Quantitative Stock Selection Based on Artificial Intelligence

Yihong Li*, Feiran Wang

School of Economics and Management, University of Science and Technology, Beijing, China

2305924928@qq.com*,1085230559@qq.com

Abstract. Artificial intelligence has emerged as a prominent catalyst for technological advancements in recent years. As a cross-disciplinary domain encompassing computational mathematics, statistics, and informatics, artificial intelligence holds significant potential for applications in financial trading and quantitative analysis. This research paper focuses on the utilization of machine learning techniques to analyze seven major fundamental factors of Chinese listed companies from September 2013 to September 2022. Specifically, five machine learning algorithms, including linear regression, Lasso regression, ridge regression, random forest, and decision tree models, are employed to identify the top 30 stocks that offer the most promising expected returns for an equal-weight holding strategy. The performance of this strategy is then compared with that of the CSI 300 index, a widely recognized benchmark. The findings demonstrate that, within the same time period, the aforementioned algorithms outperform the CSI 300 index in terms of returns and drawdowns. Notably, the ridge regression model exhibits the most favorable performance, boasting an annualized return of 14.45% and a maximum drawdown of 0.95% among all selected models. This study, by employing a diverse range of linear and non-linear machine learning algorithms for modeling purposes, contributes to the advancement of the quantitative investment field and provides valuable theoretical insights for the formulation of novel trading strategies.

Keywords: Factor analysis, Quantitative trading, Machine learning

1 Introduction

In recent years, the financial industry has witnessed significant advancements due to the implementation of cutting-edge technologies such as artificial intelligence and machine learning, undergoing significant transformations. One notable domain benefiting from this technological revolution is quantitative stock selection. Fundamental quantitative investment, which combines quantitative and value analysis, has emerged as a modern investment approach. Its key objective involves examining the relationship between fundamental information of listed companies and excess stock returns. In stock selection strategies, this primarily entails identifying a set of factors capable

A. Bhunia et al. (eds.), *Proceedings of the 2023 International Conference on Finance, Trade and Business Management (FTBM 2023)*, Advances in Economics, Business and Management Research 264,
https://doi.org/10.2991/978-94-6463-298-9_29

of generating Alpha returns and subsequently constructing a multi-factor stock selection model. This approach has garnered considerable attention within the last decade.

Gu Shiyong, drawing inspiration from the Barra E3 model, conducted extensive tests on selected factors from both macro and micro perspectives, using 8 factors that passed the tests to build a multi-factor model [1]. Chen Dehua further expanded the model by incorporating macroeconomic and industry factors, highlighting the positive impact of industry factors on the model's performance [2]. The CNE5 version, released by Barra in 2012 as an expansion of the USE4 version, gained significant popularity among Chinese investors [3]. This version categorized factors into eleven groups, including market value, momentum, beta, liquidity, volatility, leverage, etc. In an effort to explain financial phenomena such as momentum effects and address emerging demands in the field, Fama augmented the three-factor model by introducing investment and profitability. Subsequently, Barra (2018) introduced the CNE6 model, which featured a more detailed segmentation of style factors, expanding the number of layers from two to three. Notably, quality factors, sentiment factors, and dividend factors were incorporated into the first-level factors of this model [4]. Overall, existing research on multi-factor model primarily focuses on discovering new factor combinations, while their application effects in the Chinese stock market remain largely unexplored [5]. Therefore, this study aims to investigate the relationship between factors derived from company financial data and historical trading data, and stock excess returns. A specific focus is placed on utilizing machine learning algorithms for factor-based stock selection, with the objective of attaining higher excess returns and improved portfolio performance in the market. To achieve this, the study collected 15 factors from the A-share market in the period from September 2013 to September 2022 and employed both linear and nonlinear machine learning algorithms, including linear regression, Lasso regression, ridge regression, random forest regression, and decision tree regression, for predictive stock selection.

The research presented in this paper holds significant practical significance. The integration of financial trading and algorithms has been a focal point in recent years among researchers in academia [6]. The findings of this study can serve as a theoretical reference for this fusion. Additionally, as an empirical analysis using actual data from the Chinese securities market, this paper enriches the practical cases in the field of financial quantification. It offers potential practical value for financial institutions operating in the market.

2 Factor Processing

2.1 Model Design

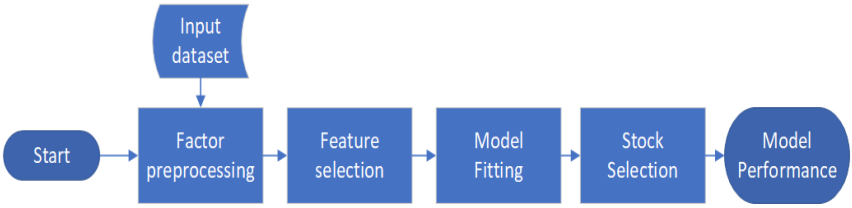

Fig. 1. Model Flowchart

Fig 1 illustrates the overarching process of the AI-based quantitative stock selection model. This paper commenced by selecting securities from the A-share market in "normal trading" [Normal trading: Refers to the securities listed for over 12 months that are neither subject to risk warnings nor suspended] as the stock pool. Historical trading data and financial analysis data were used as factors, which were then subjected to IC/IR tests to screen out factors that possessed adequate information for modeling. The feature engineering module's task was to convert the original set of factors into combinations that effectively captured the essence of the problem, thereby enhancing stock selection accuracy when applied to the predictive model [6]. Consequently, this improved the predictive performance of the model. Feature importance evaluation was accomplished through the utilization of two methods: PCA (Principal Component Analysis) and random forest for feature selection.

As fig 2 shows below, this paper adopted a rolling forecast method to divide the training and prediction sets, which was necessary for the time-series nature required by the stock selection model. The main steps of model training and testing are as follows: <1> Utilizing the experimental data starting from September 2013 as an illustration, each quarter was treated as a distinct period [In this context, a quarter does not imply a natural quarter, but rather takes into account the lag in financial data publication. In China, the deadline for publishing Q1 and annual report data is April 30th, Q2 data is August 31st, and Q3 data is October 31st. Factor data for period "t" refers to the most recent available financial report data, specifically: ① For data from January to March, the previous year's Q3 report is utilized; ② For data from April to July, the current year's Q1 report is employed; ③ For data from August to September, the current year's semi-annual report is used; ④ For data from October to December, the current year's Q3 report is adopted]. The subsequent 10 quarters, encompassing September 2013, were selected as the training set for model fitting and obtaining model parameters. <2> The trained model was then applied to the factor data as of September 2016 to forecast stock returns for March 2017. <3> Based on the predictions from step <2>, the expected returns of stocks were ranked, and the top 30 stocks were selected for an equal-weight holding portfolio. <4> This portfolio was held for one period, and the investment returns for March 2017 were calculated. <5> As the time

advanced to March 2017, steps <1> to <4> should be repeated until the end of the data period.

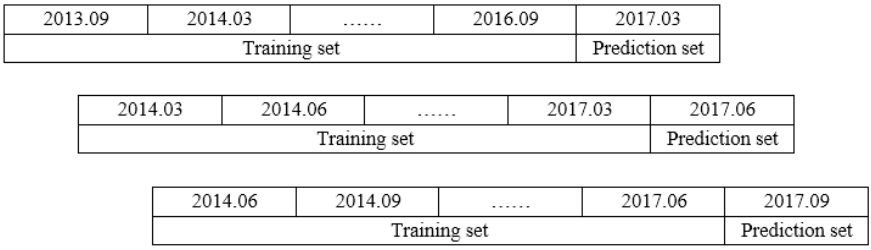

Fig. 2. Rolling Forecast Method Illustration

Rolling forecasting aligns more closely with realistic investment behavior, and the fixed lengths of the training and forecast sets contribute to standardizing the model[7]. In accordance with the aforementioned segmentation, the dataset could be partitioned into 12 sets for training and forecasting purposes. Furthermore, in order to ensure the scientific validity of the length selection, this study examined the model performance for training set lengths ranging from 1 to 21 periods. For fine-tuning the machine learning algorithms, this paper adopted a method known as "randomized grid search." This involved initially setting hyperparameters for each machine learning algorithm, and then selecting the optimal parameters from a predefined parameter set. The chosen optimal parameters were subsequently applied to the test set to determine the final model performance. By using this approach, the study systematically explored the parameter space and identified the most suitable configuration for each algorithm. This optimization process enhanced the models' performance, rendering them more applicable in predicting stock returns and increasing their practical value in investment decision-making.

2.2 Dataset Processing

The sample population of this study comprised all publicly listed companies in the A-share market of the Shanghai and Shenzhen Stock Exchanges, spanning from September 2013 to September 2022. Data collection was conducted on a quarterly basis using financial reports. To mitigate trading risks and account for the generally weak financial performance associated with ST (Special Treatment) stocks, all stocks labeled as ST and *ST were excluded from the study. Additionally, stocks with less than one year of listing were omitted due of the observed phenomenon of price suppression leading to significant price declines in their initial listing year [8].

Regarding factor selection, this paper first identified seven primary factors listed in table 1, namely volatility factor, momentum factor, profitability factor, growth factor, size factor, liquidity factor, and value factor. These primary factors were further subdivided, with specific indicators outlined in the table below.

Table 1. Selected factors and their calculation methods

Primary indicator	Secondary indicator
Volatility factor	Average True Range
Momentum factor	1-month past performance 3-month past performance 6-month past performance" 12-month past performance
Liquidity factor	Monthly turnover rate
Value factor	Earnings per share Net cash flow per share Net asset value per share
Growth factor	YOY EPS growth rate YOY net profit growth rate
Profitability factor	ROE ROIC
Size factor	Total market capitalization Float

After excluding stocks that failed to meet the criteria, a total of 427,203 data points were obtained. Within this dataset, certain data points were found to be missing, primarily due to trading halts and incomplete financial indicators corresponding to the selected factors. To ensure data integrity, stocks with missing values were eliminated from the overall stock pool, resulting in a final dataset of 242,460 valid data points.

2.3 Feature Selection

Feature selection plays a pivotal role in identifying informative factors while excluding irrelevant ones. Evaluation of the correlation between feature items and target items, referred to as feature importance, aids in simplifying the model. Feature selection can be carried out through either correlation or divergence, scoring each feature accordingly. Thresholds or the desired number of selected thresholds can be set to guide the feature selection process. This method, known as the "Filter" method, offers computational efficiency and robustness against overfitting [9]. However, it is prone to selecting redundant features. To ensure the scientific validity of the selection results, this study simultaneously employed the random forest algorithm to train the model, obtaining the weight coefficients for each feature, thereby assisting in feature selection process. By utilizing different filtering conditions for both methods, filtered factors were obtained, and these filtered factors were subsequently employed in a linear regression model for stock selection, leading to the following outcomes showed in table 2 and table 3.

Table 2. Random Forest selection results

Filtering criteria	Annualized return	Maximum draw-down	Return/Drawdown
No filtering	12.88%	1.47%	8.7731
Remove factors with importance<=0.05	13.8%	0.95%	14.5646
Remove factors with importance<=0.06	4.25%	2.75%	1.5484
Remove factors with importance<=0.07	1.13%	2.65%	0.4254

Table 3. PCA dimension reduction results

Dimension	Annualized return	Maximum draw-down	Return/Drawdown
14	11.24%	2.17%	5.1746
13	5.32%	4.8%	1.1082
12	5.18%	4.55%	1.1363
10	4.98%	2.83%	1.7609
7	4.05%	2.5%	1.1617
2	1.08%	3.9%	0.2757
Control group	12.88%	1.47%	8.7731

The above results indicate that, across all filtering criteria, employing the random forest algorithm to identify factors exhibiting an importance level exceeding 0.05 demonstrates superior efficacy in the context of stock selection. Consequently, in this paper, we opted to incorporate all indicators except for earnings per share and net asset value during the model fitting phase.

3 Model Fitting

3.1 Experimental Environment

The software environment used in this study is Python 3.8, while the computer hardware configuration employed is detailed in table 4. All data mentioned in this research were sourced from publicly available information of listed companies, which were collected via the Choice Financial Terminal database.

Table 4. Experimental hardware setup

Hardware	Hardware configuration
CPU(12 cores)	Intel(R) Core(TM) i7-10750H CPU @ 2.60GHZ
RAM	16G
GPU	NVIDIA GeForce GTX 1660 Ti

3.2 Experimental Results

By running diverse machine learning algorithms, the study successfully achieved the optimal performance of the ensemble showed in table 5.

Table 5. Model performance

Algorithm	Annualized return	Maximum drawdown	Return/Drawdown
Linear Regression	13.8%	0.95%	14.5656
Lasso Regression	7.21%	4.58%	1.5753
Ridge Regression	14.45%	0.95%	15.2406
Random Forest	8.01%[1]	5.72%	1.4004
Decision Tree	6.03%	5.26%	1.146

During the corresponding time period, the performance of the CSI 300 Index is showed in table 6.

Table 6. Performance of CSI 300 Index

	Annualized return	Maximum drawdown	Return/Drawdown
CSI300	5.2781%	6.7519%	0.7817

The findings indicate that the selected five linear and nonlinear algorithms in this paper exhibited superior performance compared to the CSI 300 Index within the same time frame. Among all the models, the Ridge Regression algorithm demonstrated the highest performance, yielding an annualized return of 14.45% while effectively managing the maximum drawdown of 0.95% with an Alpha parameter of 19.34. This outcome signified a consistent and favorable return for the portfolio. The remaining models showcased commendable risk-balancing capabilities, wherein the ratio of annualized return-to-drawdown surpassed1, exemplifying their ability to strike a balance between returns and risks.

4 Conclusion

This research study adopts an empirical approach to investigate the effectiveness of five distinct machine learning algorithms in stock selection based on 15 indicators in the A-share market from September 2013 to September 2022. The primary objective of this study is to evaluate the returns and risk performances of these algorithms, compare their trading outcomes, and assess the feasibility and effectiveness of using

[1] This section may yield better results, as the random forest algorithm involves significant computational demands. Due to hardware limitations, this study has set the maximum number of learners (n_estimators) to 600 (under this condition, a single run already exceeds 160,000 seconds). During the research process, the study found that the returns of the random forest algorithm increased as the number of learners (from 25 to 600) increased. Therefore, it is not ruled out that when a larger value is chosen for the number of learners, the model may produce even higher returns.

the aforementioned indicators in stock trading. Based on the research findings, the following conclusions can be drawn: Firstly, linear machine learning algorithms outperform nonlinear ones under the same model complexity. Secondly, the analysis of historical market data enables the generation of excess returns, thereby confirming the weak efficiency of our domestic market.

The quantitative stock selection model designed in this study holds significant practical implications. The model is founded upon rigorous algorithms and historical data, thereby mitigating the influence of fluctuations in investor sentiment and countering the temptation of making irrational decisions driven by emotional factors. Furthermore, leveraging comprehensive data analysis, the model possesses the capability to unearth latent patterns within the market, empowering investors with enhanced comprehension of prevailing market trends. Additionally, this model is amenable to risk management applications encompassing diversification, position management, and volatility control, which underpin the foundations of both long-term investments and short-term trading strategies.

In future research, further adjustments can be made to the parameters of the random forest algorithm, and additional algorithms such as XGBoost can be included for comparison. These endeavors hold the potential for developing trading strategies that exhibit superior performance while minimizing associated risks, promoting the advancement of the asset management industry and facilitating the convergence of finance with algorithmic power.

References

1. Gu S. (2007) Construction of Multi-Factor Model in China's A-Share Market (Master's Thesis, Central China Normal University).
2. Chen D, Sun C, Shi J. (2009). Multi-Factor Model in the Securities Market and Its Application in the Shanghai and Shenzhen A-Share Market. Productivity Research 21: 112-115.
3. Rosenberg B. (2012) The Estimation of Stationary Stochastic Regression Parameters Reexamined. Journal of the American Statistical Association, 67.
4. Rosenberg B, Guy J. (1976) Prediction of Beta from Investment Fundamentals: Part One. Financial Analysts Journal, 32.
5. Li B, Shao X, Li Y. (2019). Machine Learning-Driven Fundamental Quantitative Investment Research China Industrial Economics, 08:61-79. DOI:10.19581/j.cnki.ciejournal.2019.08.004.
6. Yuan Z L, Long X L. (2021) Hierarchical ensemble learning method in diversified dataset analysis. Journal of Physics: Conference Series, 2078(1).
7. Pasupulety U, Abdullah Anees, A, Anmol, S, Mohan, B. R. (2019). Predicting stock prices using ensemble learning and sentiment analysis. Proceedings - IEEE 2nd International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2019, 215–222. https://doi.org/10.1109/AIKE.2019.00045
8. Tang C, Yu Q, Li X, Lu Z, Yang Y. (2023). Application of DeepForest-CQP multi-factor model in quantitative stock selection. Journal of Intelligent and Fuzzy Systems, 44(3), 5425–5436. https://doi.org/10.3233/JIFS-222328
9. Li P, Xu J. (2022). A Study of Different Existing Methods for the Stock Selection in the Field of Quantitative Investment. Wireless Communications and Mobile Computing, 2022.