Research on Machine Learning Driven Stock Selection Strategy

Keran Wang(✉)

Faculty of Information Technology, Beijing University of Technology, Beijing, China
Keran_Wang@emails.bjut.edu.cn

Abstract. As a representative technique of artificial intelligence, machine learning could explore the relationship between stock market anomalies and excess returns, and hence develop investment strategies with high performance. This paper provides a comparative analysis of machine learning algorithm applications in the field of quantitative stock selection through detailed and solid empirical evidence. Based on 36 anomalies in the Chinese stock market from January 2011 to March 2022, this paper adopts random forest regression for feature selection and nine machine learning algorithms, including Lasso regression, Ridge regression, Elastic Net regression, SVM, GDBT, XGBoost, LGBM, and neural network, to construct stock return prediction models and portfolios. The empirical results show that the machine learning algorithms can effectively assist in the formulation of quantitative investment strategies in the A-share market, and the long-short portfolio predicted based on the LGBM algorithm can obtain the highest annualized return of 69.33%. This study further examines the importance of the A-share market anomaly factor and finds that the momentum factor and the trading-friction factor have strong predictive power on the excess returns of the A-share market. It also integrates anomaly factors used in academia and industry, and further tests the effectiveness of machine learning algorithms in investment management problems, providing a reference for academic research and practical operation of asset management.

Keywords: stock selection · machine learning · anomaly factors · Chinese stock market

1 Introduction

As the Artificial Intelligence boosted in recent years, intelligent quantitative investing has become a popular investment approach in the financial industry, where the basic problem is to find factors that have the ability to predict excess returns (i.e., "anomalies"). However, with more than 300 verified anomalies by far [1, 2], it is important to consider how to make selection from this large collection to find the suitable one for target stock market [3]. As for the Chinese Stock Market ("A-share Market"), its special characteristics determine the differences between A-share market anomalies and other capacity markets represented by US stock market [4]. For researchers of quantitative investment, it is necessary to find the suitable anomaly portfolio specific to the A-share market [5].

K. Hemachandran et al. (Eds.): ICAID 2023, AHIS 9, pp. 151–159, 2023.
https://doi.org/10.2991/978-94-6463-222-4_14

The stock excess returns can be predicted using a variety of approaches based on the anomaly set. However, Traditional asset pricing methods are not able to adequately consider the function of each anomaly and the potential interactions between them when the number of anomalies is large. As a representative technique of artificial intelligence, machine learning (ML) algorithms are well adapted to the characteristics of big data in financial markets [6, 7], learning reproducible patterns from large amounts of data and making predictions accordingly [8, 9].

In this paper, 36 company fundamental characteristic anomalies commonly used in academia and industry are selected to construct an anomaly pool [10]. Then, Machine learning algorithms are used to select a portfolio from it. An excess return prediction model is constructed, its predictions are used to perform monthly frequency trading in the simulation phase.

The main contributions of this paper are: a) A comparative analysis of the application of machine learning algorithms in the field of Chinese quantitative stock selection is made through detailed empirical evidence, covering variable preprocessing, hyperparameter tuning and algorithm selection. b) Based on the existing asset pricing theory, this paper organically integrates machine learning methods, validates the existing theories, integrates the anomalies that frequently used in financial industry, and conducts systematic tests to make the results more convenient for practical use. c) This paper selects factors that hold strong correlation with excess returns of Chinese stock market, which makes the factor selection interpretable and indicative for future academic research and practical investment.

2 Methodology

2.1 Research Scheme

Figure 1 illustrates the overall framework of the research design of the essay. In the "Stock Pool", all securities in the A-share market from January 2011 to March 2022 are selected.

The "Asset Pricing" module firstly uses the machine learning model to select the optimal portfolio of anomalous factors, and then uses a variety of machine learning algorithms to integrate the anomalies to forecast stock returns; the "Trading Strategy" model constructs investment portfolios based on the excess forecast; the "Trading Simulation" conducts simulation trading according to the strategy given by the trading strategy module.

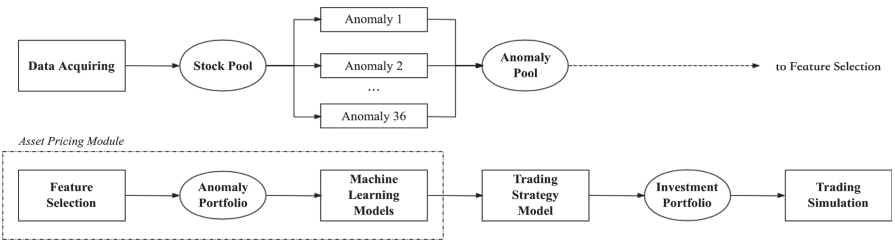

Fig. 1. The overall design of the research.

The asset pricing module has two main tasks, a) "Feature Selection", i.e., constructing the optimal portfolio of anomalies by selecting the anomalies that are significantly correlated with excess returns and reducing the feature dimension through the feature selection algorithm in machine learning. b) "Excess Return Prediction", i.e., acquiring the relationship between the anomaly portfolio and excess returns in the past and using the present data to forecast stock excess. The task of return prediction is a standard supervised learning task, the goal of which is to find functional form (1):

$$E_{t,i} = f\left(x_{t-1,i},\ k\right) + e_{t,i} \tag{1}$$

f is defined as a function with parameter k, represents for several machine learning and deep learning methods in the form of functions. $E_{t,i}$ is the excess return of stock i of period t. $x_{t-1,i} = (x_{t-1,i,1},\ x_{t-1,i,},\ \ldots, x_{t-1,i,N})$ is the array of anomalies of company i of period $t-1$. $e_{t,i}$ is the random perturbation term. After determining a specific functional form f, the parameters of the model will be decided in this paper by fitting, using data before the decision point in time. To ensure the validity of the calculation and the feasibility of the investment, the sliding window method shown in Fig. 2 will be used to divide the training and testing datasets.

2.2 Data

In this paper, all companies in A-share market from January 2011 to March 2022 are selected as the research sample, and the data frequency is chosen as monthly frequency in order to be consistent with related studies.

Referring to the anomalies in existing studies, 36 firm characteristic anomalies were selected in this paper. According to the factor attributes, they are divided into seven categories: value, profit, growth, liquidity, capitalization, momentum and trade-fraction. Data are obtained from Wind database.

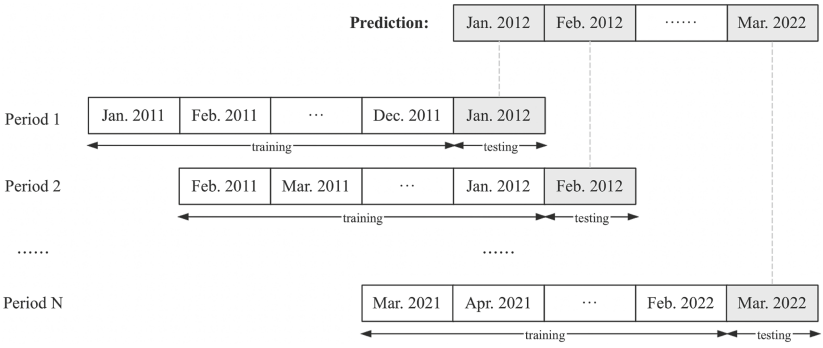

Fig. 2. The "sliding window" method.

3 Empirical Analysis

3.1 Single Machine Learning Algorithm Performance

This paper examines the empirical performance of machine learning algorithms in the A-share market. Table 1 shows the risk-return scenarios of nine machine learning methods for a 12-month sliding window long portfolio. As a comparison, the performance of the market index (HS300 Index) is presented as well. All measurements are annualized.

Table 1 shows that: a)The annualized returns of all portfolios with machine learning algorithms exceed the market index returns, showing the effectiveness of machine learning algorithms in factor investment research in the A-share market; b)Nonlinear machine learning algorithms can achieve better performance than linear machine learning algorithms in general, among which, the LGBM has the most significant performance improvement, with annualized return achieves 37.24% and Sharpe ratio reaches 1.10, making improvement of 45.81% and 23.60% compared with OLS regression respectively. The nonlinear machine learning algorithm SVM and XGBoost are the second most effective, with annualized returns improving by 43.43% and 38.88%, and Sharpe ratios improving by 34.83% and 22.47%, respectively, compared with OLS regression.

Table 2 shows the risk-return scenarios of nine machine learning methods for a 12-month sliding window short portfolio. Observing Table 3, it can be found that the investment model developed in this paper is also able to achieve good results when trading short. Nonlinear algorithms GBDT and LGBM achieve best performance with an annualized return around 17.13%, followed by the Neural Network.

Table 3 shows the risk-return scenarios of nine machine learning methods for a 12-month sliding window long-short portfolio. Observing Table 3, it can be found that the investment model have best performance trading long-short portfolio. LGBM and XGBoost reaches annualized rate of returns higher than 60% and Sharpe ratio of

Table 1. Investment Performance – Long Portfolio

	Rate of Return (%)	**Volatility (%)**	**Maximum Drawdown (%)**	**Sharpe Ratio**
OLS	25.54	32.55	37.08	0.89
Lasso	27.15	33.11	36.22	0.86
Ridge	28.53	32.43	34.81	0.98
Elastic Net	32.95	35.96	42.96	1.03
SVM	36.63	32.38	32.81	1.20
GBDT	31.11	33.64	41.01	1.03
XGBoost	35.47	35.94	33.54	1.09
LGBM	37.24	37.66	29.38	1.10
NN	30.94	32.32	31.15	1.05
MKT	4.79	40.56	40.56	0.21

Table 2. Investment Performance – Short Portfolio

	Rate of Return (%)	**Volatility (%)**	**Maximum Drawdown (%)**	**Sharpe Ratio**
OLS	12.93	12.60	66.23	0.70
Lasso	9.12	35.28	66.61	0.39
Ridge	12.63	35.20	67.27	0.50
Elastic Net	−4.33	35.49	74.30	−0.03
SVM	10.13	38.93	77.77	0.43
GBDT	17.15	36.78	73.95	0.62
XGBoost	12.89	35.11	74.01	0.50
LGBM	17.13	34.49	68.79	0.63
NN	14.39	34.08	72.41	0.56
MKT	4.79	22.61	40.56	0.21

Table 3. Investment Performance – Long-Short Portfolio

	Rate of Return (%)	**Volatility (%)**	**Maximum Drawdown (%)**	**Sharpe Ratio**
OLS	39.42	28.99	39.10	1.39
Lasso	35.54	29.47	44.20	1.26
Ridge	39.42	28.99	39.10	1.39
Elastic Net	29.84	36.35	58.37	0.96
SVM	51.75	34.68	40.10	1.51
GBDT	54.53	28.80	34.09	1.82
XGBoost	60.27	25.89	23.18	2.15
LGBM	69.33	29.38	17.46	2.15
NN	40.60	27.97	34.69	1.46
MKT	4.79	22.61	40.56	0.21

2.15. Followed by other nonlinear algorithms likewise GDBT (54.53%, 1.82) and SVM (51.75%, 1.51).

3.2 Integrated Machine Learning Algorithm Performance

To illustrate the effectiveness of machine learning algorithms for anomaly investment in the A-share market, this paper integrates several machine learning algorithms and constructs an investment portfolio based on the integrated algorithm forecasts. Specifically,

Table 4. Investment Performance – Integrated

	Rate of Return (%)	Volatility (%)	Maximum Drawdown (%)	Sharpe Ratio
Long Integrated	37.67	35.12	29.65	1.16
Long Mean	31.75	33.96	35.44	1.03
Short Integrated	19.45	33.99	70.73	0.70
Short Mean	11.33	33.11	71.26	0.32
L-S Integrated	63.09	32.77	30.42	1.84
L-S Mean	46.74	32.57	41.21	1.59

the integrated prediction is the arithmetic average of several machine learning algorithms' predictions. The results show that the integrated machine learning algorithm can achieve better return and lower risk than a single machine learning algorithm, as shown in Table 4.

4 Important Anomalies of Chinese Stock Market

The specificity of the A-share market determines the specificity of the anomaly factors associated with the excess returns of A-share stocks. This paper examines the factors influencing Chinese stock returns from a machine learning perspective to identify the set of anomaly factors with the strongest predictive power in the A-share market. In this paper, for each monthly sliding window of anomalies data, the contribution scores to stock excess returns of each anomaly are calculated by random forest regression. The top 20 anomalies are selected as important factors by calculating the average of each anomaly factor for all months and ranking in descending order, as shown in Table 5.

Observing Table 5, it is found that among the 20 selected anomalies, 6 momentum-based factors are included and all of them are listed in the top 8, indicating that momentum-based factors have excellent predictive ability for excess returns in the A-share market. Trading friction factors are the second most important factors, including 7 factors, indicating that trading friction factors represented by turnover rate and volatility of turnover rate can also predict excess returns in A-share market with good results. Meanwhile, it is noted that among all the 7 types of anomalies, momentum and trading friction factors are selected as significant pairs with 100% and 78% respectively, which proves that the relatively stronger forecasting ability of momentum and trading friction factors is not due to the large proportion of trading friction factors. The growth factor and value factor occupy three and two important factors respectively, indicating that they also have certain predictive ability for A-share market excess returns.

Table 5. Important Anomalies in A-share Market

Rank	Characteristic	Contribution (Avg.)	Type
1	*mom_1m*	4.4388	Momentum
2	*mom_3m*	3.9938	Momentum
3	*std_1m*	3.9749	Trade Fraction
4	*chmom*	3.6867	Momentum
5	*mom_6m*	3.5815	Momentum
6	*beta*	3.4464	Trade Fraction
7	*mom_12m*	3.3874	Momentum
8	*mom_6m*	3.3354	Momentum
9	*std_3m*	3.1810	Trade Fraction
10	*turn_1m*	3.1473	Trade Fraction
11	*ocfg*	2.9399	Growth
12	*mvf*	2.9219	Capitalization
13	*std_6m*	2.8783	Trade Fraction
14	*rvg*	2.8391	Growth
15	*ncfp*	2.7681	Value
16	*std_12m*	2.7594	Trade Fraction
17	*ncla*	2.7294	Liquidity
18	*ocfp*	2.6363	Value
19	*turn_3m*	2.6223	Trade Fraction
20	*npg*	2.6184	Growth

5 Discussion and Conclusion

5.1 Discussion

The findings of this paper have important implications for the study of A-share market excess returns. Unlike traditional studies that examine each of the anomaly in the A-share market in detail, this paper selects a portfolio from an anomaly pool through a machine learning approach and considers the impact of the portfolio on stock excess returns in a comprehensive manner. The machine learning perspective is able to outperform the unselected anomalies portfolio. This research approach can be extended to other similar research problems.

The research method in this paper has rich insights into the application of machine learning in economics and management research. In economics and management researches, machine learning can be used to process unstructured data and extract proxy variables to construct correlation factors for hidden information. At the same time, machine learning algorithms can be used to enhance the predictive power of traditional

methods, especially for problems with nonlinear patterns, making predictions concluded to be more accurate.

The research result in this paper has rich implications for the academic practice of asset management. The findings of this paper illustrate that both the widely tested anomalies in academia and the factors frequently used in industrial investment are predictive of A-share market excess returns. In practice, the prediction of excess returns facilitated by a combination of academic and industrial anomalies can contribute to asset management practices.

5.2 Conclusion

In this paper, 36 anomalies of the A-share market from January 2011 to March 2022 are collected, and an anomaly portfolio is constructed through machine learning algorithms by conducting feature selection. Meanwhile, a quantitative investment model is constructed using nine machine learning algorithms, and its empirical performance in the Chinese stock market is tested. Comparing the empirical performance of different types of machine learning algorithms, it can be found that machine learning algorithms can achieve better results on both the A-share market excess return prediction task and the task of extracting anomalous factors applicable to the characteristics of the A-share market. In the prediction of excess return, both linear machine learning algorithms and nonlinear machine learning algorithms beat the market index returns on long portfolio, short portfolio, and long-short portfolio. Overall, the nonlinear machine learning algorithm outperforms the linear machine learning algorithm; and the integrated algorithm that includes multiple machine learning models achieves better results than the single type of machine learning algorithms.

In the task of extracting the anomalies applicable to the characteristics of the A-share market, the machine learning algorithms are able to uncover patterns of association among the anomalies of the A-share market that are difficult to identify directly, and select the best combination of anomalies based on the characteristics of the A-share market. From the machine learning perspective, in addition to the anomalies that have been widely validated by academia, the factors that are often used in practical investment, especially momentum-based factors and trading friction-based factors, also have strong predictive power for A-share market excess returns.

References

1. Hou, K., Xue, C., Zhang, L. (2020) Replicating anomalies. The Review of financial studies, 33(5): 2019-2133. https://doi.org/https://doi.org/10.1093/rfs/hhy131.
2. Harvey, C. R., Liu, Y., Zhu, H. (2016) Hundreds of papers and factors attempt to explain and the Cross-Section of Expected Returns. The Review of Financial Studies, 29(1): 5-68. https://doi.org/https://doi.org/10.1093/rfs/hhv059.
3. Feng, G., Giglio, S., Xiu, D. (2022) Taming the factor zoo: A test of new factors. The Journal of Finance, 75(3): 1327-1370. https://doi.org/https://doi.org/10.1111/jofi.12883.
4. Leippold, M., Wang, Q., Zhou, W. (2022) Machine learning in the Chinese stock market. Journal of Financial Economics,145(2): 64–82.https://doi.org/10.1016/j.jfineco.2021.08.017

5. Jiang, F., Tang, G., Zhou, G., (2018) Firm characteristics and Chinese stocks. Journal of Management Science and Engineering, 3(4): 259-283. https://doi.org/https://doi.org/10.3724/SP.j.1383.304014.
6. Gu, S., Kelly, B., Xiu, D. (2020) Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5): 2223-2273. https://doi.org/https://doi.org/10.1093/rfs/hhaa009.
7. Dong, X., Li, Y., Rapach, D. E., et al. (2022) Anomalies and the expected market return. The Journal of Finance, 77(1): 639-681. https://doi.org/https://doi.org/10.1111/jofi.13099.
8. Li, B., Shao, X. Y., Li, Y. Y. (2019) Research on machine learning driven quantamental investing. China Industrial Economics, 8: 61–79. https://doi.org/ https://doi.org/10.19581/j.cnki.ciejournal.2019.08.004.
9. Li, B., Lin, Y., Tang, W. X. (2017) ML-TEA: A set of quantitative investment algorithms based on machine learning and technical analysis. Syst. Eng. Theory Practice, 37(5): 1089–1100. https://doi.org/10.12011/1000-6788(2017)05-1089-12.
10. Green, J., Hand, J. R. M., Zhang, X. F. (2017) The characteristics that provide independent information about average US monthly stock returns. The Review of Financial Studies, 30(12): 4389-4436. https://doi.org/https://doi.org/10.1093/rfs/hhx019.