Factor-based Stock Selection and Portfolio Construction Utilizing Machine Learning Methods

Yingli Hu

Business School, East China University of Political Science and Law, Shanghai, 200333, China
210529010236@ecupl.edu.cn

Abstract. Quantitative investment becomes the future direction of the financial market. More machine learning tools are used for stock price forecasting. Therefore, this paper combines traditional multi-factor stock selection models with machine learning. It filters fundamental and technical factors to construct a new factor-based stock selection model. It predicts stock price trends using LightGBM and Random Forest models, with parameter optimization performed using Bayesian optimization. In constituents of the S&P500 index, stocks with investment value are identified according to data over the past three years, and two effective investment portfolios are constructed. The study finds that in terms of prediction, LightGBM is faster in computation, but it is less accurate in trend forecasting compared to Random Forest. Random Fores exhibits a lag in predicting sudden changes in stock prices. Portfolios constructed using both machine learning models can generate excess returns, with the portfolio built using Random Forest offering higher returns and risks, making it suitable for more aggressive investors. LightGBM, on the other hand, provides better risk management. This study proposes some ideas and approaches for investors to predict stock prices, providing individual investors with a convenient method for forecasting.

Keywords: Stock Price Forecasting, Factors, Machine Learning, Portfolio Construction.

1 Introduction

The changes in the global financial environment are becoming increasingly intense. Achieving accurate prediction of stock prices has a significant impact on achieving investors' financial goals, and individuals can also choose or adjust their investment directions based on trend information. However, accurately predicting the trend of stock prices isn't an easy task. In reality, the stock market is often influenced by complex factors. The inherent noise environment and the characteristics of large fluctuations in the market make stock prediction face many challenges and limitations [1]. With continuous iteration, improvement of algorithms and the enhancement of hardware computing power, quantitative investment has become a new trend. The key to quantitative investment is transforming complex financial theories into specific trading rules. By establishing mathematical models and using statistical analysis tools, investors can conduct quantitative analysis of the market, discover hidden patterns and trends in the data, and apply these analysis results to trading decisions to

J. Lu (ed.), *Proceedings of the International Workshop on Navigating the Digital Business Frontier for Sustainable Financial Innovation (ICDEBA 2024)*, Advances in Economics, Business and Management Research 315,
https://doi.org/10.2991/978-94-6463-652-9_92

formulate appropriate strategies to obtain stable investment returns. In terms of return forecasting, the emergence of machine learning has provided new tools. An increasing number of investors, especially institutional investors, are using various models to try to predict stock prices.

Factor models offer a statistical description of returns' cross-sectional dependence structure, and thus are used in modeling equity returns [2]. Early financial theory such as Sharpe's CAPM established exposure to systematic and market risk as significant drivers of returns. Both size and value factors were added by Fama and French [3]. They proposed the three-factor model. In 2015, the five-factor model was proposed, which includes market the premium factor, book-to-market ratio factor, size factor, profitability factor, and investment factor [4]. Based on Sharpe's research, Morningstar popularized "style box investing", including value and size factors [5]. When selecting factors, the main reliance is on economic logic and market experience. Common stock selection factors include volatility factors, analyst factors, growth factors, profit factors, and momentum reversal factors, etc. [6].

The emergence of statistical and machine learning tools has provided data-driven solutions for asset pricing and return forecasting. Machine learning emphasizes variable selection and dimensionality reduction techniques, which are highly suitable for prediction by reducing the degrees of freedom and compressing the redundant variations between predictors [1]. Gu et al. illustrated that nonlinear methods including random forest, boosted regression trees and deep learning have substantial gains in estimating expected returns [7]. Many researchers employed vast machine learning models to predict the returns of stocks. Huang applied the XGBoost model to stock price prediction. It was proved that the prediction accuracy of XGBoost was significantly higher than using neural networks and support vector machines [8]. Li used 96 factors with a Support Vector Machine (SVM) model, demonstrating that machine learning algorithms can effectively identify anomaly factors, and can achieve a monthly return of 3.41% based on factor screening [9].

In summary, most studies use machine learning in the prediction of stocks' future returns based on previous price trends. Some studies combine the factors of the five-factor model with machine learning models for prediction. However, there are some limitations. For example, due to the large number of candidate factors, which can be in the hundreds, there may be correlations between the factors, and the interactions between them are not considered [10]. In terms of the selection of objects, most studies focus on predicting the stock prices of specific individual stocks and specific industries. Compared with previous studies, this study focuses on a market index, using the index's constituent stocks as the stock pool. It has a broader investment scope than previous studies and is committed to constructing an investment portfolio to achieve excess returns over the index. In terms of factor selection, unlike brokerage research reports, this study selects factors that have been proven by relevant theories to explain returns. It aims to measure the linear and nonlinear relationships between factors and stock excess returns by combining machine learning models and to identify factors that can better explain returns.

2 Method

2.1 Dataset Preparation

This article selects the S&P 500 index's constituent stocks as the stock pool. The current constituent stocks list is obtained from Wikipedia. The volume data and adjusted closing price for each constituent stock from January 1, 2021, to July 31, 2024, are imported from the yfinance package. The company's characteristic indicators are obtained from the ifind Excel plugin. The compound growth rate of the adjusted closing price is used as the return rate. The closing price can't reflect stocks' full value, as things like cash dividends are omitted. Therefore, the adjusted closing price is used to reflect the stocks' real value.

In terms of factor selection, the preliminarily selected factor library includes 22 factors, consisting of 8 fundamental factors e.g. Price Earnings Ratio and 14 technical factors e.g. 10-day Moving Average. In addition, some other preprocessing approaches are employed in this study as shown below.

Handling Missing Values: If the missing values of a stock exceed 30%, then it is removed. For other samples, missing values in the dataset are forward-filled to ensure continuity in the time series data. Ultimately, data for 80 stocks and 22 factors are obtained.

Normalization: Z-score normalization is used to unify the scale of data, preventing some factors from dominating the analysis due to their inherent larger magnitude.

Data Splitting: The training and testing set is divided, the training set includes data from January 2021 to December 2023. The testing set consists of data from January 2024 to July 2024. Models are trained using the training set, and the testing set simulates real-world conditions for model validation.

2.2 Factor Testing

When selecting factors, it is necessary to choose those that have a strong explanatory power for stock returns. Therefore, a univariate effectiveness analysis should be conducted for each factor to select better factors from the preliminary selection for use in machine learning models. Common methods include scoring methods, correlation analysis, and regression methods. This article screens factors through regression and correlation analysis.

T-test Method. Linear regression is performed, the factor value is the independent variable x and the subsequent period's return is the dependent variable y. A T-test is conducted on the regression coefficient to obtain the significance level of the regression coefficient as the t-value. The effectiveness of a single factor is judged using the t-value.

Correlation Analysis. The information coefficient (IC) is calculated using the following equation.

$$IC^{i,k} = \mathrm{corr}(x_{k,t}^{n}, r_{i,t+1}) \quad (1)$$

X,k,t represents the k-th factor on the cross-section at time t. $r_{i,t+1}$ represents the compound growth rate of stock i from time t to time t+1.

A high degree of absolute IC proves the greater factor's predictive association with future return, and the factor's explanatory power is greater. In addition, the inter-factor correlation indices are established. Factors belonging to the same type often have strong correlations, and the factor with the best explanatory power is selected from the same type of factors in combination with the t-value.

Multivariate Analysis. A multiple linear regression is conducted on the remaining factors using logarithmic returns, and factors with insignificant regression coefficients are eliminated.

2.3 Machine Learning Model-based Stock Return Rate Prediction

Light GBM. To effectively classify stocks into 'long' or 'short' categories, LightGBM is used. It is based on historical stock data and selected features. Light GBM is an algorithmic framework within Gradient Boosting Decision Tree (GBDT). Its core concept involves using multiple decision trees as weak classifiers and optimizing the model's explanatory power through iterative training to arrive at the final model. When dealing with large datasets, it can divide the data into mini-batches for individual training, effectively avoiding the drawback of traditional GBDT. Additionally, Light GBM helps prevent model overfitting and offers fast training speeds.

In terms of parameter optimization, A Bayesian optimization technique is employed using BayesSearchCV to select the most effective combination of hyperparameters for the LightGBM classifier. The search space included parameters such as maximum number of leaves, learning rate, feature fraction and regularization terms (lambda_l1 and lambda_l2). With the optimized parameters, the LightGBM model is retrained on the full training dataset. The models were configured with a binary objective function (binary_logloss), appropriate for the classification of stock returns.

Random Forest. Random Forest constitutes an ensemble technique that utilizes multiple decision trees. It enhances predictive accuracy and stability by averaging multiple decision trees' results. It can reduce overfitting. Additionally, it is not sensitive to noise and outliers in the training data, making it highly stable.

Hyperparameter tuning is performed using Bayesian optimization, a more efficient method than traditional grid search as it identifies near-optimal parameter combinations with fewer iterations. After optimal parameters are determined, models undergo training using the full training dataset. In order to maximize their learning capacity.

3 Results and Discussion

3.1 Factor Testing

Factor Scoring. The rolling window of 60 days was used to calculate the average of the IC series. The IC value is divided by the tracking error ratio to obtain the Information Ratio (IR). After calculating the IC, IR, correlation coefficient between factor and lagged one-period return shown in Table 1, and the T-value obtained from the T-test of the return regression on a single factor for 22 factors, a comprehensive scoring is conducted for the factors. Those with higher scores are selected to construct the model. The correlation coefficient and IR are divided into 5 quantile intervals, with scores ranging from 1 to 5 for each interval from low to high. If the t-value's absolute value is larger than 1.96, the regression coefficient of the single factor on the return is significant at the level of 5%, and this factor can score 1 point. Otherwise, it scores 0 points. The final score of the factor is obtained by adding the three scores together.

Table 1. Partial Factor Scoring Table

Factor	IR Percentile	IR Metric Score	T-value Score	Correlation Percentile	Correlation Metric Score	Total Factor Score
market cap	0.90	5	1	0.86	5	11
PS	0.86	5	1	0.81	5	11
momentum	0.95	5	1	0.95	5	11
RSI (7-Day)	1.00	5	1	1.00	5	11
PB	0.81	5	1	0.76	4	10
PE	0.76	4	1	0.71	4	9

Correlation Analysis. Factors of the same type may contain identical information. Including them in the same regression model may lead to collinearity issues. The following Fig. 1 displays a heatmap of factor correlations. The red area means a strong inter-factor positive correlation. When selecting factors, they are first ranked from high to low based on the total score, and factors with the same type or an absolute correlation greater than 0.6 with the already selected factors are eliminated to ensure the best explanatory effect of the retained factors. Initially, 9 factors are screened out.

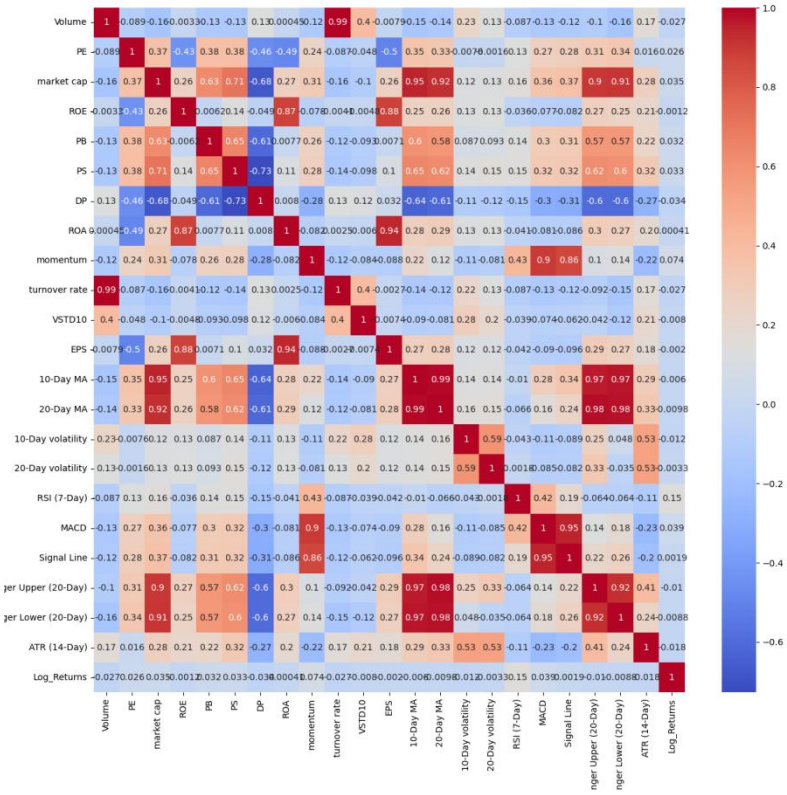

Fig. 1. All factors' correlation heatmap (Photo/Picture credit: Original).

Multivariate Analysis. In a regression model established using the least squares method, two insignificant variables are eliminated, and the logarithmic returns regression equation on seven factors proves to be significant. Therefore, the final set of selected factors is {'market_cap', 'momentum', 'RSI_7_Day', 'PE', 'ten_Day_MA', 'MACD', 'ATR_14_Day'}.

3.2 The Performance of the LightGBM Model

Through Bayesian optimization, the optimal parameters obtained for the LightGBM model are a maximum number of leaves 20, learning rate= 0.03, feature fraction= 0.9, minimum data per leaves 20, and minimum gain to split= 0.005. The best number of iterations is determined during the training process through an early stopping mechanism to prevent overfitting.

The model's accuracy for each stock is evaluated by contrasting the predicted long/short positions with the actual outcomes. Stocks are classified as 'long' if their

predicted probability exceeds a threshold (0.5), and 'short' otherwise. The top 30 stocks with the highest accuracy are selected to form the portfolio shown in Fig. 2, with the weight of each stock being proportional to its accuracy rate out of the total accuracy. The annualized return of the portfolio is 25.16%, and the maximum drawdown is -3.32%. Its Sharpe ratio is 3.00. The annualized return is 10.4% higher than the S&P 500 index's for 2024, which is 14.7%.

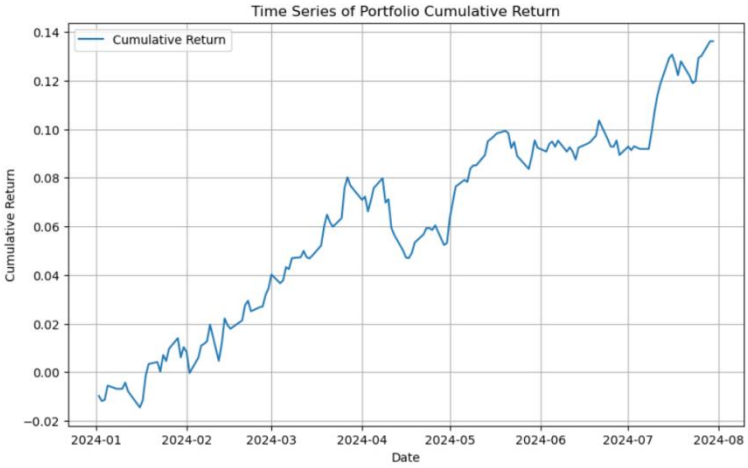

Fig. 2. Portfolio's cumulative return constructed by LightGBM model (Photo/Picture credit: Original).

3.3 The Performance of the Random Forest

In the Random Forest model, 10% of the training set's data is randomly selected for parameter tuning. Hyperparameter tuning is performed using Bayesian optimization. The optimal parameters obtained are maximum depth=11, minimum samples per leaf=8, minimum samples to split=12, and number of estimators=68. The model's evaluation metrics are Mean Absolute Error (MAE)=2.3375, Mean Squared Error (MSE)=22.0694, Root Mean Squared Error (RMSE)=4.6978, and R-squared (R2)=0.9993. Taking NVDA as an example, a comparison chart of actual and predicted stock prices is plotted shown in Fig. 3, showing that the predicted trend is basically consistent with the actual trend, and the prediction is relatively accurate.

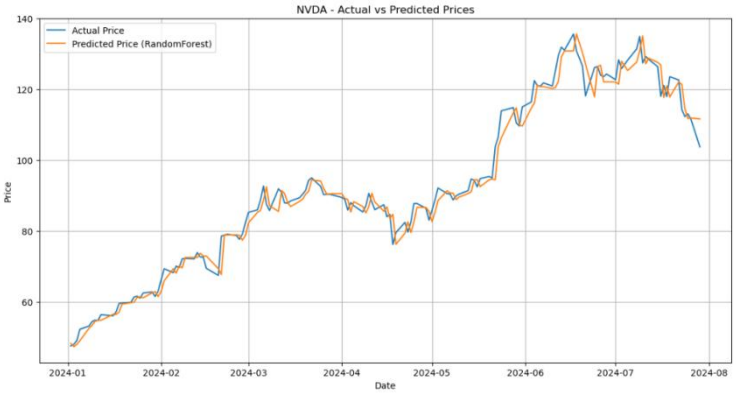

Fig. 3. Predicted price of NVDA (Photo/Picture credit: Original).

Stocks are sorted according to the predicted Sharpe ratio shown in Fig. 4, and the top 30 stocks with the highest Sharpe ratio are selected to form the portfolio, with weights allocated according to the Sharpe ratio. This portfolio's annualized return is 62.29%, and its maximum drawdown is -6.04%. Its Sharpe ratio is 4.36.

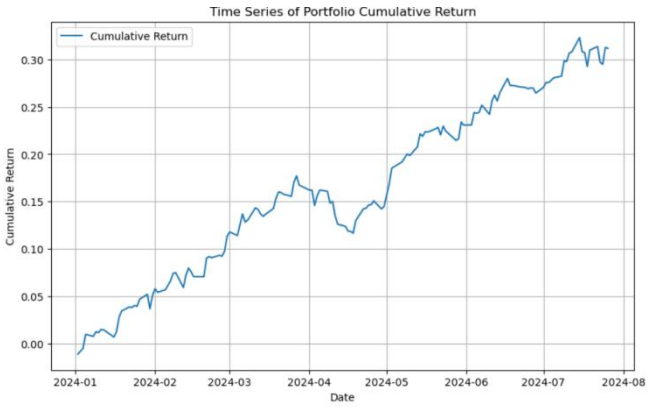

Fig. 4. Portfolio's cumulative return constructed by Random Forest model (Photo/Picture credit: Original).

In terms of prediction, the Random Forest model is more accurate in forecasting stock price trends. It has a certain degree of lag in capturing sparse mutation information, and the predicted price curve fluctuates less than the actual price. In comparison, LightGBM's highest accuracy for predicting the future upward or downward trend of individual stocks is 0.6, which may be due to data imbalance. Additionally, LightGBM requires more parameters but fits and predicts faster because Random Forest relies on deep trees, consuming a significant amount of computational resources. Portfolios constructed by both models have generated annualized returns exceeding the index. The portfolio constructed using the Random Forest model has a higher annualized return, but also a larger maximum drawdown, making it more suitable for investors who prioritize return maximization. LightGBM, on the other

hand, offers better risk management with slightly higher returns, appealing to more conservative investors.

4 Conclusion

In terms of factor selection, this study integrates both fundamental and technical aspects to build a factor pool, taking into account as many factors as possible that influence stock prices. Factors are screened through scoring and ranking for use in machine learning model predictions, providing investors with a simple and convenient method for investment and forecasting. In terms of machine learning, this paper determines the parameters of the LightGBM and Random Forest models through grid search and Bayesian optimization, respectively, to select the best possible parameters for research. This article may have certain limitations in stock selection based on the selected factors. Using appropriate factors and assigning different weights to different factors may increase the effectiveness of the model. The parameter selection of machine learning models requires optimization research. This paper uses grid search to optimize the parameters of LightGBM, with an accuracy rate of around 0.53, which is a direction that needs further study.

References

1. Vachhani H, Obiadat M S, Thakkar A, et al.: Machine learning based stock market analysis: A short survey. In: Innovative Data Communication Technologies and Application: ICIDCA 2019, pp. 12-26. Springer, Heidelberg (2016).
2. Bender, J. B., Melas, R., et al.: Foundations of Factor Investing. Publisher, Location (2013).
3. Fama, E. F., French, K. R.: A five-factor asset pricing model. Journal of Financial Economics 116(1), 1–22 (2015).
4. Fama, E. F., French, K. R.: Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33(1), 3–56 (1993).
5. Giglio, S., Kelly, B. T., et al.: Factor Models, Machine Learning, and Asset Pricing. Publisher, Location (2021).
6. Fuertes, A. M., Miffre, J., Fernandez-Perez, A.: Commodity strategies based on momentum, term structure, and idiosyncratic volatility. Journal of Futures Markets 35(3), 274-297 (2015).
7. Gu, S., Kelly, B., Xiu, D.: Empirical asset pricing via machine learning. Review of Financial Studies 33(5), 2223–2273 (2020).
8. Huang, Q., Xie, H. L.: Application research of machine learning methods in stock index futures forecasting - Based on a comparative analysis of BP neural networks, SVM and XGBoost (in Chinese). Mathematics in Practice and Theory 48(8), 297-307 (2018).
9. Li, B., Shao, X. Y., Li, Y. Y.: Research on fundamental quantitative investment driven by machine learning. China Industrial Economics, (08), 61-79. DOI: 10.19581 (2019).
10. Bali, T. G., Engle, R. F., Murray, S.: Empirical Asset Pricing: The Cross Section of Stock Returns (in Chinese). John Wiley & Sons, Hoboken, New Jersey (2016).