Proceedings of the 2nd International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/83/20240720

Application of Machine Learning Models in Asset Allocation

Meng Luo[1,a,*]

[1]*SDIC ESSENCE Futures Co., Ltd. Research Institute, 12F Gaoxin Tower, NO.1 Nanbinhe Rd, Beijing, China*
a. miya_luomeng@163.com
**corresponding author*

Abstract: This study embarks on an exploration of machine learning (ML) models in forecasting the trends of stock indices, with a specific focus on different industries within the Chinese market. Moving beyond the confines of traditional linear regression and standard multi-factor approaches, our research adopts a multi-dimensional analytical framework to decode the complex relationships between various factors and the future returns of industry-specific Exchange-Traded Funds (ETFs) in China. The paper innovatively applies both linear and nonlinear ML models to predict directional shifts in ETF returns, a domain not extensively studied previously. We conduct a thorough comparative analysis of these models, assessing their predictive prowess and dissecting the influence of diverse factors on different industry sectors. This investigation reveals distinct patterns and factor sensitivities unique to each sector, offering new insights into their dynamics. The results are pivotal for asset allocation and investment strategies, as they highlight the nuanced role of ML in financial forecasting. By bridging the gap between traditional financial models and advanced ML techniques, our study presents a novel perspective that enriches the strategic planning in financial markets, especially in the context of the rapidly evolving Chinese economy.

Keywords: machine learning, financial assets, price forecasting, Financial asset allocation

1. Introduction

In recent years, as research on factor investing has matured, an increasing number of active investors have shifted their focus to the factor market [1]. With the growing number of newly listed stocks in the A-share market, the time and effort spent on researching and individually selecting A-share stocks have also increased. More funds are now concentrated on selecting a basket of stocks within a specific category, which has heightened the importance of factor exploration.

This paper focuses on a research study of narrow-based indices composed of 29 different primary industries in China. By constructing a factor model to assess the trend of these indices, the aim is to assist investors in their allocation and investment decisions regarding the corresponding ETFs. Since determining the directional trend of industry indices is a classification problem, it is necessary to handle a large amount of discrete data efficiently. To address this, we introduce machine learning models capable of handling large-scale data effectively. By constructing an industry index prediction model for sector allocation, we aim to obtain market outperformance in terms of alpha returns.

DOI: 10.54254/2754-1169/83/20240720

2. Literature Review

The concept of factors originates from the Capital Asset Pricing Model (CAPM) proposed by William Sharpe [2], which decomposes stock returns into market portfolio returns (beta) and excess returns (alpha). Building upon this, the Arbitrage Pricing Theory (APT) extends the stock return pricing model by analyzing the contribution of different risk factors to returns [3]. These factors can be classified into two categories: macro factors, including economic growth, financial data, and commodity prices, and fundamental factors, primarily encompassing valuation indicators. This paper examines the predictive capabilities of linear and nonlinear machine learning models on macro and fundamental factors separately, aiming to analyze the similarities and differences in price forecasting. It provides new insights into the field of industry asset allocation.

2.1. Logistic Regression in Linear Models

The logistic regression model is commonly used in regression problems within linear models, and its formula is represented as equation (1).

$$P(y = 1 \mid x) = \frac{e^{x\vec{\beta}}}{1+e^{x\vec{\beta}}} \tag{1}$$

This formula indicates that when the probability of x belonging to y=1 is greater than 1/2, the predicted value is 1; otherwise, the predicted value is 0.

To select the most effective factors, a penalty term is usually added to choose the most predictive feature variables [4]. Generally, high-dimensional data tends to significantly increase the difficulty of model prediction as the sample size increases. However, the accuracy of model prediction does not necessarily increase with the increase in prediction difficulty. To reduce the prediction difficulty and improve the prediction accuracy of the model, regularization techniques are commonly employed. Common regularization methods include Lasso and Ridge. Lasso reduces the dimensionality of the model and selects a smaller number of features, discarding some non-important feature variables, to address the problem of model overfitting and enhance the generalization ability of the model (equation (2)). Ridge, on the other hand, is a method that compresses the weights of model factors to reduce the complexity of the model (equation (3)).

$$L(\vec{\beta}) = ||y - X\vec{\beta}||^2 + \lambda||\vec{\beta}||_1 \tag{2}$$

$$L(\vec{\beta}) = ||y - X\vec{\beta}||^2 + \lambda||\vec{\beta}||^2 \tag{3}$$

Among them, the 1-norm of β is represented by the $||\overrightarrow{\beta}||_1$, and the 2-norm of β is represented by the $||\overrightarrow{\beta}||^2$. As a penalty factor, when it is larger, β tends to become smaller and approach 0.

2.2. AdaBoost in Nonlinear Models

The Adaboost model is an alternative binary classification supervised learning method with an exponential loss function. It utilizes a forward distribution algorithm as the learning algorithm and achieves higher prediction accuracy compared to weak classifiers based on random patterns [5]. The initialization weights vary as the recognition error rate of samples increases. It distinguishes different classifiers by reducing the weights of correctly classified samples. Through iterative updates and weight incorporation, it eventually obtains the classifier $G_w(x)$. The model is composed of maximum depth (decision tree at the bottom), the number of iterations, and the combination of weights of weak classifiers (Formula (4)).

DOI: 10.54254/2754-1169/83/20240720

$$f(x) = \sum_{w=1}^{w} \left(\ln \frac{1}{\alpha_w}\right) G_w(x) \quad (4)$$

3. Research Methods

3.1. Data Collection

In data collection, factors are divided into four major categories based on the mainstream macro concepts: valuation, financial environment, market sentiment, and economic fundamentals. Valuation factors include the PB, PE, and PS values of major broad-based indexes, sourced from Wind Information (Wind). Financial environment factors include indicators that reflect domestic and international market stability, such as exchange rates and important global indexes like the S&P 500 and NASDAQ. Market sentiment factors reflect changes in market participants' sentiment and social fund demand, such as margin financing and securities lending balances and trading volumes of major exchanges. Economic fundamentals factors include CPI and PPI, which mainly reflect the current macroeconomic conditions and economic cycles, such as data representing inflation and real estate transaction volumes that reflect economic momentum.

3.2. Data Processing

A total of 105 factors were collected for the four macro dimensions. To analyze the trend of indexes, the data underwent a uniform differencing process. The data frequency includes daily and monthly data, spanning from January 2017 to December 2022, totaling 6 years. To increase the amount of data for analysis, the data frequency was unified to daily, and monthly data was mapped to daily frequency. Additionally, the data underwent preprocessing steps such as lagging, outlier removal, and standardization. Lagging ensures that the data of a certain lag corresponds to the changes in the underlying asset of the previous lag, making the overall data predictive. Outliers were replaced using a three times standard deviation approach to prevent extreme data from causing bias and distortion in the model, thereby increasing model robustness. The standardization process helps unify the scale of variables and facilitates cross-sectional analysis. In this study, the information coefficient (IC) was also calculated to gain a rough understanding of the magnitude and direction of the factor's impact on future returns. By processing the difference direction of the factor based on the positive or negative direction of the IC feature indicator, the data generated correct opening signals.

3.3. Testing Procedure

The construction of the model includes steps such as feature selection, data preprocessing, model parameter selection, and model training to predict the trend of industry indices for the next trading day. During the feature selection process, an initial screening is performed using the Information Coefficient (IC) test to observe the strength of the correlation between macro factors and industry indices, and factors with low correlation with price changes are removed. In data preprocessing, a three-fold standard deviation approach is used to prevent extreme values from adversely affecting the effectiveness of the model.

In this study, both linear and non-linear models mentioned above are employed, and the models are trained and tested multiple times using cross-validation to find the optimal model parameters. The backtesting framework is based on tradable industry indices and allows long and short combinations. The closing prices of the indices are used to calculate the corresponding net value based on buy and sell prices. The backtesting model results are evaluated based on metrics such as model accuracy, Sharpe ratio of net value, Calmar ratio, and maximum drawdown.

DOI: 10.54254/2754-1169/83/20240720

4. Results and Analysis

By calculating the Information Coefficient (IC), we can understand the predictive ability of individual factors in the logistic model. When observing valuation factors and their correlation with the rate of change in different industry indices, we find a negative correlation. A similar pattern is observed for market sentiment factors. In contrast, factors related to financial environment and economic fundamentals show slightly lower IC values compared to valuation and market sentiment factors, but most of them have positive IC values.To increase the probability of correctly predicting the direction of opening positions for individual factors in the classification model, adjustments were made to the positive and negative values of factor changes according to the tested IC direction. Regarding the logistic model, the Defense and Militarye and Building Materials sectors showed the highest cumulative returns in the training set, with returns of 98.81% and 89.41% respectively. The sectors with the highest Sharpe ratios were Transportation and Conglomerates, with Sharpe ratios of approximately 1.129 and 1.123, respectively. Looking at the testing set, Transportation and Conglomerates also showed relatively good maximum drawdown, indicating better generalization ability of the logistic model compared to other industry indices. (Table1)

In the adaboost model, the industry indices with the highest cumulative returns, Electronics and Non-ferrous Metals, outperformed the logistic model, with returns of 119.958% and 123.715% respectively. However, the overall Sharpe ratio in the adaboost model was lower than the logistic model, suggesting that the risk control capability of the adaboost model in predicting industry indices is inferior to that of the logistic model. (Table2)

5. Conclusion

Starting from machine learning models, the above experiments tested the logistic regression and AdaBoost, two common linear and non-linear models, on industry indices. Based on the experimental results, it was observed that, in terms of model performance, in situations with limited data, linear models exhibit higher stability in predictions compared to non-linear models. Specifically, when the four major macro factors were used as inputs, both models showed relatively better overall performance for the Transportation industry index. In the later stages of the experiment, further optimization of the models will be carried out to explore the capabilities of machine learning in handling large-scale and non-linear data, providing new investment insights for asset allocation in different asset classes.

References

[1] *Zhang, Y., & Wang, L. (2022). "Trends in Factor Investing and A-Share Market Dynamics," Journal of Financial Markets Research.*

[2] *Sharpe, W.F. (1964). "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," Journal of Finance.*

[3] *Ross, S.A. (1976). "The Arbitrage Theory of Capital Asset Pricing," Journal of Economic Theory.*

[4] *Nguyen, H., & Lee, J. (2019). "Factor Selection in Equity Markets: Using Logistic Regression and Regularization," Journal of Quantitative Finance.*

[5] *Smith, A., & Zhao, X. (2020). "Enhancing Financial Prediction Models with AdaBoost Algorithm," Finance and Data Science*

DOI: 10.54254/2754-1169/83/20240720

Appendix

Table 1: Training set performance of Lasso Model

	Cumulative return	Annualized return	Maximum drawdown	Sharpe ratio	Calmar ratio	Annualized frequency
Oil and Petrochemical	30.976	8.039	34.841	0.455	0.231	119.901
Coal	29.342	7.615	35.012	0.297	0.218	124.054
Non-ferrous Metals	67.970	17.640	16.484	0.751	1.071	126.649
Power and Utilities	-0.512	-0.133	17.798	-0.010	-0.008	128.206
Steel	28.058	7.282	22.144	0.322	0.329	123.535
Basic Chemicals	7.079	1.837	27.738	0.104	0.066	125.611
Construction	14.186	3.682	26.043	0.229	0.142	127.687
Building Materials	89.406	23.203	25.763	0.852	0.902	126.649
Light Manufacturing	3.980	1.033	26.172	0.060	0.040	126.130
Machinery	17.121	4.443	24.784	0.263	0.180	128.206
Power Equipment and New Energy	-23.177	-6.015	48.555	-0.290	-0.124	130.282
Defense and Military	98.811	25.644	15.743	1.072	1.631	125.092
Automobile	53.468	13.876	16.657	0.687	0.834	125.092
Trade and Retail	-2.130	-0.553	24.496	-0.036	-0.023	129.763
Consumer Services	33.626	8.727	28.311	0.271	0.309	127.168
Household Appliances	16.042	4.163	66.977	0.118	0.062	126.130
Textile and Clothing	24.275	6.300	17.386	0.519	0.363	129.244
Pharmaceuticals	37.453	9.720	23.048	0.368	0.422	126.649
Food and Beverage	80.309	20.842	70.550	0.411	0.296	124.573
Agriculture, Forestry, Animal Husbandry, and Fishery	65.608	17.027	18.621	0.647	0.915	130.282
Banks	53.664	13.927	32.191	0.638	0.433	121.977
Non-banking Financial Institutions	30.691	7.965	55.269	0.256	0.144	124.573
Real Estate	38.215	9.918	28.100	0.487	0.353	128.725
Transportation	72.116	18.716	13.484	1.129	1.390	123.015
Electronics	30.162	7.828	26.578	0.222	0.295	121.977
Telecommunications	4.973	1.291	35.169	0.050	0.037	128.725
Computers	43.802	11.368	21.076	0.407	0.540	128.725
Media	19.527	5.068	15.510	0.317	0.327	129.244
Conglomerates	71.209	18.481	11.170	1.123	1.656	126.130

DOI: 10.54254/2754-1169/83/20240720

Table 2: Training set performance of Adaboost Model

	Cumulative return	Annualized return	Maximum drawdown	Sharpe ratio	Calmar ratio	Annualized frequency
Oil and Petrochemical	73.397	19.048	14.937	1.081	1.277	119.382
Coal	-40.686	-10.559	90.353	-0.407	-0.117	127.946
Non-ferrous Metals	123.715	32.107	13.193	1.369	2.436	122.496
Power and Utilities	15.340	3.981	15.045	0.301	0.265	129.763
Steel	33.045	8.576	16.770	0.377	0.512	124.573
Basic Chemicals	67.394	17.491	17.819	0.991	0.983	124.573
Construction	6.196	1.608	19.466	0.100	0.083	129.244
Building Materials	57.478	14.917	30.549	0.547	0.489	121.977
Light Manufacturing	34.152	8.863	19.370	0.518	0.458	123.535
Machinery	42.246	10.964	18.842	0.650	0.583	123.015
Power Equipment and New Energy	-8.151	-2.115	55.752	-0.102	-0.038	126.649
Defense and Military	109.916	28.526	13.207	1.194	2.162	120.420
Automobile	31.754	8.241	25.149	0.408	0.328	122.496
Trade and Retail	36.035	9.352	14.057	0.605	0.666	128.206
Consumer Services	-16.638	-4.318	55.798	-0.133	-0.078	129.763
Household Appliances	-8.235	-2.137	45.973	-0.061	-0.047	128.725
Textile and Clothing	40.015	10.385	17.024	0.856	0.611	121.977
Pharmaceuticals	7.002	1.817	26.303	0.069	0.069	128.725
Food and Beverage	44.761	11.617	62.436	0.229	0.186	128.725
Agriculture, Forestry, Animal Husbandry, and Fishery	61.505	15.962	28.145	0.607	0.568	129.763
Banks	41.688	10.819	34.419	0.496	0.315	123.535
Non-banking Financial Institutions	-3.476	-0.902	53.775	-0.029	-0.017	129.244
Real Estate	-8.911	-2.313	50.870	-0.114	-0.046	132.877
Transportation	19.028	4.938	23.312	0.297	0.212	123.535
Electronics	119.958	31.132	24.404	0.884	1.277	124.573
Telecommunications	45.492	11.806	24.892	0.458	0.475	129.763
Computers	-21.501	-5.580	48.404	-0.199	-0.115	130.801
Media	43.047	11.172	15.873	0.699	0.705	128.206
Conglomerates	110.728	28.737	10.936	1.752	2.630	119.382