm of the parameter. By adjustment λ can control the weight of the regularization term, so as to balance the complexity and training errors of the model and avoid overfitting.

5. Conclusion

The application of data modeling methods has become a key tool to improve decision quality and prediction accuracy. This paper analyzes the application of regression analysis and machine learning algorithms in the financial sector, and explores challenges such as data quality, hardware resource constraints, and model overfitting in the context of increasing market volatility and data complexity. In the face of these challenges, optimization paths such as optimizing data preprocessing, introducing robust models, improving hardware resource utilization and applying regularization technology are proposed. These efforts aim to enhance the model's stability and generalizability, thereby improving the timeliness and accuracy of investment decisions. In the future, with the continuous development of technology, data modeling methods will play a greater role in the financial field, driving the industry to a more intelligent and data-driven direction.

References

1. M. Kim and X. Lu, "L2 English speaking syntactic complexity: Data preprocessing issues, reliability of automated analysis, and the effects of proficiency, L1 background, and topic," *Mod. Lang. J.*, vol. 108, no. 1, pp. 270–296, 2024, doi:10.1111/modl.12907.
2. D.-D. Dau, S. Lee, and H. Kim, "A comprehensive comparison study of ML models for multistage APT detection: focus on data preprocessing and resampling," *J. Supercomput.*, vol. 80, no. 10, pp. 14143–14179, 2024, doi:10.1007/s11227-024-06010-2.
3. F. Brunner, F. Gamm, and W. Mill, "MyPortfolio: The IKEA effect in financial investment decisions," *J. Bank. Finance*, vol. 154, Art. no. 106529, 2023, doi:10.1016/j.jbankfin.2022.106529.
4. S. Hussain and A. Rasheed, "Risk tolerance as mediating factor in individual financial investment decisions: a developing-country study," *Stud. Econ. Econom.*, vol. 47, no. 2, pp. 185–198, 2023, doi:10.1080/03796205.2023.2218053.
5. L. Özdemir, N. Uğur, A. Kaya, M. Yılmaz, F. Aydın, İ. Demir, et al., "Sovereign credit default swap market volatility in BRICS countries before and during the COVID-19 pandemic," *Sci. Ann. Econ. Bus.*, vol. 71, no. 1, pp. 21–42, 2024, doi:10.47743/saeb-2024-0005.