WHIOCE PUBLISHING PTE. LTD.
PROVIDING
FIRST-CLASS SCIENTIFIC INFORMATION
FOR TOP SCHOLARS

2025 Volume 3, Issue 5
Lecture Notes in Education, Arts, Management and Social Science
ISSN(Online): 2705-053X

Research on Stock Selection Model and Portfolio Based on Machine Learning

Yaqian Kang, Zhilong Yang, Shengsheng Zhang, Yaoguo Li, Shuyun Li
Wuwei NO.6 High School, Wuwei 733000, Gansu, China

Abstract: Stock and portfolio is an important part of the financial industry, which is affected by many factors and changes at the same time there are certain rules. The rapid development of big data and artificial intelligence technology represented by machine learning provides investors with new tools for stock selection and investment portfolio. Based on this, this paper will analyze the research background and significance of the stock selection model and investment portfolio based on machine learning, and combine the development status at home and abroad and relevant theoretical basis, to explore the stock selection model and investment portfolio application based on machine learning.

Keywords: Machine learning; Stock selection model; portfolio

Online publication: June 26, 2025

1. Introduction

Machine learning, as an algorithmic technology capable of processing large amounts of complex data, can be applied to the financial field to mine potential non-linear relationships that affect stock movements, thereby improving the predictive power and stability of stock picking models and providing investors with more accurate stock picking recommendations. At the same time, by combining stock selection model and portfolio, it can not only help investors reduce investment risks and maximize returns, but also enrich the theoretical system of quantitative investment, and further promote the innovation and development of the financial industry.

2. Research background and significance

2.1. Research background

In the 1990s, China established stock exchanges in Shanghai and Shenzhen successively, which officially kicked off the development of the domestic securities market. In the new era, the stock market, as an important investment channel, has attracted the attention of many investors. However, the volatility and uncertainty of the stock market increase the difficulty and risk of investment. With the wide application of Internet technology and artificial intelligence in the financial field, machine learning, with its strong learning ability and generalization ability, has become one of the ways to help investors to quantitatively speculate on the risks and returns of stocks, and then optimize the composition

of investment. Compared with the traditional current econometrics model, machine learning is more prominent in dealing with nonlinear variables. The application of machine learning in the field of stock selection model can provide reasonable explanations for people when they encounter uncertain problems in the financial field, and enhance the scientific nature of quantitative stock selection. However, how to design a learning model that can predict the trend of stock market and optimize the investment portfolio is still a problem that investors need to face.

2.2. Research significance

With the progress of science and technology and the continuous development of the financial field, the traditional manual stock selection has gradually failed to meet the needs of investors in the new era. Machine learning has powerful data processing ability and forecasting ability, which can provide new ideas for stock selection and portfolio optimization. On the one hand, at present, artificial intelligence and big data technology represented by machine learning are increasingly applied in the financial field. In-depth research on stock selection model and investment portfolio based on machine learning is of great practical significance to promote technological innovation in the financial field and improve the work efficiency of financial market. On the other hand, traditional stock selection and portfolio construction often rely on investors to make judgments based on professional knowledge and industry experience, which is subjective to a certain extent. Sometimes it is difficult to accurately capture market trends and individual differences, but machine learning can realize accurate prediction of market trends by collating and training massive historical data, and help investors develop more scientific and reasonable investment strategies. For example, by analyzing a stock's data and real-time market dynamics, machine learning can more accurately identify abnormal fluctuations and potential risk points in the market, so as to help investors adjust their portfolios in a timely manner, avoid unnecessary losses and reduce investment risks.

In addition, this study not only focuses on the application of machine learning in stock selection models and investment portfolios, but also discusses the relevant theoretical basis and evaluation optimization methods, providing new ideas and directions for the innovative application research of artificial intelligence technology with machine learning in the financial field, which enriches and expands the research field of finance to a certain extent.

3. Research status at home and abroad

3.1. Domestic research status

Although domestic research work on stock selection model and portfolio based on machine learning started late, many researchers have conducted research on the quantification of the domestic market by summarizing and analyzing foreign quantitative models. Li Bin, Shao Xinyue, Li yueyang reviewed the application of machine learning in quantitative investment, focusing on the research progress of stock selection model and portfolio optimization[1]. The application of support vector machine, random forest, deep learning and other models in stock selection was analyzed in detail, and the potential of reinforcement learning in dynamic asset allocation was discussed. In addition to pointing out the problems of data quality and model explainability, the future research direction is proposed. Jiang Binxin, Zhao Shengzhe, Huang Yahe discussed the application of machine learning in the financial field from a macro perspective, including stock selection, portfolio optimization, risk management, etc[2]. The advantages of machine learning models in processing high-dimensional data and nonlinear relationships were analyzed, and the current research status at home and abroad was summarized. In addition, many scholars have also discussed the multi-factor quantitative stock selection strategy based on machine learning, and found that the multi-factor quantitative stock selection strategy optimized by machine learning can significantly improve investment returns and reduce investment risks. Some scholars have also conducted in-depth exploration on the application of stock portfolio optimization and price prediction based on machine learning, and put forward the relevant discussion that optimization and price prediction of stock portfolio through machine learning algorithm can significantly improve the return rate and risk control ability of portfolio.

3.2. The status quo of foreign research

The research on quantitative investment in foreign countries is relatively long, and the research field is more comprehensive. For example, the research on the construction of multi-factor stock selection model not only considers the pricing of various assets, but also considers the application of feature engineering and machine learning methods.

Fama and French proposed the famous FamA-French three-factor model in the 1990s, which laid the foundation for the research of fact-based stock selection[3]. While the study did not directly use machine learning methods, its multifactor framework provided a theoretical basis for subsequent machine learning stock picking models. The research shows that factors such as market capitalization and book-to-market ratio have significant explanatory power to stock returns. Tang Guohao, Zhu Lin, Liao Cunfei, Jiang Fuwei systematically evaluated the application of machine learning in asset pricing[4]. After comparing the performance of various machine learning models such as random forest, gradient lift tree and neural network in predicting stock returns, it is found that nonlinear models, especially neural networks, are significantly better than traditional linear models. The study also highlights the importance of feature engineering and provides an empirical framework for the application of machine learning in the financial field. Bianchi et al. applied machine learning to bond risk premium forecasting. By comparing the performance of various machine learning models in predicting bond returns, it is found that deep learning models have significant advantages in processing time series data, which provides an important reference for the application of machine learning in fixed income asset pricing.

4. Relevant theoretical basis

4.1. Basic concepts and main methods of machine learning

Machine learning, a branch of artificial intelligence, is an artificial intelligence technique that enables computers to learn from data and improve their performance without explicit programming. In recent years, by applying machine learning to finance, investors and financial institutions have achieved a better understanding of market dynamics, identification of investment opportunities and management of risks. In practical applications, different machine learning approaches are able to solve different types of problems. Here are some of the main methods of machine learning. Supervised learning refers to training by learning input features and output labels during training. In the financial field, it can be used to predict the stock price trend and build forecasting models. Common supervised learning methods include linear regression, logistic regression, support vector machines, decision trees, etc. Unsupervised learning models have no label information during training and are designed to uncover hidden structures or patterns in the data[5]. They can be used in market factor analysis, cluster analysis, etc. Common unsupervised learning methods include clustering of K-means clustering, reduction and anomaly detection represented by principal component analysis, etc. Reinforcement learning models learn behavioral strategies by interacting with the environment in order to maximize some kind of cumulative reward. By simulating the stock-picking or investment decision process, it optimizes investment strategies and increases returns through trial-and-error learning. Deep learning refers to a type of machine learning method that uses deep neural networks for learning. This method has significant advantages in dealing with massive and complex data, and can efficiently process financial data such as stock trading data and market sentiment data. In short, different machine learning methods have their respective advantages and disadvantages in the application of the financial field, therefore, in practical applications, investors should choose the right machine learning method according to the specific needs and problems.

4.2. Decision tree and model fusion theory

Decision tree is a kind of machine learning algorithm based on tree structure, which is widely used in classification and regression tasks. Decision trees divide a dataset by recursively selecting the best features, making the divided subset more pure. Model fusion, on the other hand, is the idea of merging multiple weak models together into a single strong model. Through model fusion, the prediction accuracy, stability and generalization ability of the model can be improved.

However, two problems need to be paid attention to in this process: first, how to select suitable multiple machine learning algorithm models; Second, several machine algorithm models are usually independent of each other, how to integrate them into a powerful algorithm model.

There are three common methods of model fusion integration: First, bagging self-help method. This is a sampling statistical method to be recovered by estimating the equation of the statistic to get the interval in non-parametric statistics. It follows that "the distribution provided by the data is the best guess when the population distribution is unknown." "Principle. Based on the bagging method, it is possible to integrate the learning subsamples of different base models from the training set to the subsampling group, so as to obtain the final prediction result. By combining Bagging with decision tree algorithm, a random forest is morphed. Boosting, which focuses more on classification errors in the final sample. By giving more weight to the classification error sample and setting the corresponding training goals, it is easier to determine the final sample error by training multiple iterations from the weak classifier to weight one strong classifier. It should be noted that in the sample selection, the training set remains the same, but the weight of each sample will be adjusted according to the classification results of the previous round. At the same time, each weak learner has a different weight, and the weight is related to its performance on the validation set. And the individual weak learners must be trained sequentially, because the latter model parameter needs to depend on the results of the previous round of models. Thus, Boosting can reduce the bias of the model and improve the accuracy of the model by gradually correcting the errors of the previous learner. The hierarchical model integration framework is stacking. The prediction results of multiple base models are input into a sub-model as new features, and the sub-model outputs the final prediction results. Stacking combines the predictions of multiple base models to integrate the benefits of different models and improve overall forecasting performance.

4.3. Application of machine learning in portfolio

In the Internet era, machine learning has become an information tool in investment portfolio that can improve decision-making efficiency and predict market development and risks based on complex models, which helps investors reduce risks and improve returns to a certain extent. The application of machine learning method to predict the trend of stock prices, exchange rates, etc., is the most intuitive one in portfolio applications. The future trend of the stock market is predicted by using linear regression, support vector and random forest in the supervised learning algorithm. This kind of prediction is the key to formulating investment strategy, which can help investors judge when it is appropriate to buy or sell a certain asset in order to get the maximum profit. In asset selection, machine-learning algorithmic models sift through vast amounts of data to select the optimal portfolio. For example, using cluster analysis, it is possible to find some asset groups with similar characteristics, so as to build a diversified and risk-diversified portfolio. In addition, reinforcement learning algorithms can simulate different investment situations, and then dynamically adjust the investment allocation ratio, so that investors can better cope with the complex and changeable market environment.

In portfolio management, risk management is a very important aspect, and machine learning provides investors with a new method, through the establishment of a forecast model for risk assessment of various assets, to help investors understand the possible downside risks in the market, so as to take effective preventive measures. In addition, the use of anomaly detection technology can identify extreme events in the market and give early warning of possible huge losses. With the rapid development of artificial intelligence technology, the application of machine learning in investment portfolio not only changes the traditional portfolio management mode, but also brings new development opportunities for the financial field. However, it should be noted that machine learning is not a "master key". The predictive models and strategy recommendations it provides are all based on high-quality data resources and appropriate model algorithm selection. Therefore, in practical application, investors need to have good professional knowledge and investment experience, so that machine learning can become an effective tool to improve the quality of investment portfolios.

5. Stock selection model and portfolio application analysis based on machine learning

5.1. Prediction design of stock selection model

5.1.1. Data source and pre-processing

The sample data selected in this study are all from the standardized panel data of 74 enterprises and the corresponding income data of individual stocks, and the industry to which a stock belongs will significantly affect its fundamental structure, so this study constructs the industry average index of each indicator simultaneously. The data span from January 2003 to December 2017.

5.1.2. Divide the training set validation set and the test set

Divide the sample data according to different proportions, as shown in the figure below. Among them, the data from 2003 to 2006 is used as the training set, the data from 2007 and 2008 is used as the verification set, and the data from 2009 is used as the prediction set for testing. After the model training, the length of the verification set and the test set remained the same, and the training set was increased by one year. The training set data is used for parameter training, and the verification set and test set are used to evaluate the accuracy of the trained model to evaluate the generalization ability of the model.

The sample data is divided according to different proportions

2003	2004	2005	2006	2007	2008	2009
Training Set					Verification Set	Prediction Set

At the same time, in order to reflect the prediction ability of different models, the out-of-sample prediction accuracy index R is also constructed

$$R^2 = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{N_t}(r_{i,t}\hat{r}_{i,t})^2}{\sum_{t=1}^{T}\sum_{i=1}^{N_t} r_{i,t}}$$

Where, $r_{i,t}$ and $\hat{r}_{i,t}$ are the actual and predicted returns of each period of stocks or portfolios respectively, and the value range of R^2 is $(-\infty, 1]$. The higher the value, the better the prediction ability of the model, and $R^2 = 1$ when the model fully predicts the return of each period of stocks.

5.1.3. Empirical analysis of earnings prediction

5.1.3.1. Linear model

Linear-class machine learning model by using small square linear regression OLS, LASSO, RIDGE regression RIDGE, elastic network ENet, principal component analysis PCA and partial least square PLS. After plugging in the corresponding data, the linear regression model R^2 is negative, indicating that the predictive ability of the model is lower than that of the random walk model[6]. Except the linear regression model, the prediction accuracy of other machine learning methods is positive, which indicates that the linear machine learning method can alleviate the overfitting problem of OLS model to a certain extent, and can effectively use the data information to complete the prediction. In order to further verify the prediction ability of the model, 500 stock sub-samples with market value top500 (the largest) and bottom500 (the smallest) were selected monthly to predict the results. Among the Top500 samples, the forecasting ability of all models has declined, especially PCA and PLS, and the T-value is also low, which indicates that the linear model has a relatively stronger forecasting ability for small-market stocks to some extent.

Table 1. Out-of-sample prediction R of the linear model

		OLS	LASSO	RIDGE	ENet	PCA	PLS
all	R^2	-5.21	0.20	0.21	0.23	0.36	0.37
	T-value		1.62	1.67	1.68	2.62	2.68
	P-value		0.05	0.05	0.05	0	0
top500	R^2	-6.40	0.16	0.16	0.20	0.18	0.15
	T-value		1.92	1.95	2.01	2.78	2.01
	P-value		0.03	0.03	0.02	0.02	0.02
bottom500	R^2	-3.35	0.25	0.29	0.28	0.54	0.56
	T-value		2.28	2.36	2.37	1.94	2.32
	P-value		0	0	0	0.03	0

5.1.3.2. Nonlinear model

By using enhanced gradient regression tree GBRT, random forest RF, feedforward neural networks with one to four hidden layers FFN1, FFN2, FFN3, FFN4, long short-term memory networks with one to four hidden layers LSTM1, LSTM2, LSTM3, LSTM4 And a machine learning model for generative adversarial networks GAN.

Table 2. Nonlinear model out-of-sample prediction R

		GBRT	RF	FFN1	FFN2	FFN3	FFN4	LSTM1	LSTM2	LSTM3	LSTM4	GAN
all	R^2	0.40	0.46	0.51	0.56	0.59	0.56	0.58	0.64	0.65	0.71	0.89
	T-value	2.61	2.99	2.58	2.62	2.74	2.77	3.3	4.06	2.45	2.88	4.13
	P-value	0	0	0	0	0	0	0	0	0.01	0	0
top500	R^2	0.21	0.35	0.43	0.42	0.46	0.52	0.56	0.56	0.59	0.60	0.71
	T-value	2.04	2.73	2.26	2.23	1.78	3.53	4.05	4.36	3.15	3.35	3.69
	P-value	0.03	0	0.01	0.01	0.04	0	0	0	0	0	0
bottom500	R^2	0.62	0.66	0.83	0.93	0.89	0.84	1.04	1.23	1.01	1.39	1.42
	T-value	2.79	3.13	2.79	2.95	2.17	2.49	2.98	3.71	2.19	2.76	3.01
	P-value	0	0	0	0	0.01	0	0	0	0.01	0	0

After taking into account the nonlinear information between the variables, the nonlinear model performs better overall than the linear model. The R^2 of two types of tree models reaches more than 0.40%, and the prediction accuracy of the small market capitalization sample is also higher than that of the market capitalization sample. Among neural network models, GAN model has the best performance. Multi-layer network not only brings higher complexity, but also increases the risk of overfitting. In the top500 and bottom500 subsamples, the nonlinear model is similar to the linear model, that is, for small-market stocks with higher predictive power, the GAN model is still the best performance.

5.1.3.3. Portfolio forecasting model

When building a portfolio prediction model, investors can adopt a variety of machine learning algorithms such as decision trees, random forests, support vector machines and artificial neural networks according to actual needs. These algorithms have different characteristics and advantages, and are suitable for different types of data and scenarios.

For example, decision trees and random forests are good at dealing with high-dimensional data, while support vector machines are good for nonlinear relationships. The first step in model design is to determine the objectives of the portfolio, such as maximizing the expected return or minimizing risk. Based on the current financial market development and investment strategy, the constraints including but not limited to the industry distribution of assets, liquidity requirements, the maximum investment proportion and so on are formulated. Then, the collected data were pre-processed such as data cleaning, feature selection and normalization to ensure the data quality and accuracy of the model training. Subsequently, the data is trained according to the selected machine learning algorithm. During the training process, techniques such as cross-validation are used to evaluate the performance of the model, and methods such as hyperparameter adjustment are used to optimize the model parameters. After completing the above steps, the model's performance is evaluated by calculating its prediction error, yield, and risk. Commonly used evaluation indicators include but are not limited to mean square error, mean absolute error, etc. Finally, the optimal portfolio forecasting model is used to provide investors with more comprehensive risk management and decision support.

5.1.3.4. Model evaluation and optimization

After discussing the prediction design of stock selection model and the application of portfolio prediction model, model evaluation and optimization is also an indispensable step. This stage emphasizes to measure the prediction accuracy and stability of the model through various evaluation indexes, and make necessary adjustments and improvements to the model accordingly. First, investors should identify key indicators such as accuracy rate, recall rate and F1 score according to actual needs to fully reflect the performance of the model. Among them, the accuracy rate can reflect the proportion of the number of samples correctly predicted by the model to the total predicted samples; The recall rate refers to the proportion of samples correctly identified as positive by the model in all the actual positive samples; The F1 score takes into account both accuracy and recall, avoiding possible contradictions between them. In addition to traditional evaluation indicators, machine learning-based methods also include cross-validation techniques, which can also effectively reduce the selection bias of training data sets and enhance the ability of model generalization. In practical application, by dividing the data set into several parts, using one part each time as the test set, and the rest as the training set, iterating repeatedly, more reliable results can be obtained. In this process, if the shortcomings of the evaluation results are found, the model needs to be optimized accordingly. Optimization strategies include, but are not limited to, feature selection and ensemble learning methods. Feature selection aims to select the feature subset with the most strong correlation and low redundancy from many candidate features, thereby improving the predictive performance and interpretability of the model. By fusing the prediction results of multiple independent models, ensemble learning can not only reduce the errors caused by the randomness of a single model, but also improve the overall stability and accuracy of the model. In addition, considering the volatility of the financial market environment and the stock price itself, the construction of dynamic models is also one of the effective optimization paths.

5.1.3.5. Empirical research design and analysis

In ROC space, the X-axis is the false positive case rate, and the Y-axis is the true case rate. These two values are calculated from the classification results as follows:

$$TPR = TP/ (TP + FN)$$

$$FPR = FP /(FP + TN)$$

In this study, many samples show that when XGBoost is selected as a quantitative investment tool for investors, the win rate and accuracy rate of expanding the number of transactions can reach more than 50%, which can achieve a certain income in a large number of transactions, which also indicates that the model of XGBoost algorithm has a strong forecasting ability.

Although machine learning shows good predictive ability in stock selection models and portfolios, in the actual investment process, investors also need to consider other non-linear factors and the impact of abnormal market

fluctuations, and based on the constant changes in the financial market environment and development trend, the corresponding model algorithm is regularly updated and verified.

5.1.3.6. Model risk assessment

When using machine learning to design a stock selection model or portfolio prediction model, it is also crucial to evaluate the risk of the corresponding model. For example, some models are prone to overfitting risk, that is, a model that performs well on training data but is poor at generalizing on new data. This is usually due to issues such as the limited amount of data selected and poor selection of features. This can be done by introducing regularization terms to limit the complexity of the model and prevent overfitting. Or select features that have a significant impact on the target variable, reduce redundant features, and improve the explanatory and predictive power of the model.

6. Conclusion

As one of the core technologies of artificial intelligence, the deep integration of machine learning with the financial field is currently a cutting-edge trend in academic research. This study contributes to a deeper understanding of the application mechanisms and effects of machine learning technology in the field of financial investment, providing practical experience and theoretical support for interdisciplinary research, and promoting the innovative application and development of artificial intelligence technology in the financial sector.

Funding

Special project of Shaanxi Provincial Department of education, project type: natural science special project, research category: applied research; project name: application of machine learning algorithm in quantitative investment (Project No.: 23jk0638.)

Disclosure statement

The author declares no conflict of interest.

References

[1] Li B, Shao X, Li Y, 2019, Research on Fundamental Quantitative Investment Driven by Machine Learning. China Industrial Economics, 0(8): 61-79.

[2] Jiang B, Zhao S, Huang Y, 2023, Research on Machine Learning Applications in Financial Risk Prediction. Journal of Wuhan University of Technology (Information and Management Engineering Edition), 45(05): 785-789.

[3] Luo W, 2024, Application of Fama French Three Factor Model in China's Financial Stock Market. Green Finance and Accounting, 2024(11): 27-31.

[4] Tang G, Zhu L, Liao C, Jiang F, 2024, Research on Asset Pricing Based on Self coded Machine Learning: A Financial Big Data Analysis Perspective of China's Stock Market. Journal of Management Science, 27(9): 82-97.

[5] Gu L, 2023, Research on application of Machine Learning in portfolio. Industrial Innovation Research, 2023(05):127-129.

[6] Zhang J, Li Z, 2023, Whether a company's high quality development level can predict its stock market performance--based on machine learning. Research of Financial Economics, 38(06):50-65.