Frontiers in Economics and Management Volume 5 Issue 12, 2024
ISSN: 2692-7608 DOI: 10.6981/FEM.202412_5(12).0015

Analysis and Research on Cross Sectional Returns of Hedge Funds based on Machine Learning

Gaofeng Xu[1], Chencheng Li[2]

[1] Shanghai Urban Construction Vocational College, Shanghai 2101415, China

[2] Shanghai Guanghua Middle School, Shanghai 201105, China

Abstract

With the rapid development of the domestic fund market, research on fund evaluation and screening continues in academia and industry. Currently, the commonly used method is to establish an evaluation model based on the return and risk characteristics of fund product performance, and use historical data regression to predict future fund performance. However, the reality of the financial system often exhibits more complex nonlinear characteristics, which poses significant challenges for fund evaluation modeling and prediction work. The emergence of machine learning methods and the rapid development of IT technology have provided a new turning point for people to solve such problems. Using economic theory as a prior and machine learning methods, we aim to uncover hidden insights from data in order to better understand the complex relationship between investor expectations and portfolio performance, and design and develop investment strategies that better meet investor needs.

Keywords

Fund Market; Machine Learning; Investment Strategies.

1. Introduction

Currently, the financial industry is undergoing a major transformation. With the advent of the big data era [1-3] and continuous technological advancements, European countries have developed real-time, borderless settlement systems that can process 40000 instructions in less than 20 seconds after receiving them, making settlement transactions more efficient. In this context, data has become a key driving factor in the asset management industry. With the explosive growth of data, managers should grasp the trend of big data development and seize the opportunities it brings. Compared to traditional econometric and statistical methods, machine learning algorithms are data-driven models [4-6]. With the rapid development of Internet technology, machine learning and deep learning technologies have been widely used in various fields of life, including computer science, statistics, engineering and other fields. Machine learning, as an emerging technology in the financial field, is receiving increasing attention and research. The complexity and variability of financial markets make traditional statistical methods difficult to cope with, while the powerful data processing and analysis capabilities of machine learning make it an effective tool for solving these problems [7-9]. At present, machine learning is mainly applied in the financial field for risk management, investment decision-making, credit evaluation, fraud detection, and customer service. By analyzing and mining large amounts of data, machine learning can help financial institutions better understand market dynamics, predict risks and opportunities, improve decision-making efficiency, and provide better services to customers [10-12]. The application prospects of machine learning in the field of financial investment are broad, but continuous exploration and improvement are needed in technology, data, and regulation to realize the value and role of machine learning in investment decision-making.

ISSN: 2692-7608 DOI: 10.6981/FEM.202412_5(12).0015

2. Related Work on Hedge Fund Analysis

2.1. Research on the Factors Influencing the Performance of Public Funds

The research results indicate that the liquidity and liquidity risk of funds have good predictive ability for fund performance and can be used to evaluate the proactive management ability of fund managers. Due to the influence of market and investor sentiment, the asset flows of different market funds also vary. In the experiment, it was found that the flow of public funds changes with the market, and there is a significant correlation between the capital flow of open-ended stock funds and their performance. However, under different market cycles, investment decisions exhibit heterogeneity. In bear markets, investors are less sensitive to the historical performance of funds, while in bull markets, they tend to invest more in funds with better historical performance. Numerous studies have shown that it is possible to use certain characteristics of funds to predetermine the performance of public funds. Based on the reading and summary of existing domestic and foreign literature, scholars have conducted research on the impact of fund performance from the aspects of fund characteristics, fund manager capabilities, fund companies, and risk factors.

2.2. Research and Exploration of Investment Portfolios

A more effective stock selection model was constructed based on a weighted scoring model, and a performance prediction model was established using simulated performance data sets and artificial neural networks, achieving better test results than Standard & Poor's . A study has been conducted on the risks that companies need to bear when making investment decisions, and an improved quantitative risk assessment model has been proposed. This model can accurately identify the relationship between specific risks and investment decisions, making the work of risk managers more efficient. A directional change trading strategy was proposed for the investment portfolio of China's stock market, and the sample and out of sample performance were compared in different methods of model parameter estimation. The results showed that the average upward overshoot was greater than the average downward direction change. Therefore, statistical methods can capture important fluctuations in the stock market and obtain returns. Based on the stochastic dynamic programming technique, the corresponding Hamilton Jacobi Bellman equation is derived for dynamic investment portfolios in benchmark processes that are subject to dynamic process risk constraints. Finally, the Lagrange multiplier method is used to obtain the optimal investment portfolio strategy. We have developed an algorithm based on alternating least squares and used it to determine the eigenvectors representing investment portfolios with low variability. Research has shown that centralized investment strategies in international markets, supported by the theory of information advantage, can achieve optimal results, which traditional asset pricing theories do not possess.

2.3. Multi Factor Model Theory

In a multi factor model, factors are an indispensable part. These factors serve as the basis of the stock selection model and are the conditions and classification criteria used by the model for stock selection. As a commonly used stock selection method, establishing a multi factor model requires consideration of two aspects: how to select appropriate factors to apply to the model, and what method to use to make decisions based on the factors after selecting the appropriate ones. Generally speaking, most multi factor models use two methods when making decisions: scoring or regression. The steps of establishing a multi factor stock selection model mainly include selecting factors, testing effectiveness, establishing the model, and evaluating the model. In the process of selecting a model, many factors that have an impact on the stock market and are likely to cause future stock price fluctuations can be considered. In addition to commonly used technical and fundamental indicators, as well as economic and market information that

ISSN: 2692-7608 DOI: 10.6981/FEM.202412_5(12).0015

can affect stock market volatility, other factors related to the stock market can also be included in the factor pool. The selection of factors is not necessarily better, but increasing the number of factors as much as possible can increase the information that the model can summarize and extract, and enable more accurate analysis and decision-making, resulting in more reliable and effective conclusions.

2.4. Naive Bayes algorithm

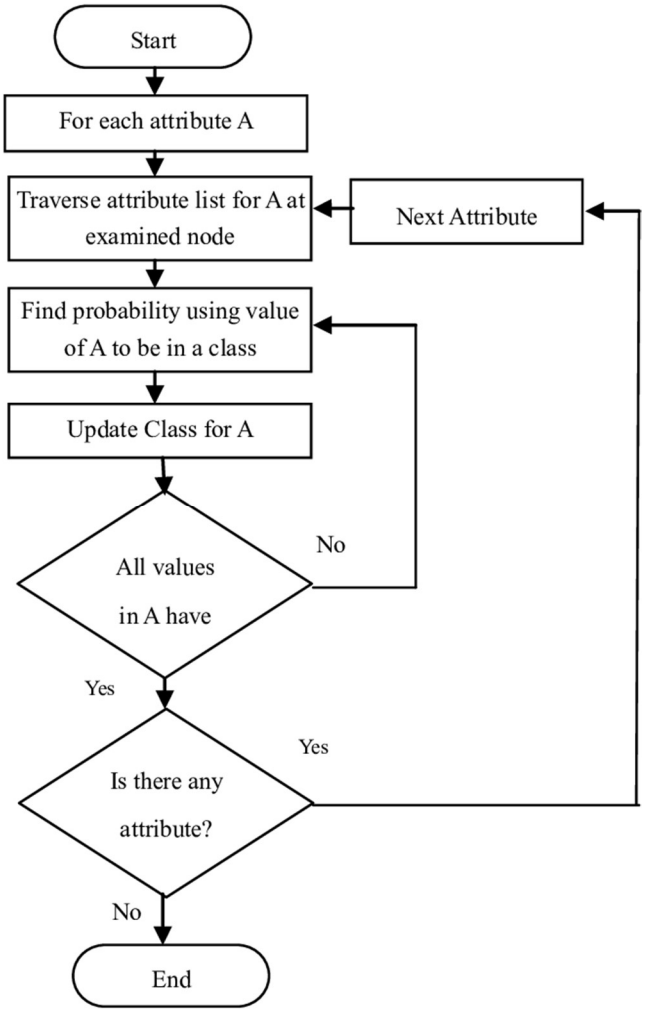


Figure 1. The workflow of Naive Bayes algorithm.

Naive Bayes algorithm was born in the 1950s and is mainly used in the field of text classification. Its main idea is that when the dimensionality of features is high, it can exhibit very efficient classification performance. In the current field of quantitative finance, Naive Bayes mainly models stock selection through linear discriminant analysis and quadratic discriminant analysis, which are improved by expanding itself. Naive Bayes algorithm is a classification method based on Bayes' theorem, whose core idea is to classify by calculating the conditional probability of each category under given features. Naive Bayes classifier assumes that features are independent of each other, meaning that the presence or absence of one feature is independent of the presence or absence of other features. The characteristic of Naive Bayes algorithm is that it assumes that all features appear independently and do not affect each other, and that each feature is equally important. But in fact, this assumption does not hold in the real world: firstly, the necessary connection between adjacent words cannot be independent; Secondly, for an article, some of its representative words determine its theme, and there is no need to read the entire article or look at all the words. So it is necessary to adopt appropriate

ISSN: 2692-7608 DOI: 10.6981/FEM.202412_5(12).0015

methods for feature selection, so that the Naive Bayes classifier can achieve higher classification efficiency. As an extension of the Gaussian Naive Bayes method, LDA (Linear Discriminant Analysis) abandons the assumption that features are independent of each other and estimates the correlation between features through covariance, while keeping other conditions constant. The secondary discriminant analysis method assumes that different categories will lead to different correlations in the characteristics of things. According to relevant reports and summaries, we can see that model tests based on Naive Bayes, Linear Discriminant Analysis, and Quadratic Discriminant Analysis have all yielded good conclusions. Generally speaking, the accuracy of the LDA model's test set and backtesting performance are better than other models. QDA's assumptions are the most detailed, but there are too many parameters to consider, while Naive Bayes has the best performance due to its concise assumptions. According to the relevant research reports of securities firms on the Wind website, it can be seen that their strategy performance in modeling based on naive Bayes algorithm. The workflow of Naive Bayes algorithm is shown in Figure 1.

2.5. Machine Learning Algorithms

The financial market is a complex and constantly evolving system, and low signal-to-noise ratio is an important characteristic of the financial market. The performance of the stock market directly affects the performance of equity securities investment funds, which in turn directly manifests as the returns and risks of equity investments in the FOF asset allocation. There are many analysis methods for the stock market, and numerous analysis methods and asset pricing models have analyzed various factors that affect the stock market. Nowadays, the number of potential covariates that can be used to predict returns is increasing day by day. Traditional data, including historical quantity and price data, financial statement data, analyst consensus data, as well as alternative data such as news and public opinion data, text analysis data, satellite image data, etc., can all be processed into various predictive variables. With the increasing variety of influencing factors and the enrichment of observation dimensions, the concept of factor zoo has emerged. At present, in terms of financial asset research, the industry has already entered the era of high-dimensional data for predicting variables. With the explosion of data that can serve as influencing factors in the financial market, the observation range and factor dimensions of market data are becoming larger and larger, leading to a sharp increase in the complexity of asset pricing prediction functions. Traditional econometric methods are no longer suitable for the current big data environment. For example, when modeling, OLS is powerless when the number of covariates that can be set exceeds the number of observation samples. Traditional econometric methods and linear methods are no longer suitable for non-linear, high-dimensional, and high noise financial market data sequences. The development of machine learning algorithms and their successful application in the field of data processing have brought us new ideas for solving financial problems. Compared to traditional analytical models, machine learning algorithms have a natural ability to process unstructured data, extract deep underlying features, and eliminate the assumption of data feature relationships and covariance matrix calculations in traditional econometric methods. They can more comprehensively describe financial laws and are an important supplement to structured data analysis. Financial time series data has characteristics such as nonlinearity, non-equilibrium, high-dimensional properties, and high noise. The research paradigm of traditional econometric models is difficult to make further breakthroughs, and machine learning does not make strict assumptions about the form of model functions, canceling assumptions such as variable probability distribution in financial markets, making it simpler and more advantageous. The research on asset pricing using machine learning has many obvious advantages such as good algorithm performance, strong applicability, and ease of handling big data. Machine learning methods can not only effectively handle large amounts of financial data, but also represent a new mode of thinking and research.

ISSN: 2692-7608 DOI: 10.6981/FEM.202412_5(12).0015

3. Dataset

3.1. Experimental Dataset

The performance of the four basic strategies and the core strategy of this article were compared horizontally with the performance benchmark using the most commonly used indicators in fund performance evaluation, including interval return, volatility, maximum drawdown, and Sharpe ratio. The comparison results are shown in Table 1 below:

Table 1. Comprehensive comparison of fund selection strategies

Strategy code	interval return	Volatility	maximum drawdown	Sharpe ratio
Benchmark	1.86	17.82	34.44	0.11
Tube01	30.62	20.87	25.07	1.47
Tube02	32.13	21.08	19.52	1.52

From Table 1, it can be seen that as the goal of selecting sub funds, it is not only necessary to track the performance standards of major asset classes, but also to avoid significant fluctuations and performance benchmark differences, while striving to obtain excess returns. Through direct comparison of strategies, we can see that machine learning methods for fund selection cannot meet the overall requirements of fund selection with a single predictive strategy. We need to use a dynamic data-driven approach to set different prediction targets for different market environments, in order to achieve a balance between risk and return by selecting investment portfolios composed of sub funds through machine learning methods. This strategy chose sub funds with low volatility risk in the adverse environment of a general market downturn, while selecting sub funds with better performance in the optimistic environment of a market rise. This objectively suppresses the overall volatility of the investment portfolio while achieving excess returns in a generally rising environment, meeting the overall requirement of selecting the best among the best for the fund.

3.2. Data Processing

According to the steps in the public fund prediction model, construct a training dataset up to September 2020 using data from multiple cross-sections of the fund, including input features and model input variables. Missing data in the dataset can be directly deleted or filled in based on the degree and importance of the missing data. Directly delete data with missing dependent variables; In order to reduce the impact of missing explanatory variables on model training, data with less than 15 feature data for some funds will be directly deleted. For other missing data, such as Sharpe ratio, annualized return, stock market value to fund net asset value ratio, investment concentration, etc., the missing data ratio is less than 5% of the total data. Therefore, the mean of the same fund is used for filling. To address the comparability between feature indicators and eliminate the impact of different input feature dimensions on model training, normalization is applied to the data.

3.3. Data Construction

To evaluate whether the performance prediction model for public funds constructed in this article has the ability to distinguish between "good" and "bad" funds and to screen for "good" funds. Firstly, verify the performance of the model under extreme market conditions of continuous rise and continuous fall. Using the public fund performance prediction model constructed and trained above, predict the fund performance during three months of continuous market rise and three months of continuous market fall. In the analysis, DOWN is used to represent the market's downtrend range, UP is used to represent the market's uptrend

ISSN: 2692-7608 DOI: 10.6981/FEM.202412_5(12).0015

range, Y1 is used to use the benchmark yield as the predictive indicator, and Y2 is used to use the cumulative yield as the predictive indicator

4. Experiments and Results

To evaluate the effectiveness of fund selection outside the sample of the performance prediction model, firstly, the ability of the evaluation model to distinguish the potential of fund performance is evaluated. Within a prediction cycle, the performance prediction results of the fund exceeding the benchmark return rate and the interval cumulative return rate are sorted and evenly divided into ten groups. The funds in each group are equally weighted to construct investment portfolios. The ability of the prediction model to distinguish between "good and bad" funds is evaluated by comparing the actual cumulative returns of each portfolio in one investment cycle. Secondly, evaluate the screening effect of the model on "good" funds, select the investment portfolio constructed by the first group of funds with equal weights, compare the cumulative return of the portfolio constructed by different models, the cumulative return of the portfolio composed of randomly selected funds, and the rise and fall of the Shanghai and Shenzhen 300 Index. And use various fund performance evaluation indicators such as standard deviation of returns, maximum drawdown, Sharpe ratio, Sortino ratio, and Alpha returns based on the Fama French five factor model to compare the performance of investment portfolios composed of funds selected through various prediction models. Finally, considering the performance of the model under different investment cycles, the first group of funds is selected from multiple prediction cycles to form a dynamic investment portfolio. The actual interval performance of the portfolio and the actual cumulative return after multiple cycles are evaluated in each prediction cycle, and the portfolio is constructed using randomly selected funds and the Shanghai and Shenzhen 300 Index as controls.

5. Conclusion

A fund selection strategy based on machine learning methods, targeting domestic stock funds and equity mixed funds as the selection targets for equity assets, with the Shanghai and Shenzhen 300 Index returns as the performance benchmark for fund selection. The strategy adopts machine learning algorithms to use multi-dimensional feature data from fund risk return data, fund managers, and fund companies to make multi-level predictions on multiple evaluation indicators such as future returns, downside risk, upside volatility, and maximum drawdown of sub funds. The strategy first takes the overall performance of the fund as the prediction target, obtains the first level prediction results, determines more targeted second level sub fund prediction targets based on the prediction of the future market, and uses machine learning methods to predict again. Based on the results, the evaluation and selection of sub funds are carried out to cope with different market environments. To achieve the overall goal of tracking the performance benchmark indicators of major assets and obtaining excess returns in fund selection, multi-level and multi-objective machine learning methods are used to select sub funds, especially different selection evaluation and prediction models are used to cope with extreme market changes in different market environments. Compared with traditional fund evaluation and selection methods, the results show that in the selection of equity sub funds, it is necessary to track the benchmark indicators of asset performance and obtain excess returns. In terms of aspects, The multi-level and multi-objective approach adopted in this article has more advantages and can provide institutional investors with a new idea when constructing sub fund selection strategies.

ISSN: 2692-7608 DOI: 10.6981/FEM.202412_5(12).0015

References

[1] Zou W .The Construction and Implementation Path of the Audit System for Quality Control in the Big Data Era[J].Proceedings of Business and Economic Studies, 2023, 6(6):111-115.

[2] Campos G D L , Gianola D .Genomic Prediction in the Big Data Era[J].American scientist, 2023(5):111.

[3] Zhang R , Yang Z , Liu X .Exploring Random Forest Modeling of Accounting Disclosure Quality and ESG Performance in the Face of Big Data Era[J].Applied Mathematics and Nonlinear Sciences, 2024, 9(1).DOI:10.2478/amns-2024-3273.

[4] Jan D P , Liesbeth T , Serge B ,et al.Validation of a machine learning algorithm to identify pulmonary vein isolation during ablation procedures for the treatment of atrial fibrillation: results of the PVISION study[J].Europace, 2024(5):5.DOI:10.1093/europace/euae116.

[5] Computing C M .Retracted: Traditional Ceramic Sculpture Feature Recognition Based on the Machine Learning Algorithm[J].Wireless Communications and Mobile Computing, 2023.DOI:10.1155/2023/9812659.

[6] Colic N , Milosevic P , Dragovic I C M S .IBA-VNS: A Logic-Based Machine Learning Algorithm and Its Application in Surgery[J].mathematics, 2024, 12(7).

[7] Seo B K .Financial Market Crash and Phase Transition: Through Model-Free Framework with Machine Learning[J]. 2020.

[8] Goel T , Chawla T , Verma N .Machine Learning Approach for Financial Analysis of Multiple Stock Market Return[J]. 2020.

[9] Feng J , Liu X , He Q ,et al.No more free lunch: The increasing popularity of machine learning and financial market efficiency[J].Economic and Political Studies, 2024, 12(1):34-57.DOI:10.1080/20954816.2023.2230622.

[10] Miragaia D A M , Ferreira J J M ,Cédric T. Vieira.Efficiency of Non-profit Organisations: a DEA Analysis in Support of Strategic Decision-Making[J].Journal of the Knowledge Economy, 2024, 15(1):3239-3265.

[11] Song H H , Zamora D G , Martinez W L .Handling multi-granular hesitant information: A group decision-making method based on cross-efficiency with regret theory[J].Expert Systems with Application, 2023, 227(Oct.):120332.1-120332.14.

[12] Jin F , Cai Y , Zhou L .Multiplicative Consistency and DEA Cross-Efficiency-Driven Decision-Making Method with Fuzzy Preference Relations[J].International Journal of Fuzzy Systems, 2023.