A Study of Investment Strategies Based on Artificial Intelligence and Machine Learning Forecasting

Cheng Peng[1*#a], Yiming Fang[2#b], Yinzhe Chang[3#a]

[1*]Corresponding author:19150221575@163.com

[2]xinledun2008@sina.com

[3] 750149063@qq.com

[a]College of Water Resource and Hydropower, Sichuan University China

[b]National University of Defense Technology Hunan China

[#]All authors contribute the same amount

Abstract: Artificial intelligence has developed rapidly since the 21st century and its integration with various fields has contributed to the rapid development of each field. This paper focuses on the quantitative application of artificial intelligence in China's financial market, by introducing 20 indicators covering value, technical, momentum and sentiment reversal and 8 machine learning algorithms to forecast stock returns in Shanghai and Shenzhen markets. In terms of the degree of contribution of each indicator to the model, this paper finds that momentum, reversal and technical indicators have the highest degree of influence on the future stock returns. The paper then ranks these stocks according to their predicted returns and develops a trading strategy. Comparing the results of each model, it is found that the trading strategies formed by the predicted returns can achieve significant excess returns in the Chinese market, with deep neural networks being the best predictors and regularised linear machine learning models the second best. Deep machine learning is used to explore the impact of each factor indicator on the Chinese stock market, providing some implications for policy makers, as well as a better understanding of the irrational factors in Chinese market trading.

Keywords: Machine learning, artificial intelligence, deep neural networks, decision tree models

1. Introduction

Artificial intelligence is an important development strategy for countries since the 21st century. Due to the wide range of applications of artificial intelligence, it can be closely intertwined with various disciplines such as the internet, big data, sensor networks, brain science, finance and image processing. Machine learning, with its advantage of non-linear data fitting, can better capture the impact of individual features on financial return forecasting. The Chinese market is dominated by individual investors and the trading behaviour of these traders has a huge impact on the volatility of the Chinese stock market. Unlike institutional investors, the trading logic of individual investors is driven by factors such as technical indicators and

ICIDC 2023, June 02-04, Nanchang, People's Republic of China

DOI 10.4108/eai.2-6-2023.2334655

momentum. Given that individual traders are a force to be reckoned with, this paper focuses more on the impact that individual traders' trading strategies have on market prices. This paper focuses on using artificial intelligence methods to provide insight into the factors that influence stock return forecasting in the Chinese market, taking into account the characteristics of the Chinese market.

2. Materials and Methods

2.1 Data sources

The data are obtained from all A-shares on the SSE and SZSE from January 2002 to June 2022 (data from WIND data), with non-compliant stocks excluded by number of trading days and listing time.

2.2 Introduction to the method

Decision tree models [1] are often used to solve complex decision problems and are particularly effective for data with high dimensionality [2] that cannot be classified by ordinary logistic regression models [3]. The decision tree model uses its complex tree classification nodes to efficiently classify the data from each parent to child node to achieve the optimal decision result. Decision trees can be classified into CLS, ID3, C4. 5 and CART algorithms depending on their attribute classification and internal node classification.

SVR [4] is an effective method for predicting financial time series as it uses a risk function consisting of empirical errors [5] and a regularisation term derived from the structural risk minimisation principle [6]. It has certain advantages for high-dimensional data. It is valid even if the data is huge and high-dimensional; the training samples used in the decision function have a certain memory effect; and it has many kernels for different purposes. However, in some cases the network also has some disadvantages. For example, different kernels and parameters may lead to overfitting. When the data set is large, it can increase the time consumption.

The multi-layer perceptron [7] starts with initial random weights and minimises the loss function by iteratively updating the weights. After calculating the loss, backward passing propagates it from the output layer to the previous layers, providing an updated value for each parameter of the weights to reduce the loss. By choosing different iteration steps and learning rates, iterations are continuously iterated and learned, and the algorithm stops when the number of descent steps reaches a preset maximum number of iterations, or when the improvement in the loss function [8] falls below a set value.

2.3 Variable selection

In this paper, based on previous research and summarising most of the existing literature, both domestic and international, indicators that have an impact on price prediction are selected and applied to each machine learning model with the aim of exploring the importance of these trading indicators on the prediction of equity risk premiums. These indicators are: market capitalisation outstanding, market capitalisation to earnings, market capitalisation to cash flow, market capitalisation to book, heterogeneous volatility, 20/240 day average turnover, ROE,

daily closing price/1 month closing price, 20 day cumulative return, 3/9/12/18/24 moving average, 3/6/12 cumulative return, 1 month maximum return, 20 day average turnover/240 day average turnover. Average daily turnover rate.

3. Machine learning portfolio returns

3.1 Long-short portfolios

The machine learning model [9] has some predictive power for stock returns in the Chinese market. To further explore the quantitative investment ability of machine learning forecasts, this paper groups the predicted returns for investment separately. Specifically, the predicted returns of the eight models and the average of these predicted returns are sorted from smallest to largest and divided equally into 10 equal parts P1, P2, ..., P10, and finally the average returns of the lowest 1 equal part, the highest 1 equal part bought and the long-short portfolio strategy with zero cost construction are found. In order to obtain the excess returns adjusted for risk factors, the Fama-French five-factor model (FF5) and the L-S-Y four-factor model (CH4) are introduced to regress the portfolio returns.

3.2 Model selection

In terms of model performance, Bayesian and Ridge perform the best, with portfolio returns of 1. 875 (2. 883) and 1. 837 (2. 830) respectively for the buy 10 equivalents. Similarly, the FF5 and CH4-adjusted returns were also the highest, at 0. 771 (3. 350) and 1. 135 (4. 554) for Bayesian and 0. 771 (3. 525) and 1. 098 (4. 528) for Ridge. The SGD model was the worst performer, with an average portfolio return of 1. 432. In terms of portfolio returns for the long-short strategy, Bayesian and Ridge continued to perform best, with average returns of 0. 842 and 0. 789 for their long-short portfolios, and 0. 859, 0. 890 and 0. 858, 0. 774 after adjustment for FF5 and CH4. To investigate the overall performance of the machine learning strategies, this paper also includes the mean, which is a metric that averages the predicted returns of all models before portfolio strategy construction. The results from the average show that it is an average performer, with a pure long portfolio having an average return of only 1. 541 and a long-short portfolio having an average return of only 0. 430.

In terms of the grouping of forecast results, the regularised linear model works best for the forecast grouping with the highest returns, with traditional machine learning such as Decision Tree, Random Forests and SVM all being less effective than the linear model, with the average return for a long 10-equivalent portfolio being 1. 514 compared to 1. 667 for the linear model. similarly, the average return for constructing Similarly, the average return on investment for constructing a long-short portfolio is only 0. 379 for machine learning compared to 0. 682 for the linear model. as seen in Figure 1, the cumulative return on buying a 10-equivalent portfolio.

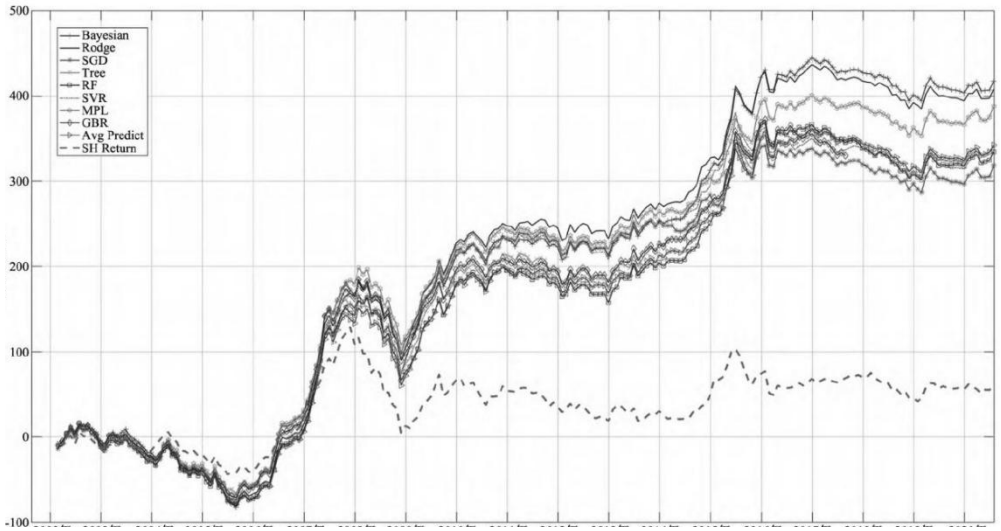

Figure 1.Cumulative earnings graph

Looking at the cumulative return charts for each model, Bayesian, Ridge and MPL performed the best, with cumulative returns of around 400% over the last 20 years for their buy 10 equivalents, while the rest of the models performed second best, with cumulative returns of around 350%. The forecast portfolio returns of these models are similar to the trend of the CSI 300, for example, from 2002 to 2005, there was a slow correction cycle in the market and the models' returns were also slow to the downside, similarly, the 2008 financial crisis and the 2016 stock market pullback were also consistent with the market trend. However, the strategy portfolio obtains higher excess returns than the SSE index, which indicates that machine learning forecasting has significant effects in the Chinese market. To investigate whether the machine learning forecasting portfolio [10] long-short strategy can withstand the risks associated with frequent market volatility, the long-short trading strategy and its cumulative returns can be seen in Figure 2.

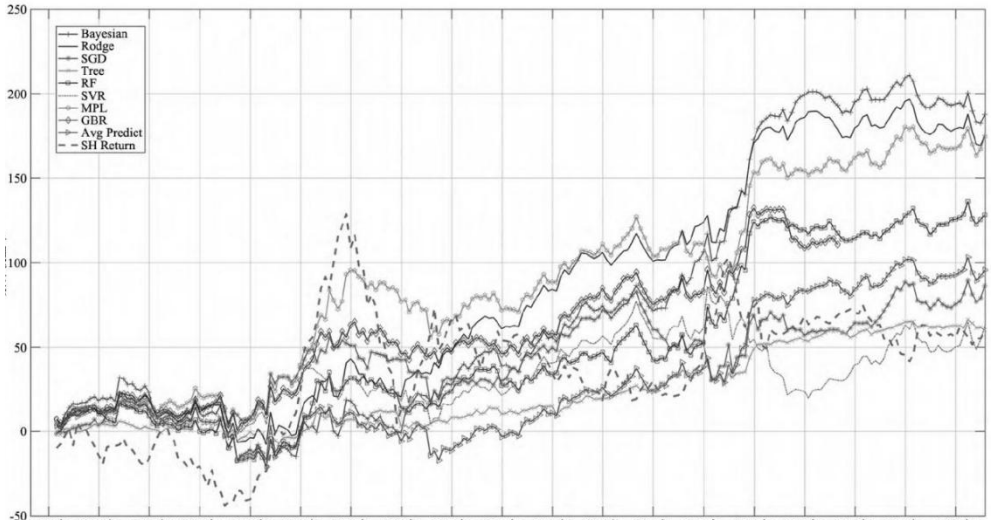

Figure 2.Cumulative earnings graph

The cumulative returns of the long/short portfolios show that Bayesian, Ridge and MPL continue to perform best, although the cumulative returns of their long/short portfolios are not as high as the cumulative returns of the long-only portfolios, but their risk has been reduced in relative terms. As can be seen from the graphs, the long-only portfolios of each model had a retracement of nearly 80% between 2002 and 2005, while the long-short portfolios had a retracement in the range of 0-20%. Similarly, in the period 2008 to 2009, the long-only portfolio had a retracement of over 100%, while the long-short portfolio had a retracement of between 0 and 50%. Thus, the long-short portfolio returns are effective in reducing risk, but their returns also decline.

3.3 Analysis of results

In order to compare the forecasting performance of each machine learning model more intuitively, several return-risk performance indicators are used, namely Sharpe Ratio:$\frac{E(R_p)-R_f}{\sigma_p}$, where $E(R_p)$ is the annualised return of the portfolio, R_f is the risk-free rate and σ_p is the standard deviation of the annualised return. Therefore, the higher the ratio, the better the strategy; Calmar Ratio:$\frac{E(R_p)-R_f}{MaxDD}$, where $E(R_p)-R_f$ is consistent with the Sharpe Ratio and is the excess return and $MaxDD$ is the maximum retracement rate; Omega Ratio:$\frac{\int_r^{\infty}[1-F(x)]dx}{\int_{-\infty}^{r}F(x)dx}$,F is the cumulative distribution function of asset returns and r is the target return used to determine whether an asset's return is positive or negative. The higher the ratio, the greater the portion of the asset's return r over the target return is than the portion of the return below the target return r; Sortino Ratio:$\frac{E(R_p)-R_f}{DR}$, where $E(R_p)-R_f$ is the excess return and DR is the downside risk, is the standard deviation of the portion of the return below zero. This ratio is an improvement on the Sharpe Ratio, i.e. it only calculates the risk of negative returns, while positive returns are not calculated. the larger this indicator is, the higher the return of the

portfolio strategy and the lower the downside risk. annual Ret is the annualised return and STD is the standard deviation.

Table 1.Machine learning and other results

Long	Sharpe	Calmar	Omega	Sortino	Annual Ret	Std.Dev
Bayesian	1.910	0.281	1.020	1.209	0.185	0.097
Ridge	1.862	0.278	1.011	1.182	0.180	0.098
SGD	1.289	0.199	0.905	0.882	0.124	0.096
Tree	1.470	0.215	0.933	0.946	0.140	0.093
RF	1.464	0.204	0.919	0.955	0.137	0.093
SVM	1.548	0.226	0.924	0.961	0.141	0.091
MLP	1.850	0.263	0.985	1.161	0.171	0.094
GBR	1.449	0.201	0.926	0.959	0.138	0.096
AVG.model	1.550	0.222	0.930	0.987	0.143	0.093
Long-Short	Sharpe	Calmar	Omega	Sortino	Annual Ret	Std.Dev
Bayesian	1.702	0.211	0.604	0.928	0.088	0.052
Ridge	1.736	0.293	0.563	0.931	0.084	0.048
SGD	0.767	0.082	0.413	0.457	0.035	0.046
Tree	1.492	0.218	0.153	0.707	0.032	0.022
RF	1.199	0.169	0.491	0.709	0.057	0.048
SVM	0.405	0.039	0.432	0.296	0.020	0.049
MLP	1.602	0.237	0.583	0.901	0.082	0.053
GBR	1.692	0.321	0.517	0.928	0.076	0.045
AVG.model	0.957	0.108	0.405	0.515	0.043	0.043

As can be seen from Table 1, the long-only (Long) machine learning strategy slightly outperforms the long-short (Long-Short) strategy, and the long-only strategy achieves higher Sharpe, Calmar, Omega and Sortino ratios than the long-short strategy, despite its higher risk (larger standard deviation), due to its higher return. In terms of long portfolio performance, the Bayesian model continues to perform best, with Sharpe, Calmar, Omega and Sortino ratios of 1. 909, 0. 280, 1. 021 and 1. 200 respectively, and an annualised return of 18. 5%, but it also has the highest risk, with a standard deviation of 9. 7%.

The Ridge and MLP models are the next best performers, with the Sharpe Ratio of these two models also reaching over 1. 8 for the forecast return portfolio. In terms of the long/short portfolio strategy, the strategy sells a portfolio of stocks with poorer forecast returns, which significantly reduces the downside risk associated with these stocks due to the hedging mechanism. As a result, after hedging out the risk, the long-short portfolio has a lower risk of return with half the standard deviation of a pure long portfolio. Again, the Baysian and Ridge regressions continue to perform best in terms of results. This indicates that these two models are better at predicting stock returns, both for strong stocks and for weak stocks, and their predicted returns combine to form a portfolio of stocks that achieve higher returns. In terms of the other ratios, the three models Baysian, Ridge and MLP also performed better, confirming that these models combine strategies with relatively low risk while achieving high returns.

3.4 Robustness tests

As a robustness check, this paper repeats the above approach to verify the robustness of the machine learning model by excluding stocks in the bottom 30% of the market capitalization according to their approach. There are three main reasons for this approach: (1) In the Chinese stock market, small-cap stocks are known for their high volatility, which makes it more difficult to predict the model; (2) The bottom 30% of stocks have a "shell effect" problem and are more likely to operate in the dark, and they are difficult to fit the model with market indicators; (3) In general, large-cap stocks have higher level of liquidity and lower price volatility, therefore, these stocks are less affected by the 10% daily price limit in China. In summary, this paper removes the bottom 30% of stocks and derives the respective ratios for the long and short portfolios. From the results, the machine learning portfolios based on the top 70% of large-cap stocks perform similarly to the full sample in terms of their results. However, due to the exclusion of the more volatile small-cap stocks, all of the modeled portfolios achieved lower average monthly returns, Sharpe ratios, standard deviations and annualised returns. However, the predicted portfolio returns of the machine learning algorithms were higher than those of the CSI 300. Of these, the deep network portfolio had the best return performance, followed by the regular linear model and the tree model. These models have the highest long returns and long-short portfolio returns, and therefore the robustness test results also confirm that machine learning methods have excellent forecasting capabilities in the Chinese stock market.

4. Conclusion

This paper explores the relationship between value, momentum, reversal and trend following factors and future stock returns in the Chinese stock market by introducing classical machine learning models, and finds that historical stock information has some predictive power for their future returns; by comparing the full sample data it is found that small capitalisation stocks are more predictive, and the full sample returns with the inclusion of small capitalisation stocks are higher than the predicted returns with the exclusion of these samples higher. Despite the impact of the new crown pneumonia epidemic after the end of 2019, the results of the study remain robust and the constructed machine learning quantitative strategy achieves higher positive returns in 2020, which illustrates the effectiveness of the selected factors and can better improve the formulation of investor education in China and strengthen investors' concept of value investment plays an important role in the stable development of the Chinese stock market.

References

[1] Chen Boyu,Xu Ming,Yu Hongmei,He Jiachuan,Li Yingmei,Song Dandan,Fan Guo Guang. Detection of mild cognitive impairment in Parkinson’s disease using gradient boosting decision tree models based on multilevel DTI indices[J]. Journal of Translational Medicine,2023,21(1).doi:10.1186/S12967-023-04158-8.

[2] Shiravani Anita,Sadreddini Mohammad Hadi,Nahook Hassan Nosrati. Network intrusion detection using data dimensions reduction techniques[J]. Journal of Big Data,2023,10(1).doi:10.1186/S40537-023-00697-5.
[3] Soh Chin Gi,Zhu Ying,Toh Tin Lam. A regularised logistic regression model with structured features for classification of geographical origin in olive oils[J]. Chemometrics and Intelligent Laboratory Systems,2023,237.doi:10.1016/J.CHEMOLAB.2023.104819.
[4] Kou Lei,Sysyn Mykola,Liu Jianxing,Fischer Szabolcs,Nabochenko Olga,He Wei. Prediction system of rolling contact fatigue on crossing nose based on support vector regression[J]. Measurement,2023,210.doi:10.1016/J.MEASUREMENT.2023.112579.
[5] Endo Yutaka,Alaimo Laura,Sasaki Kazunari,Moazzam Zorays,Yang Jason,Schenk Austin,Pawlik Timothy M. Liver Transplantation for Colorectal Liver Metastases: Hazard Function Analysis of Data from the Organ Procurement and Transplantation Network.[J]. Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract,2023.doi:10.1007/S11605-023-05672-2.
[6] EL Qate Karima,El Rhabi Mohammed,Hakim Abdelilah,Moreau Eric,Thirion Moreau Nadège. Hyperspectral Image Completion Via Tensor Factorization with a Bi-regularization Term[J]. Journal of Signal Processing Systems,2022,94(12).doi:10.1007/S11265-022-01817-9.
[7] Farouk Naeim. Applying machine learning based on multilayer perceptron on building energy demand in presence of phase change material to drop cooling load[J]. Engineering Analysis with Boundary Elements,2023,150.doi:10.1016/J.ENGANABOUND.2023.02.003.
[8] Tolpadi Aniket A.,Han Misung,Calivà Francesco,Pedoia Valentina,Majumdar Sharmila. Region of interest-specific loss functions improve T2 quantification with ultrafast T2 mapping MRI sequences in knee, hip and lumbar spine[J]. Scientific Reports,2022,12(1).doi:10.1038/S41598-022-26266-Z.
[9] Huang Weihua,Lyu Yang,Du Minghao,He Can,Gao Shangde,Xu Renjun,Xia Qunke,ZhangZhou J. Estimating ferric iron content in clinopyroxene using machine learning models[J]. American Mineralogist,2022,107(10).doi:10.2138/AM-2022-8189.
[10] Consunji P M. EvoTrader: Automated Bitcoin Trading Using Neuroevolutionary Algorithms on Technical Analysis and Social Sentiment Data[C]//International Association of Applied Science and Engineering.Conference proceedings of 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence (ACAi 2021).ACM,2021:688-696.DOI:10.26914/c.cnkihy.2021.055307.