Multi-factor Stock Forecasting Based on Regression and Machine Learning Models

Jiawen Liu*

Department of Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China

*Corresponding author: p930026084@mail.uic.edu.cn

Abstract. In some stock predictions, shallow learning is better than deep learning. With this in mind, this study will adopt a machine learning model and a regression model to predict Chinese stocks to help the stock timing strategy. Different from the previous usage of plenty of indicators as stock features for prediction, this study selected specified multi-factor data and compared the prediction effects of foreign multi-factor data models and China's multi-factor data models to verify. It is found that the prediction effect of the Chinese new salience multi-factor data model which has a better ability to explain market anomalies is better. However, factor models with higher correlations have poor results. Compared with randomly selecting factors or selected factors weight by model, choosing a specific factor model in advance can not only improve calculation efficiency, but also improve prediction accuracy. In addition, compared with the regression model, the prediction results of machine learning also have better performance results.

Keywords: Machine learning; time series forecasting; multi-factor.

1. Introduction

Discussed from the machine learning perspective, in economic research, theory-driven reasoning and data-driven analysis are always complementary and indispensable parts. Machine learning provides a strong analytical technique that allows researchers to distinctly obtain the content of the data and obtain calculation results that cannot be achieved manually. When faced with big data, supervised machine learning technology in machine learning can select a more flexible function form based on the data, has good out-of-sample prediction capabilities, and at the same time avoids the problem of overfitting. Moreover, on the issue of causal inference that is of greatest concern to empirical research in economics, machine learning can use covariates to obtain the nonlinear relationship between them [1]. In the relationship between machine learning and asset pricing, machine learning takes into account the interaction of nonlinear predictors. Tree models and neural networks perform better than regression models, especially in the prediction of large-capitalization stocks and high-liquidity stocks. Shallow learning models such as neural network models perform better than deep learning models. Therefore, in the training model selected in this article, this study will use common regression models and machine learning models.

The factor stock selection model is one of the most broadly adopted stock selection models in quantitative investment strategies. Alternative factors usually need to capture economic information. Securities with the same factor have the same behavioral path, are more differentiated in different markets and samples, and can perform stably in the time dimension.

One of the widely used empirical asset pricing methods is the multi-factor model in academic research. Its primary objective is to identify a set of factors that can effectively interpret the cross-sectional variations in stock forecast returns. By constructing long-short portfolios based on fundamental characteristics, and price indicators, one can evaluate the performance of these portfolios using the multi-factor model for asset pricing. If the return of a portfolio cannot be adequately explained by this model, it indicates an anomaly [2]. Over several decades of rigorous academic exploration, numerous anomalies have been discovered across different stock markets. The abundance of anomalies can be attributed to two main reasons: firstly, setting a low t-statistic threshold for anomaly detection may result in more false positives; secondly, the choice of pricing

models has a significant impact on the number and significance level of identified anomalies. It is believed that selecting an appropriate and effective pricing model will help explain more anomalies accurately. For instance, raising the t-statistic threshold to 3.0 as suggested by Harvey et al., renders many previously identified anomalies insignificant; furthermore, employing Hou et al 4-factor model as a pricing framework diminishes' significance while only retaining a small portion [3, 4]. However, it should be noted that most prediction models heavily rely on numerous factors or adjust their weights accordingly. Alternatively, some models solely focus on time-series predictions solely on stock prices themselves which may lead to overfitting issues and introduce redundant factors with excessive correlation into consideration unnecessarily. In reality, though not all factors possess sufficient explanatory power for predicting returns across all stocks. Thus, they fail to provide accurate predictions for other stocks. Therefore, rather than allowing the model to select an effective model among many factors or determine the weights between different factors, it may be possible to directly select a pricing model of effective factors that can better explain the anomalies to explain the expected return of the asset which is a more efficient and accurate method.

2. Methodology

2.1. Data

This article mainly studies the multi-factor model of China's stock market. This study will use China A-share market stock data from December 2009 to December 2020 as the measurement data in this article. the stock data source is the Resset database. Collected stock ticker, date in months, closing price, and yield. When experienced COVID-19, one measured stock data on China's A-share market before and after the epidemic, and found that a Chinese four-factor model containing emergent factors was very interpretable, which explained 33 significant anomalies in China [5]. In addition, this study will select the traditional Fama-French three-factor model, and the Carhart four-factor which adds the momentum effect, based on the Fama-French three factor model and the PMO contained in the Chinese four factor model (CH4) predicts stock data under regression and machine learning models to contrast the effectiveness of several multi-factor models in stock prediction [6-8]. Among them, the relevant multi-factor data all come from the database of central University of Finance and Economics. This study will study and compare the stock prediction results based on domestic and foreign multi-factor data models. After obtaining the stock data and factor model data, in order to make the calculations more convenient, this study merged the two data together.

2.2. Data preprocessing

There will be issues with incomplete and mismatched packets in the collected data. Since there are obvious differences in the magnitude and distribution of the values of different factors, it may lead to prediction bias, causing features with larger magnitudes to dominate in prediction or slowing down the iterative convergence of machine learning. Therefore, this study will Standardize the data. In this experiment, one set the value between -1 and 1. In addition, this study chooses to use the mean to fill in the blank data. Finally, this study will delete the data that has a lot of missing information, such as some stocks that have been listed for more than 16 years data.

2.3. Model Description

After 1964, The Capital asset pricing model (CAPM) was defined and developed by Sharpe, Lintner, and Black believed that stock returns only have a linear relationship with the systemic risk of the entire stock market [9-13]. However, in 1993, Fama and French introduced a three factor model to account for stock returns. This model suggests that the stock excess return can be explained by market asset portfolio Rm−rf, market capitalization factor SMB, and book-to-market ratio factor HML. This multi-factor model can be expressed as

$$R_{it} - r_{ft} = a_i + b_i[(R_{mt} - r_{ft})] + c_i(smb_t) + d_i(hml_t) + \varepsilon_{it} \quad (1)$$

The following variables are used in the analysis: rft denotes the risk-free rate of return at a specific time t, Rit represents the return market rate at the same time t, and Rit represents the rate of return of a specific asset i also at time t. The market risk premium is represented by Rit-rft, and smbt is the market value or size at time t. The simulated portfolio return for the Small minus Big factor is represented by hml, while the High minus Low factor is represented by hmlt. The coefficients of the three factors are represented by b, $c_i$and d_i.

The anomalies of efficient markets were explained through the binomial model based on arbitrage pricing theory (APT) and the three factor model supposed by Fama-French. To build a four-factor model, Carhart added the one-year return momentum anomaly factor to the Fama and French three-factor model [7]. The formula for this model is as follows:

$$R_{it} - r_{ft} = a_i + b_i(mkt_t) + c_i(smb_i) + d_i(hml_t) + e_i(umd_t) + \varepsilon_{it} \quad (2)$$

On the basis of FF-3, the alternative China three factor model data based on China's A-share market was calculated, and added the fourth factor PMO (pessimistic minus optimistic), they used abnormal turnover. It means they attempted to use the past month's share turnover divided by the past year's turnover, building the turnover factor with the same method to calculate the value factor, again neutralizing with respect to size. This allows the model to explain more anomalies[8]. The model formula is such as Eq. (3):

$$R_{it} - r_{ft} = a_i + b_i(mkt_t) + c_i(smb_i) + d_i(hml_t) + e_i(umd_t) + f_i(pmo_t) + \varepsilon_{it} \quad (3)$$

Sun et al. created a new four-factor model, which replaces the PMO factor in CH4 to PMOR (prominent-minus-ordinary returns) and PMOV (prominent-minus-ordinary trading volumes) which based on Chinese stock salient return and stock salient trading volume respectively [5]. This new four-factor model can explain more significant anomalies than the CH4 model. This study uses the new four-factor which contains PMOR. The formula is as follows:

$$R_{it} - r_{ft} = a_i + b_i(mkt_t) + c_i(smb_i) + d_i(hml_t) + e_i(umd_t) + f_i(pmov_t) + \varepsilon_{it} \quad (4)$$

The stock return prediction model employs the OLS model, which uses the least squares regression method to determine the best function match for the data. The aim is to minimize the sum of squared errors, thereby obtaining unknown data and reducing the difference between obtained and actual data. The OLS model formula is utilized in this process.

$$h_\theta(x) = \theta^T \cdot x + b = \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots + \theta_3 \cdot x_3 + b \quad (5)$$

When one compares the regression model with the multi-factor model formula, one can find that when using the value of return risk-free return as the prediction target, the excess return rate as the intercept, and the factor corresponds to the x value in the formula, the unknown factor coefficient one obtains is θ in this formula.

The process of gradient boosting involves iteratively reducing loss by fitting the last residual. To achieve this, the gradient descent of the loss function is used as the approximate value of the residual for regression problems in the lifting tree algorithm. The loss function in this case employs square loss. Unlike the OLS model that fits the original value, gradient boosting fits the error value between the predicted value and the actural value, continuously adjusting it using the gradient descent method.

Random forest is an algorithm that leverages ensemble learning to combine multiple trees result. The decision tree is the fundamental unit of this algorithm, which falls under the major branch of machine learning called Ensemble Learning. The algorithm randomly selects samples and features to build each decision tree, increasing the model's diversity and robustness. The final regression result is obtained by averaging or taking the weighted average of multiple decision trees' prediction results.

The Recurrent Neural Network (RNN) is one kind of neural network model that excels in handling sequential data. Unlike traditional neural networks that have a direct connection from the input layer collecting the variable and feature value to the hidden layer, after then to the output layer where can get the outcome, an RNN stores and remembers previous information in its hidden layer. This allows the RNN to input the previous hidden layer's output into the current hidden layer unit. In other words,

the nodes in the hidden layer are not only independent, but also convey the information to each other. Consequently, the input of the hidden layer can be broken down into three parts: the output of the input layer and the hidden layer from the past moment, and the state of the previously hidden layer. The RNN model's structure is illustrated in Fig. 1.

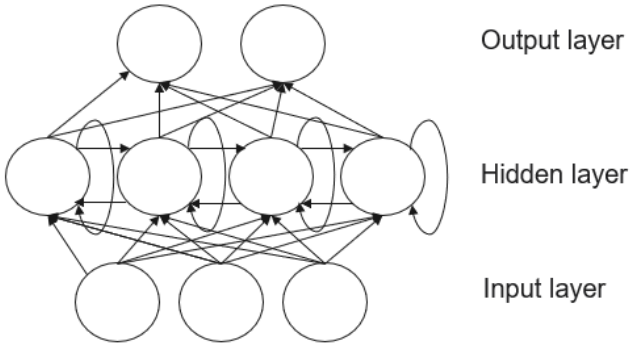

Fig. 1 RNN model structure

Although RNN have a memory function, they may not be able to solve long-term problems effectively. When the prediction point is too far from the dependent information, understanding the relevant content can become difficult. Additionally, during the training process, the model may face problems with gradient explosion or disappearance, which can hinder effective training and learning. However, Long Short Term Memory Network (LSTM) is a specialized type of RNN that can handle long-term dependency problems and gradient issues well. Compared to RNN, LSTM has a memory block that consists of a forgetting gate (rt), an input gate (it), and an output gate (ot), and a memory cell that selectively decides whether to pass on information to the next moment. The gate control unit determines the importance of information and decides whether to store or delete it. The operating principle of LSTM is shown in the figure below, which illustrates the structure of the LSTM model (seen from Fig. 2).

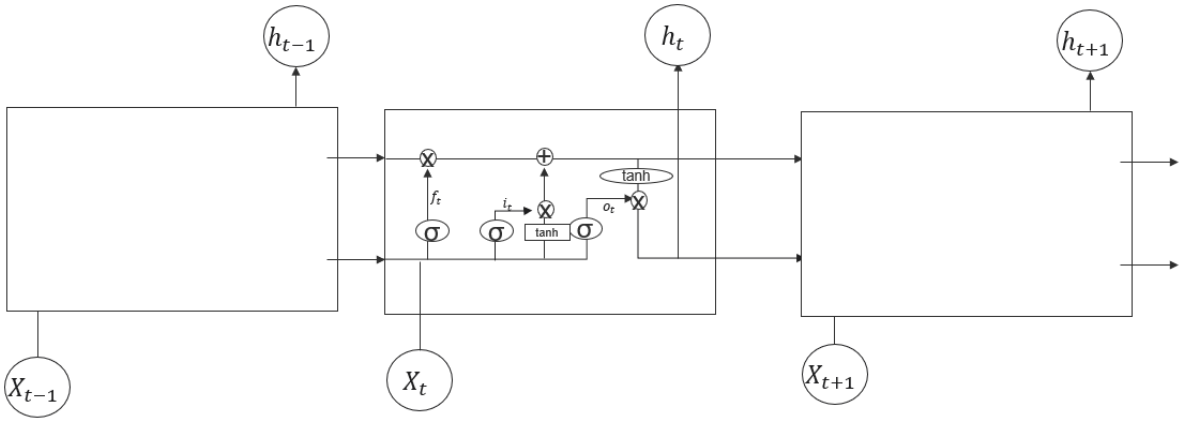

Fig. 2 LSTM model structure.

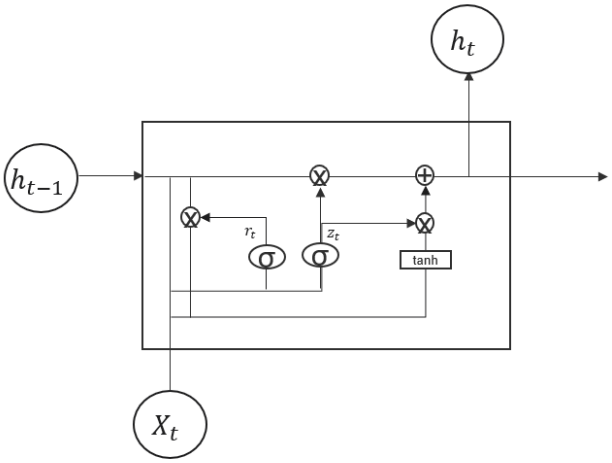

Fig. 3 GRU model structure.

The Gated Recurrent Unit (GRU) is a type of RNN that was introduced by Cho et al. in 2014. It is an improvement on the LSTM network, which is a popular model in the field of deeper machine learning. The GRU integrates the input gate and forget gate into a "update gate", while also making modifications to the cell state and hidden state. The main difference between the GRU and LSTM is the way the gating mechanism is designed. The GRU uses only one update gate, which controls the flow of information and updates its status while reducing the number of parameters and calculations required. Additionally, the GRU model is more concise and easier to train than LSTM, and has achieved good results in some tasks. The simple graph structure of the GRU is shown in Fig. 3.

3. Results and Discussion

First, this study conducts a correlation analysis between factors in the multi-factor model. One can find that in the Carhart four-factor model, except for the two factors smb and hml, which have a high correlation of about -0.5, the correlations between other factors are low. Among the four Chinese factors, the correlation between vmg and smb is very high, around 0.73. Compared with other factors, the new highlighted four factors also have higher values in the correlation between smbr and vmgr. This study will divide the data into train data and test data, with a ratio of 1:2. In the machine learning model, this study will continue to divide the training data into the train set and validation set, the proportion of validation is 0.2. Meanwhile, one set the Epoch number of all machine learning models to 100, the batch size to 32, and the loss function to MSE. also. the optimizer will be set to Adam. The prediction results of all stock data are calculated. This study first selects one of the stocks to show the prediction results of different models on the test data of different factor models. Then, one calculates the RMS of the forecast data for all stocks. The result is illustrated in the Table 1.

Table 1. RMSE result of different model and muti-factor

Test	OLS	Gradient boosting	Random forest	RNN	LSTM	GRU
Fama-French	0.080395	0.093132	0.097328	0.111565	0.093281	0.095962
Carhart	0.071138	0.083278	0.092378	0.075643	0.074069	0.070181
CH4	0.113891	0.129329	0.113289	0.123561	0.133748	0.140814
Salience factor	0.070363	0.081267	0.079326	0.074150	0.057634	0.065261

4. Conclusion

Based on the previous related tests, it is found that the newly highlighted four factors performed better than other factor models in China's A-share market stocks. At the same time, it was found that

the effect of China's four factors was the worst among all results. It is speculated that there may be factors with high correlation that affect the effect of the model. In the training model, most machine learning models perform better than simple linear models in predicting results.

This article only uses four multi-factor models for comparison, and only uses common time series machine learning models for prediction. Although the effect of deep learning models on time series is not particularly advantageous so far, deep learning models can capture more implicit feature relationships. In the future, you can try to select appropriate deep learning models for prediction, as well as other multi-factor models. Compare.

Using specific multi-factor model data helps us improve the efficiency of calculations and also increases the accuracy of predictions. At the same time, it was found that better prediction results can be obtained when the new prominent four factors are selected to have more explanations for market anomalies. On the premise of having better prediction effects, the multi-factor model and machine learning model can help us select stocks and gain profits.

References

[1] Huang N, Yu M. Research progress on the impact of machine learning on economics research. Economic Perspectives, 2018, (7):115-129.

[2] Lee C M C. Technological links and predictable returns. Journal of Financial Economics, 2019,132(3): 76-96.

[3] Harvey C R, Liu Y, Zhu H et al. The cross-section of expected returns. The Review of Financial Studies, 2016, 29(1): 5-68.

[4] Hou K, Chen X, Lu Z. Digesting anomalies: An investment approach. The Review of Financial Studies, 2015, 28(3): 650-705.

[5] Sun K, Zhu Y. Salience Theory Based Factors in China. SSRN 4342607, 2023.

[6] Fama E F, Kenneth R F. Common risk factors in the returns on stocks and bonds. Journal of financial economics, 1993, 33(1):3-56.

[7] Carhart M M. On persistence in mutual fund performance. The Journal of finance, 1997, 52(1): 57-82.

[8] Liu J, Robert F S, Yu Y. Size and value in China. Journal of Financial Economics, 2019, 134(1): 48-69.

[9] Sharpe W F. Capital asset prices: A theory of market equilibrium under conditions of risk. The journal of finance, 1964,19(3): 425-442.

[10] Lintner J. Security prices, risk, and maximal gains from diversification. The journal of finance, 1965, 20(4): 587-615.

[11] Black F, Michael C J, Myron S. The capital asset pricing model: Some empirical tests, 1972.

[12] Chai T F, Roland R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development, 2014,7(3): 1247-1250.

[13] Chung J Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.