Machine learning-driven factor fitting model for stock data and its future trend prediction

Binghui Wang

Northeastern University at Qinhuangdao, Qinhuangdao, China

1847781952@qq.com

Abstract. Stock data factor model fitting and forecasting has been a hot topic in fintech and quantitative investment research. As a representative technology of artificial intelligence, machine learning can greatly improve the effect of forecasting research in economics and management. This paper aims to analyze stock data and predict fu-ture trends by using a variety of factor fitting models. First, we ex-tract multiple related factors from the data, such as volatility factor, growth factor, momentum factor, size factor, value factor, liquidity factor, profit factor, etc. Then, we use these factors to build fitting models, in this paper, eight machine learning algorithms, including linear model, Lasso regression, Ridge regression, decision tree model, random forest model, GBDT model and XGBT model, are used to build stock return prediction model and investment portfolio. Through these models, we can make pre-dictions about the stock data and come up with future trends. Final-ly, through empirical analysis, we verify that the forecast of these factor fitting models is better than that of CSI 300, and the annual-ized return rate and Sharpe ratio are both higher than that of CSI 300. Among them, the GDBT model has the best forecast results, with Sharpe ratio reaching 1.07 years and annualized return rate reaching 0.92, far higher than the 0.15 annualized return rate of CSI 300.

Keywords: Factor analysis, Stock picking strategy, Machine learning.

1 Introduction

With the increase of household income and assets available for investment, how to maintain and increase the value of assets is a problem that every family and individual will face. Among them, the stock market, as a high-risk and high-return investment mode, is very popular among families with investment assets. Moreover, the stock market is a barometer of economic development, which has an important impact on the national economy and the global market. Therefore, the research and analysis of the stock market is particularly important. In the stock market, price fluctuations are affected by a variety of factors, including the company's financial condition, macroeconomic factors, policies and regulations, market sentiment, and so on. The interaction between these factors makes the stock market forecast very complicated.

T. Ramayah et al. (eds.), *Proceedings of the 2024 International Conference on Applied Economics, Management Science and Social Development (AEMSS 2024)*,
Advances in Economics, Business and Management Research 284,
https://doi.org/10.2991/978-2-38476-257-6_45

In recent years, a large number of scholars have begun to explore the applica-tion of machine learning methods to stock model fitting prediction. The main research results are as follows: The problem of using machine learning to predict the trend of stock market. For example, Al-Ahmadi et Al. (2021) summarized the results of using traditional machine learning algorithms (such as linear regression, decision tree, support vector machine, etc.) in stock market prediction in their research[1]. They found that these algorithms can effectively use historical data to predict the trend of the stock market, and provide a certain reference value of the forecast results. The problem of using deep learning algorithm to predict the trend of stock market. Zhang et al. (2022) proposed a stock market prediction model based on deep learning algorithm. Through deep learning of a large number of historical data, the model automatically extracts the features that have influence on the stock price, and makes use of these features to forecast the stock price[2]. In addition, Chen et al. (2023) also proposed a stock market prediction model based on convolutional neural networks (CNN), which can effectively extract information related to stock prices from a large amount of unstructured data[3]. Use text mining and sentiment analysis to predict stock market movements. Ali et al. (2022) proposed a stock market prediction model based on text mining and sentiment analysis[5]. By analyzing the text in social media, the model extracts the emotional information related to the stock price and incorporates this information into the prediction of the stock price. How to extract the features related to stock price prediction from a large number of data. L. Chen, W. Zhang et al. (2023) proposed a feature extraction method based on machine learning[6], using a variety of different types of data (such as historical price data, news articles, social media posts, etc.) as input to extract features related to stock price movements. How to use machine learning to assess the risk of the stock market. R. Kaushik et al. (2021) conducted a comprehensive review and analysis of existing relevant studies[4], and discussed the application and effect of various machine learning algorithms (including linear regression, support vector machines, neural networks, etc.) in stock market risk assessment. They also discuss various factors that affect the accuracy of risk assessment, such as data quality, model selection, hyperparameter adjustment, etc.

In order to predict the trend of stock market better, this paper adopts a common method - factor fitting model. By fitting the historical data, the factor fitting model finds the main factors that affect the stock price and uses these factors to predict the future trend. This approach effectively reduces the dimensionality of the data and captures the main features that affect the stock price. Factor fitting model is also the main research method of stock data forecasting in this paper.

This paper aims to explore the application of multi-factor fitting models in stock data, and verify the validity and reliability of these models through empiri-cal analysis. By building different fitting models such as linear regression model, random forest model, we forecast and analyze the stock data and get the future trend. We will first introduce the relevant factor analysis of stock data, including volatility factor, growth factor, momentum factor, size factor, value factor, etc. Then, we will introduce the building process of factor fitting model in detail, including linear regression model, random forest model and so on. Then, we will verify the validity and reliabil-

ity of these models through empirical analysis. Finally, we will forecast and analyze the future trend.

2 Research design

2.1 Model overall design

Figure 1 shows the overall process of a quantitative stock selection model based on artificial intelligence. This paper first selects securities from the A-share market that are "normally traded" [Normal trading refers to securities that have been listed for more than 12 months and have not been subject to risk warnings or suspensions] to create the stock pool. Factors for modeling are selected using historical trading data and financial analysis data, ensuring the availability of sufficient information. The task of feature engineering is to transform an original set of factors into a combination that effectively captures the essence of the problem, thus enhancing the accuracy of stock selection when applied to predictive models. Ultimately, a variety of machine learning algorithms are utilized to predict stock data, filtering out high-yield stocks for backtesting of stock selection, and observing the strategy's return performance of the model.

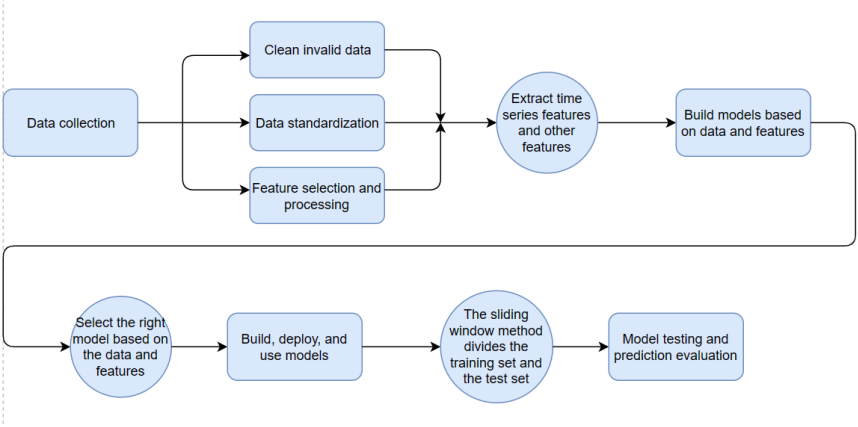

Fig. 1. Machine learning model data preprocessing, feature extraction, model building, model fitting process

In this paper, the sliding window method will be used to fit the parameters of the model by using the data before the decision point, so as to ensure the efficiency of calculation and the feasibility of investment. After determining the function form f (•), we will divide the training and test data set as shown in Figure 2.

The steps of model training and testing are as follows:

1. Assuming that it is currently in early January 2020, the model will fit the machine learning model and get the model parameters according to the anomaly factor-excess return data of the past 8 years (i.e. 2012 --2020) as a training set.

2. the trained model is used to predict the stock return in December 2020, based on the anomaly factor data in November 2020.

3. Based on the forecast results, sort the stocks and construct a long-short portfo-lio on the cross section, that is, long the 10% stocks with the best expected return and short the 10% stocks with the worst expected return, with equal weight allo-cation for long and short positions.

4. hold the portfolio for one month to get the portfolio return for December 2020.

5. the time reaches the beginning of February, repeat the above steps until the end of the data period.

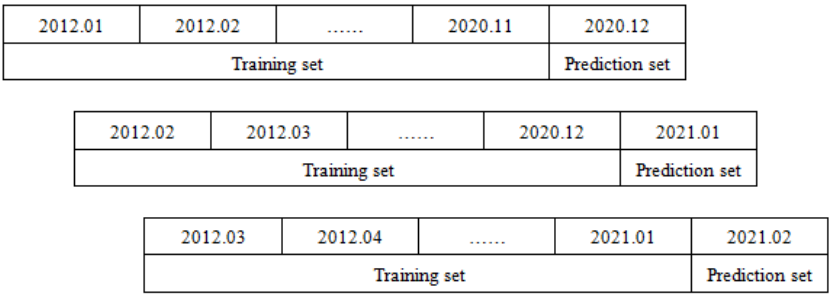

Fig. 2. Sliding window diagram

2.2 Data source and sample selection

Stock data from January 31, 2012 to December 30, 2022 are selected and processed on a monthly frequency. 16 corporate characteristic variables are selected to represent the anomaly factors, which are divided into seven categories according to their attributes: volatility factor, growth factor, momentum factor, scale factor, value factor, liquidity factor and profit factor. (1) Volatility factor includes historical volatility index; (2) Growth factors mainly focus on the growth potential of the company, including: main business income growth rate, net profit growth rate, internal growth rate; The momentum factor mainly focuses on the trend of the stock price, including the following indicators: the stock return over a period of time (such as the past week, one month, three months, six months or one year), the relative strength of the stock; (4) The scale factor reflects the size of the company, including the following indicators: total market value, circulation market value, free circulation market value; The value factor mainly focuses on the valuation level of the company, including the following indicators: dividend yield, price-to-book ratio, price-to-sales ratio, earnings per share to price ratio; The liquidity factor mainly focuses on the realization ability of assets, including the following indicators: trading volume, turnover rate; The earnings factor mainly focuses on the company's profitability, including the following indicators: return on equity (ROE), return on total assets (ROA), gross profit margin on sales, and net profit margin on sales.

Most of the earnings data is published on a quarterly basis, so this article us-es quarterly data and makes monthly padding. When populating data, we follow the principle of populating only after all the specified reports are available. All data are

obtained from the CSMAR database, where the factor and indicator data need to be downloaded from the corresponding data source and calculated using the appropriate calculation methods and formulas, as shown in Table 1. For example, the volatility factor can be calculated by calculating the historical volatility of the return on assets, the growth factor can be calculated by comparing the growth rate of the main business income, net profit and other indicators in the history of the company, the momentum factor can be calculated according to the stock return rate in the past period of time, and the size factor can be calculated by the total market value and current market value of the company. The value factor can be obtained by calculating the company's dividend yield, price-to-book ratio and other indicators, the liquidity factor can be obtained by analyzing the company's trading volume and turnover rate, and the profit factor can be obtained by calculating the company's return on equity, return on total assets and other indicators. In order to get a better understanding of the database, this paper uses the standard asset pricing method to test all the anomaly factors.

Table 1. What indicators are included in the seven categories of factors and how to obtain data for these indicators

Factor	Index	Calculation method
Volatility factor	Historical volatility	The standard deviation of calculated asset return or volatility
Growth Factor	Main Business	Growth Rate Compares the company's main business revenue for the past several quarters or years
	Net profit growth rate	Compare the company's net profit for the past several quarters or years
	Internal growth rate	Calculate the growth rate of the company's retained earnings and total assets
Momentum factor	Stock returns over a period of time	Calculate the return on the stock over a period of time
	Relative strength	Compare the performance of a stock over a period of time with the performance of an industry or market
Size factor	Total market value	The number of shares in the company multiplied by the stock price
	Circulating market value	The number of shares outstanding in a company multiplied by the stock price
	Free-float market value	Excluding shares held by controlling shareholders, the number of shares outstanding is multiplied by the stock price
Value factor	Dividend yield	The ratio of a company's dividend payments to its stock price
	Price/book ratio	The ratio of a company's market value to its book value (net assets)
	Market-to-sales ratio	The ratio of a company's market value to its

		sales revenue
	Earnings per share to price ratio	The ratio of a company's earnings per share to its stock price
Liquidity factor	Volume of transaction	The daily trading volume reflects the level of activity in the market
	Turnover rate	The ratio of volume to total issuance over a period of time reflects the liquidity of an asset
Profit factor	Return on equity (ROE)	The ratio of net profit to net assets of a company reflects the level of return on shareholders' equity

3 Empirical results and analysis

3.1 Yield curve

The construction of various machine learning algorithm models for backtesting experiments has yielded results for various metrics as shown in Table 2. Observation reveals that the Random Forest and GBDT models perform the best.

Table 2. The performance of each model on different financial indicators

Model	Annual return	Maximum pullback	Annualized fluctuation	Annual rate of return	Sharpe ratio
Linear model	12.85%	-47.28%	20.98%	0.27	0.52
Decision tree	6.84%	-55.07%	20.61%	0.12	0.24
Random forest	36.93%	-65.03%	37.12%	0.57	0.94
Ridge regression	14.88%	-46.30%	22.02%	0.32	0.58
Lasso	4.7%	-59.78%	24.07%	0.08	0.11
GBDT	46.97%	-50.85%	42.09%	0.92	1.07
XGBOOST	28.46%	-65.3%	41.38%	0.44	0.64

3.2 Model performance evaluation

The performance of the CSI 300 Index synchronized with the Shanghai and Shenzhen markets is shown in Table 3.

Table 3. Fitting results of CIS 300

Annual return	Maximum pullback	Annualized fluctuation	Annual rate of return	Sharpe ratio
6.09%	-40.56%	22.02%	0.15	0.19

From our data, the annualized returns, maximum pullback, annualized volatility, and Sharpe ratios are all different for each model.

These models generally show some differences. Among them, the linear model performs well when fitting the data, with relatively low annualized returns and maximum pullback, and a high Sharpe ratio. The fit data for the decision tree and Lasso models performed poorly, with lower annualized returns, maximum pullback, and Sharpe ratios. The Random Forest and GBDT models perform well in terms of annualized returns and Sharpe ratios, but the maximum pullback are also large. The performance of ridge regression in prediction is between linear model and random forest. The XGBT model has a higher annualized return and Sharpe ratio, but also a larger maximum pullback.

Compared with the CSI 300, the annualized returns and Sharpe ratios of all models are higher than the CSI 300. This could mean that these machine learning models are better at predicting the stock market than the CSI 300. However, it is important to note that the return volatility of machine learning models (such as maximum pullback) may also be high.

Overall, GBDTis the best performer among all models, with the highest annualized return, maximum pullback, and Sharpe ratio. This means that the model best fits the data and gives the best predictions.

4 Conclusion

By collecting 16 anomaly factors in the A-share market from January 31, 2012 to December 30, 2022, and constructing 7 fitting models driven by machine learning algorithms, this paper systematically compares these machine learning models with the CSI 300 model predicts the results of the Chinese stock market. The empirical results show that the performance of linear machine learning algorithm is better than that of single factor and linear regression model, while the overall performance of machine learning algorithm is better than that of Hushen 300. The prediction of these factor fitting models is better than that of Hushen 300, and the annualized return rate and Sharpe ratio are higher than that of Hushen 300, among which GDBT model achieves the best effect on this problem. The Sharp ratio reached 1.07 years, with an annualized return of 0.4697, much higher than the 0.0609 annualized return of the CSI 300. This shows that there are nonlinear patterns between the factors of each algorithm that are difficult to be recognized by traditional linear regression, and the machine learning algorithm can automatically recognize these patterns, so as to obtain better prediction effects and combinatorial returns. After the optimality test, we found that the machine learning-driven GBDT model showed excellent performance under various conditions, including maximum fluctuation, multiple factors, and different indicators. Different from the traditional single factor test, from the perspective of machine learning, we can find new important factors, such as liquidity factors represented by turnover rate and trading volume, which have strong predictive ability for stock cross-sectional returns. At the same time, compared with the CSI 300 invest-

ment strategy, the machine learning-driven model performs better in processing important factor data.

References

1. "Predicting Stock Market Movement Using Machine Learning: A Review" by S. Al-Ahmadi, H. T. Lam, and K. Mao, published in Expert Systems with Applications, 2021.
2. "Machine Learning for Stock Market Forecasting: A Deep Learning Approach" by J. Zhang, Y. Li, and H. Liu, published in Neural Computing and Applications, 2022.
3. "Using Machine Learning to Extract Stock Price Predictive Features from Big Data" by L. Chen, W. Zhang, and J. Wang, published in IEEE Transactions on Knowledge and Data Engineering, 2023.
4. "Machine Learning for Risk Assessment in the Stock Market: A Review" by R. Kaushik and J. F. Dantzig, published in European Journal of Operational Research, 2021.
5. "Using Machine Learning to Model Stock Market Sentiment: A Text Mining Approach" by S. M. Ali, A. H. Johnson, and T. Zhang, published in Journal of Business Economics and Management, 2022.
6. "Application of Machine Learning in Portfolio Management: Improving Returns and Reducing Risk" by Y. Chen and L. Xu, published in Journal of Portfolio Management, 2023.