Stock Price Return Prediction Based on Multifactorial Machine Learning Approaches

Xingtong Wang[1], Wen Wang[2(✉)], and Shuya Zhang[3]

[1] School of Accounting, Southwestern University of Finance and Economics, Chengdu, China
[2] School of Finance, Southwestern University of Finance and Economics, Chengdu, China
41928018@smail.swufe.edu.cn
[3] School of Finance, Dongbei University of Finance and Economics, Dalian, China

Abstract. Contemporarily, the combination of artificial intelligence and financial theory is a hot topic. In this paper, the multifactorial machine learning models for stock price prediction are implemented and compared after screening the effective factors. Specifically, four different linear models (OLS regression, Lasso regression, Ridge regression, Elastic Network regression) and nonlinear model XGBoost are applied. Based on the analysis, nonlinear model has better performance than different linear models, and the expected return rate constructed by this method has a higher correlation with the real rate of return. In terms of factor selection, this paper refers to the classification and construction of factors in relevant literature, including basic information factor, volume price factor, valuation factor, financial statement factor. In terms of data selection, the daily data from October 2018 to October 2021 among the four indexes with different industries, scales and market sentiment are selected to prevent extreme situations when using a single index. These results shed light on some extent that machine learning combined with quantitative investment has certain application value.

Keywords: Multifactorial Prediction · Linear Regression · Machine Learning Model

1 Introduction

Traditional quantitative investment methods are unable to obtain excess returns from the constantly updated financial market. In order to solve this problem effectively, more and more researchers choose the optimal quantitative investment method from the perspective of artificial intelligence through the effectiveness analysis of multi-factor model. Big data and artificial intelligence, as a new data analysis and prediction tool combined with the current financial background, have efficient processing ability for financial data in the field of quantitative investment. In this context, by combining the application of artificial intelligence in quantitative investment direction, this paper compares OLS linear regression method, Lasso regression, ridge regression, elastic network regression and XGBoost regression with MSE, R^2, and other empirical research methods, trying to

X. Wang, W. Wang and S. Zhang—Contributed equally.

D. Qiu et al. (Eds.): ICBBEM 2022, AHIS 5, pp. 324–333, 2023.
https://doi.org/10.2991/978-94-6463-030-5_34

find the optimal method to predict stock return rate. In terms of data selection, this paper mainly investigates the four underlying stocks in a-share market with different industries, different company sizes, different market sentiment and different core themes, so as to avoid the limitations of the model.

With regard to factor selection, this paper refers to the construction and testing of factor classification in relevant literature, including basic information factor, volume price factor, valuation factor, financial statement factor, etc. Firstly, the correlation analysis of 22 factors in these four categories was carried out to determine the low correlation among factors. Secondly, the validity of these factors is tested to screen those factors whose values have a good correlation with stock returns. Finally, the validity factor combined with Lasso regression, ridge regression, elastic network regression and other machine learning methods are used to simulate robust stock returns, and the optimal model is selected for comparative analysis.

Stock price prediction is one of the most common research topics in financial economics. The way to make a reasonable estimate of the expected return of stock in the uncertain financial market is the core of this topic. Among the theoretical models in the field of asset pricing that include stock price forecasts, the capital Asset Pricing Model (CAPM), developed by William Sharpe et al., in the modern portfolio selection theory of Harry Max Markowitz, undoubtedly plays an important role. This model describes the linear relationship between a single risk factor and return on assets [1, 5, 7]. With the continuous development of research on asset pricing theory, the academic circle has gradually found that in addition to a single risk factor, the return on asset is also affected by the company's market value and book-to-market ratio. Combined with these new research findings, Fama and French proposed a three-factor model combining risk factor, scale factor (SMB) and value factor (HML), which has a certain improvement in the explanatory power of excess return on assets compared with CAPM model [2].

With the deepening of the research on asset pricing, the academic circle has gradually discovered more factors affecting the return on asset. Zura Kakushadze presented 101 effective alpha factors in his research on the factor construction model of Alpha101, including content price factor and basic information factor [12]. In the quantitative stock selection analysis of multi-factor model, Xu Jingzhao tested the effectiveness of candidate factors and obtained value factors including earnings ratio and price-to-book ratio, growth factor including ROE growth rate, quality factor including asset-liability ratio, and momentum factor including monthly average turnover rate [4]. In their research on China's stock market reform, Li Zhibing et al. found that the efficiency of capital market was significantly improved after share reform, among which the scale effect and book-to-market ratio effect were significant [10]. With the increasing application of multiple factors in the return on assets, the academic circle is gradually trying to introduce machine learning methods to the study of asset pricing, so as to give play to its advantages in multi-feature input and prediction. In recent years, Gu applied a variety of machine learning models to the American stock market based on multiple factors, and Rapach, Strauss and Zhou used Lasso regression to predict global stock prices, all of which showed good prediction effects [6, 8]. This paper will further study the prediction effect of multi-factor machine learning model in Chinese stock market.

In this paper, OLS regression, Ridge regression, Lasso regression and elastic network regression are selected as models to predict stock returns. OLS is the most basic linear regression model. Hoerl and Kennard proposed ridge regression based on OLS models with regular terms of L2 norm in order to improve the collinearity problem and improve the generalization ability of the model [3]. Later, Tibshiran proposed Lasso regression with regular terms of L1 norm, which also improved the generalization ability of OLS model [9]. In 2005, elastic network regression was proposed by Zou and Hastie as a combination of Ridge regression and Lasso regression [11].

The rest part of the paper is organized as follows. The Sect. 2 will introduce data origination and factor selection methods as well as regression models. The Sect. 3 will demonstrate the correlation analysis and regression results for different models while the Sect. 4 present the discussion. Eventually, a brief summary is given in Sect. 5.

2 Data and Method

2.1 Data

To make a comprehensive study of A-share market, the sample targets are selected among different industries among the A-share market, which is imported from Wind. A total of four bids were selected for data extraction and research in the factor correlation assessment, namely: (i) COSCO Shipping (601919), in the Industrial Marine industry, (ii) CATL (300750), in the Industrial Power Equipment industry, (iii) PharmaBlock Science (300725), in the Pharmaceutical and Biological industry, and (iv) Yili (600877), in the Food and Beverage industry. The daily data of the companies from September 2018 to September 2021 were selected as the base data for further factor screening. The Figs. 1, 2, 3 and 4 illustrate the k lines of PharmaBlock Science, CATL, Yili and COSCO Shipping, respectively.

2.2 Methodology

According to the alpha 101, the factors are selected and divided into basic information factor, volume price factor, valuation factor, and financial statement factor. The specific classification of factors is shown as Table 1.

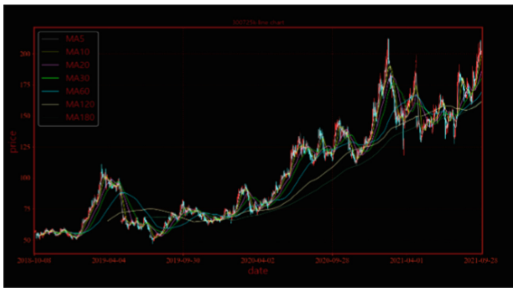

Fig. 1. K-line chart of Pharmablock science.

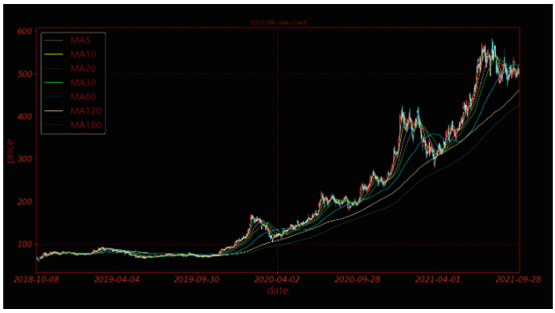

Fig. 2. K-line chart of CATL

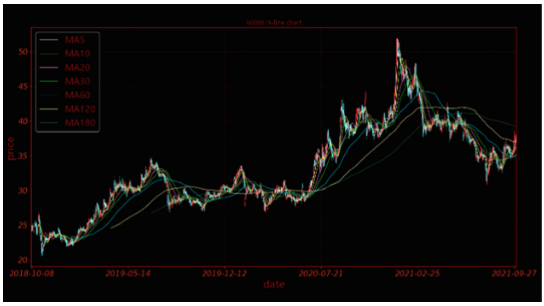

Fig. 3. K-line chart of Yili.

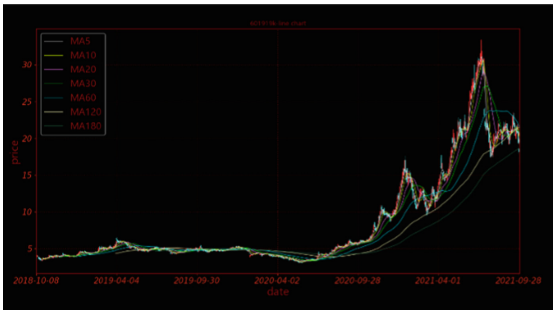

Fig. 4. K-line chart of COSCO Shipping.

In the initial screening, the basic information factor and the cross-sectional data earnings factor, which are more difficult to quantify, were excluded, while the daily data valuation factor and the volume factor were retained for further validity testing. Table 2 lists the final selected factors.

Factors with a correlation of more than 60% and those with a correlation of less than 5% were eliminated to obtain the effective factors. Then, the effective factors were used to construct the multi-factor model, combining OLS regression, Lasso regression, ridge regression, elastic network regression, and XGBoost. During the training process, 70%

Table 1. Initial classification of factors

basic information factor	days listed, version listed, days of incorporation of the company, the and industry category, ST status, whether it is part of the SSE 50 Index, whether it is part of the CSI 300 Index
volume price factor	Total market value, P/E ratio, P/N ratio, P/S ratio
valuation factor	current day's opening price, previous day's opening price, current day's closing price, current day's high price, current day's low price, current day's volume and current day's turnover
financial statement factor	Current ratio, total assets, net assets, gearing ratio, earnings per share

of the selected data is taken as the training set and 30% as the verification set. The data is substituted into the model. R^2 and MSE is then calculated on the verifier.

a) OLS linear regression: OLS linear regression is the most common and basic integrated analysis model. For example, assume that certain indicators of the underlying stock have a linear relationship with future returns, and let the linear equation be $y = b + w_1x_1 + w_2x_2 + w_3x_3 + ... + w_nx_n$. Then the indicator is chosen as "x_1, x_2, x_3...x_n", which is the factor x_1 to x_n, and the expected return is y. The purpose of this regression is to minimize the sum of squares of the straight distance from each point to the optimal fitting curve.
b) Ridge regression: Ridge regression is a linear regression method for fitting. L2 norm regularization term is added on the basis of general linear regression to enhance the generalization ability of the model while ensuring the best fitting error. The formula is: $y = b+w_1x_1+w_2x_2+w_3x_3+....+w_nx_n+b$. In ridge regression, the factors are chosen not only to give good predicted returns on the data, but also to fit additional constraints. This means that the effect of each feature on the output should be as small as possible (i.e., the slope is small) and this constraint is also known as the regularization process. The ridge regression function with L2 regularization added is: $J = \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - y_i\right)^2 + \lambda||\omega||_2^2$
c) Lasso regression: By constructing a penalty function, Lasso regression can compress the coefficients of variables and make some regression coefficients become 0, so as to achieve the purpose of variable selection. The Lasso regression function with L1 regularization is: $J = \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - y_i\right)^2 + \lambda||\omega||_1$
d) Elastic network regression: Elastic network regression is based on the above analysis, combining ridge regression and Lasso regression, and adding L1 and L2 norm regularization terms to ensure that the objective function has a unique optimal solution. The objective function of elastic network regression is: $\min(||y - X\beta||) + \lambda_1|\beta|_2^2 + \lambda_2|\beta|_1$
e) XGBoost: XGBoost adds L1 and L2 regulars on the basis of GBDT error function. The advantage of adding regular terms is to prevent overfitting.

Table 2. Final Selected Factors

y3	Three Day Yield
y5	Five Day Yield
y10	Ten Day Yield
x1	Closing price minus opening price 5 days ago
x2	Highest price minus lowest price
x3	First order differential of the highest price
x4	First order differential of the lowest price
x5	Three-day average closing price minus five-day average closing price
x6	Five-day average closing price minus ten-day average closing price
x7	Ten-day average closing price minus twenty-day average closing price
x8	Price-to-sales ratio
x9	P/B ratio
x10	Average of three-day traded shares minus average of five-day traded shares
x11	Average of five-day traded shares minus average of ten-day traded shares
x12	Average of ten-day traded shares minus average of twenty-day traded shares
x13	Closing price minus average price
x14	First order difference in market sales rate
x15	First order difference of turnover rate
x16	First order difference of P/E ratio
x17	First order difference in market rate
x18	The change of value divide the three-day average of the high price minus the low price
x19	The change of value divide the five-day average of the high price minus the low price
x20	The change of value divide the ten-day average of the high price minus the low price
x21	the change of value* the turnover rate
x22	First order difference of the highest price times the lowest

3 Results

3.1 Results of Factor Selection

In order to sift effective factors, the correlation between different factors and rate of return was tested. Figure 5 presents the correlation heatmap of factors based on data from PharmaBlock Sciences. According to the results of correlation analysis, we select effective factors of each company, as shown in Table 3.

Table 3. Effective factors of each company.

CATL	x1, x5, x6, x10, x12, x13, x18, x19
PharmaBlock Sciences	x1, x2, x3, x4, x5, x6, x10, x11, x12, x14, x18, x19, x20, x21, x22
Yili	x1, x2, x5, x6, x12, x18, x19, x20
COSCO	x1, x5, x6, x11, x15, x17, x18, x19, x20

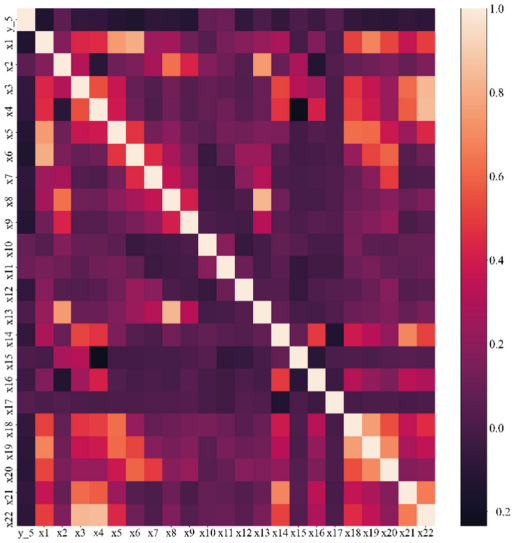

Fig. 5. Correlation heatmap of PharmaBlock Sciences.

Table 4. Modeling results of CATL.

	R^2	MSE	Correlation with rate of return
OLS	−0.0729	309.6876	−0.0165
Ridge	−0.0398	300.1376	−0.0123
Lasso	−0.0458	301.8754	−0.0189
ElasticNet	−0.0398	300.1388	−0.0123
XGBoost	0.0561	272.4742	0.2595

3.2 Results of Modeling

Tables 4, 5, 6 and 7 illustrate R^2, MSE and correlation with rate of return of different models based on data from four companies. It is apparent that XGBoost model has best performance in terms of all four companies, which possesses the highest R^2, highest correlation and lowest MSE.

Table 5. Modeling results of PharmaBlock Sciences.

	R^2	MSE	Correlation with rate of return
OLS	−0.0729	309.6876	−0.0165
Ridge	−0.0398	300.1376	−0.0123
Lasso	−0.0458	301.8754	−0.0189
ElasticNet	−0.0398	300.1388	−0.0123
XGBoost	0.0561	272.4742	0.2595

Table 6. Modeling results of Yili.

	R^2	MSE	Correlation with rate of return
OLS	0.0234	14560.5093	0.1680
Ridge	0.0203	14591.3369	0.1899
Lasso	0.0113	14724.7728	0.1698
ElasticNet	0.0174	14901.2923	0.1865
XGBoost	0.0587	14019.5641	0.2815

Table 7. Modeling results of COSCO.

	R^2	MSE	Correlation with rate of return
OLS	0.0118	1.4891	0.1803
Ridge	0.0207	1.4758	0.1466
Lasso	0.0171	1.4812	0.1379
ElasticNet	0.0202	1.4765	0.1437
XGBoost	0.0613	1.4146	0.2831

4 Discussion

4.1 Comparison Between Models

Little was found in the literature on the application of machine learning in the A-share market. The results of this study show performance of different machine learning models on A-share market.

As can be seen in results section, OLS regression, the most basic linear model, shows the poorest performance in all four cases. In terms of other linear models including ridge regression, LASSO regression and elastic net regression, results indicate better performance than OLS regression. We believe that this improvement can be explained by the fact that regularization items are added to all three linear models which show

better performance. The existence of regularizer help prevent overfitting thus increase generalization ability of model.

XGBoost model, on the other hand, outperforms all four linear models. As shown in Table 7, R^2 of XGBoost is even one order of magnitude larger than that of OLS regression. The most intuitive explanation for this significant difference may be the nonlinearity of XGBoost model. This finding is in consistent with most previous studies on application of machine learning in other stock markets.

In general, regularizers help improve performance of linear models and nonlinear model shows better performance than linear models. Consequently, although contrary to the classic CAPM model which uses linear relationship to calculate the asset pricing, it seems that nonlinear relationship can be a better choice when it comes to predicting stock returns.

4.2 Limitations

A number of caveats need to be noted regarding the present study. First, only 8 to 9 effective factors are chosen for three out of four companies in our study. More effective factors may be out of consideration. In terms of models used above, we only choose one nonlinear model (XGBoost) to predict stock returns. As a consequence, whether all nonlinear models outperform linear models remains a problem. In addition, another limitation of this study is that the sample size was relatively small. Since only four listed companies and marketing data from a three-year period are concerned in the study, caution must be applied, as the findings might not be transferable to the whole A-share market.

5 Conclusion

In summary, return prediction based on multi-factor model is investigated in terms of linear model as well as machine learning method. Primarily, the regression results of the non-linear model XGBoost are significantly better than the linear regression method. Meanwhile, the consideration of the relationship between the factors and the returns is not limited to univariate regressions, as multiple regressions in a non-single coordinate space have better fit results and smaller mean square error coefficients. Secondly, from the regression results of the four linear models, the three linear regression models incorporating regularization generally outperform the OLS regression models in that they enhance the generalization of the model while ensuring the best fit error. Ridge regression, Lasso regression and elastic network regression also complement and optimize the OLS linear models. The inclusion of the regularization term can better prevent overfitting problems and improve the effectiveness of solving practical problems through machine learning. In terms of the comparison of the results of these three linear regression models, the degree of fit of Ridge regression and elastic network regression tends to be consistent, and both are slightly better than Lasso regression. In brief, the non-linear model has some advantages over other different types of linear regression models in terms of constructing multi-factor models to predict returns, and achieves better results.

For further study, more nonlinear models should be put into consideration. More samples with more samples can be selected to strengthen the conclusion. The research aims to select the better regression models for investment decision selection. These results provide an alternative analytical tool for data analysis in economics and finance as well as offer a practical value in machine learning applications.

References

1. Black, Fischer. "Capital Market Equilibrium with Restricted Borrowing." The Journal of Business 45 (1972): 444–455.
2. Fama, Eugene F. and Kenneth R. French. "Multifactor Explanations of Asset Pricing Anomalies." Journal of Finance 51 (1996): 55–84.
3. Hoerl, Arthur E. and Robert W. Kennard. "Ridge Regression: Applications to Nonorthogonal Problems." Technometrics 12 (1970): 69–82.
4. Jingzhao Xu. Quantitative stock selection analysis based on multi-factor model[J]. Financial Theory exploration, 2017(03):30–38.
5. Lintner, John. "THE VALUATION OF RISK ASSETS AND THE SELECTION OF RISKY INVESTMENTS IN STOCK PORTFOLIOS AND CAPITAL BUDGETS." The Review of Economics and Statistics 47 (1965): 13–37.
6. Rapach, David E., Jack Strauss and Guofu Zhou. "International Stock Return Predictability: What is the Role of the United States?" FEN: Other International Corporate Finance (Topic) (2012): n. pag.
7. Sharpe, W.F. (1964), CAPITAL ASSET PRICES: A THEORY OF MARKET EQUILIBRIUM UNDER CONDITIONS OF RISK*. The Journal of Finance, 19: 425–442.
8. Shihao Gu, Bryan Kelly, Dacheng Xiu, Empirical Asset Pricing via Machine Learning, The Review of Financial Studies, Volume 33, Issue 5, May 2020, Pages 2223–2273.
9. Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." Journal of the royal statistical society series b-methodological 58 (1996): 267–288.
10. Zhibing Li, Guangyi Yang, Yongchang Feng, Liang Jing. An empirical test of Fama-French five-factor model in Chinese stock market [J]. Financial research, 2017(06):191–206.
11. Zou, Hui and Trevor J. Hastie. "Regularization and variable selection via the elastic net." Journal of The Royal Statistical Society Series B-statistical Methodology 67 (2005): 301–320.
12. Kakushadze, Zurab. Massive Gravity in Extra Dimensions [J]. Acta Physica Polonica B, 2014.