Genetic programming algorithms and factors

Zhihao Cui*

Abstract

The efficient market hypothesis states that in an efficient market, all available information will be quickly and accurately reflected in asset prices, making it impossible for investors to obtain excess returns by exploiting public or non-public information. Based on the efficient market hypothesis, this paper uses genetic programming algorithm to mine the efficient factors of China A-share market to test the efficiency of the market in different time spans. The study selects the A-share market OLHC data from 2019 to 2023, removes ST stocks, and considers the impact of market value factors. By mining factors on the training set and filtering out the top five factors based on IC values, the long-short portfolio is constructed in the test set, and the CH-4 factor model and Fama-French regression method are used for testing. The results show that the mined factors can significantly predict the market return in different time spans, the portfolio return is obvious, and the validity of the factors is verified by regression test, which indicates that there are effective factors that can be mined to predict the future return in China A-share market to A certain extent, and the market validity needs to be further improved. This study provides A new method and empirical basis for the study of the effectiveness of China's A-share market, and has certain reference value for investors to formulate investment strategies and market regulators to improve the market mechanism.

1 Introduction

The Efficient Market Hypothesis (EMH) is an important cornerstone of modern financial theory. Its core idea is that in efficient markets, all available information is quickly and accurately reflected in asset prices, Investors cannot obtain excess returns through public or non-public information (Fama, 1965) (Fama, 1970). According to different information sets, EMH is divided into

*School of Software, BeiHang University. zhihaocui@buaa.edu.cn

three forms: weak efficiency, semi-strong efficiency and strong efficiency (Fama, 1970). Weakly efficient markets assume that prices reflect all historical trading information. Semi-strong efficient markets assume that prices reflect all public information. However, the strong efficient market assumes that prices reflect all public and undisclosed information (Fama, 1998). Despite the theoretical importance of EMH, its applicability in real markets has been controversial, especially in emerging markets and under certain conditions, where prices may not fully reflect all information, thus providing investors with potential arbitrage opportunities (Allen et al., 2020).

In recent years, with the increasing complexity of financial markets and the explosive growth of data volume, more and more studies have begun to explore how to use advanced data analysis methods to examine market efficiency. As an optimization algorithm based on natural selection and Genetic mechanism, Genetic Programming (GP) has been gradually applied to the study of financial markets because of its powerful feature mining ability and adaptability to complex nonlinear relationships (Koza, 1992) (Koza, 1994) (Banzhaf et al., 1997). By simulating the process of biological evolution, genetic programming algorithm can automatically screen out factors with predictive ability from massive data, which provides a new perspective and tool for studying market efficiency.

In terms of the research on the efficiency of China's A-share market, the existing studies have extensively explored it using traditional methods. For example, the Fama-French three-factor model has been widely used in the empirical research of the Chinese stock market, and the results show that the model has good applicability in the Chinese market and can effectively explain the changes in stock return rates (Fama and French, 1988) (Fama and French, 1996) . In addition, the CH-4 factor model has also been mentioned in related studies, which further optimizes its ability to explain market returns by introducing more factors (Liu et al., 2024) . However, there is still debate about the effectiveness of China's A-share market. Some studies have shown that the Chinese stock market does not meet the weak efficiency hypothesis in some cases, especially after the 2004 reform of non-tradable shares, the market efficiency has not been significantly improved (Gao et al., 2021). In addition, the study also found that the momentum strategy still has significant profitability in the Chinese market, further indicating that the market may not achieve full efficiency (Kang et al., 2002).

This paper aims to mine the effective factors of China A-share market by genetic programming algorithm, and test the effectiveness of the market in different time spans (one day, one week and one month). The study selects the opening price (O), highest price (H), lowest price (L) and closing price (C) data of the A-share market from 2019 to 2023, removes ST stocks, and considers the influence of market value factors. By mining the factors on the training set (2019-2020) and selecting the top five factors based on the Information Coefficient (IC) values, the long-short portfolio is constructed in the test set (2021-2023). And the CH-4 factor model (Liu et al., 2024) and Fama-French regression method (Fama and French, 1996) were used for testing. The results show that the mined factors can significantly predict the market return in different time spans, the portfolio return is significant, and the validity of the factors is verified by regression test. This indicates that there are effective factors that can be mined to predict future returns in China's A-share market to A certain extent, and the market effectiveness needs to be further improved.

The contributions of this paper are as follows: firstly, genetic programming algorithm is used to mine market efficiency factors, which provides a new method for testing market efficiency. Secondly, combined with the CH-4 factor model (Liu et al., 2024) and Fama-French regression method (Fama and French, 1996), the validity of the factor was further verified. Finally, this study provides A new empirical basis for the study of the effectiveness of China's A-share market, and has certain reference value for investors to formulate investment strategies and market regulators to improve the market mechanism.

The structure of this paper is as follows: the second part is a literature review and reviews the related research progress. The third part introduces the research methods and data. Section 4 shows the empirical results and analysis. The fifth part is the conclusion and prospect.

2 Review of Selected Literature

In this section, we address several studies that have attempted to develop models to predict financial market movements and optimize trading strategies. These studies employ various methodologies, including genetic algorithms, support vector machines, and machine learning techniques, to enhance the accuracy and efficiency of financial forecasting and portfolio management. (Ruiz-

Torrubiano and Suárez, 2008) Routledge (1996) investigated adaptive learning in financial markets, particularly the imitation and experimental processes within the Grossman-Stiglitz model. The study demonstrated that under a fixed proportion of informed traders, monotonic selection dynamics converge to rational expectations equilibrium. Additionally, the robustness of the learning process to noise experimentation was explored, and conditions under which adaptive learning leads to Grossman-Stiglitz equilibrium were proposed. (Dunis et al., 2013) proposed a hybrid model combining Genetic Algorithm (GA) and Support Vector Machine (SVM) to forecast and trade the daily and weekly returns of the FTSE 100 and ASE 20 indices. The hybrid model outperformed other models, such as higher-order neural networks and ARMA models, in terms of trading performance and directional prediction accuracy. (Ruiz-Torrubiano and Suárez, 2008) introduced a hybrid optimization approach that combines genetic algorithms and quadratic programming to address the index tracking problem. By selecting a subset of assets and optimizing their weights, this strategy constructs near-optimal tracking portfolios at a lower computational cost. (Gupta, Mehlawat, and Mittal, 2011) developed a multi-criteria portfolio optimization method using Support Vector Machines (SVM) and Real-Coded Genetic Algorithm (RCGA). SVM was employed for asset classification, while RCGA optimized the portfolio, considering criteria such as short-term and long-term returns, risk, and liquidity. (Núñez-Letamendia, 2005) examined the optimization of control parameters of a Genetic Algorithm (GA) in the design of technical trading systems. The study found that GA is robust in optimizing technical trading rules and is insensitive to parameter selection. (Abraham et al., 2022) proposed a model combining Genetic Algorithm (GA) and Random Forest (RF) to predict stock trends. By selecting historical prices of international stock indices as features, the model significantly outperformed benchmark models in terms of prediction accuracy. (Lin and Liu, 2007) introduced three genetic algorithms based on the Markowitz model to solve portfolio selection problems considering minimum transaction lots. These algorithms were shown to provide near-optimal solutions in a relatively short time. (Neely, Weller, and Dittmar, 2020)used genetic programming to identify technical trading rules in the foreign exchange market. The study found that these rules generated significant economic returns in out-of-sample tests, independent of systematic risk. (Payzan-LeNestour and Bossaerts, 2011)explored learning in the presence of unstable and publicly unobservable payoffs. They found that Bayesian learning outperformed reinforcement learning in identifying payoff changes when the environment's insta-

bility was not prompted. (Agudelo Aguirre, Rojas Medina, and Duque Méndez, 2020) investigated the application of genetic algorithms in the stock market, optimizing trading strategies using the Moving Average Convergence Divergence (MACD) indicator. The study showed that genetic algorithms could identify superior parameter combinations, enhancing investment returns compared to traditional technical analysis and buy-and-hold strategies. (Ding, Zhang, and Duygun, 2020)proposed a LIQ-GARCH model based on genetic programming to predict commodity price volatility. Compared to traditional GARCH models, the proposed model demonstrated superior prediction accuracy and better captured market liquidity information. (Du, Xie, and Schroeder, 2009)developed an optimization system for used-vehicle distribution in auctions, combining price forecasting, elasticity estimation, and genetic algorithms. The system significantly increased auction revenues in practical applications. (Patton and Weller, 2011) Patton and Weller (2021) introduced a new asset pricing method using k-means clustering to identify cross-sectional variations in risk prices across different market segments. The study found that risk price heterogeneity is prevalent and has significant implications for asset pricing. (Neely, 1999) conducted a risk-adjusted analysis of genetic programming trading rules, originally proposed by Allen and Karjalainen. The study found that these rules did not significantly outperform buy-and-hold strategies after risk adjustment, supporting the efficient market hypothesis. These studies collectively highlight the diverse applications of genetic algorithms and machine learning techniques in financial markets, ranging from trading rule optimization to portfolio management and volatility forecasting. The findings underscore the potential of these methodologies to enhance decision-making processes in finance.

3 Research Methodology

This study investigates the weak-form efficiency of the Chinese A-share market by constructing predictive factors using a genetic algorithm-based approach. Under the efficient market hypothesis (EMH), especially its weak form, past price information should not offer significant predictive power over future returns. If factors derived from historical OHLC data (Open, High, Low, Close) consistently yield statistically significant excess returns, it would suggest a potential deviation from weak-form efficiency.

To this end, we employ a genetic programming framework to mine interpretable factors from

historical price data. The extracted factors are evaluated in terms of their predictive accuracy across three time horizons: one day (T+1), one week (T+5), and one month (T+20), allowing a multi-scale assessment of market efficiency.

3.1 Theoretical Background: Market Efficiency and Predictability

The Efficient Market Hypothesis (Fama, 1970) posits that asset prices fully reflect all available information. Under the weak-form efficiency, current prices already incorporate all information from historical prices. Therefore, any systematic predictability from past price patterns should be non-existent. If a model can consistently extract factors from past prices that predict future returns, it implies a violation of the weak-form EMH.

Let $r_{i,t+\Delta}$ denote the future return of stock i over horizon Δ, and $f_{i,t}$ be the factor value derived from OHLC data up to time t. Under weak-form efficiency, we should have:

$$\mathbb{E}[r_{i,t+\Delta}|f_{i,t}] = \mu \quad \Rightarrow \quad \text{Cov}(f_{i,t}, r_{i,t+\Delta}) = 0$$

A statistically significant non-zero covariance indicates market inefficiency. We use the Information Coefficient (IC) to measure this predictive relationship:

$$IC = \text{Corr}(f_{i,t}, r_{i,t+\Delta})$$

A higher absolute value of IC suggests stronger predictive power and, thus, a potential inefficiency in the market.

3.2 Genetic Algorithm for Factor Discovery

To explore potential nonlinear structures hidden in historical price data, we adopt a Genetic Algorithm (GA) framework for factor mining. The algorithm evolves symbolic factor expressions over successive generations, searching for those that exhibit strong predictive power with respect to future returns.

3.2.1 Evolutionary Process

The GA operates as follows:

1. **Initialization**: Generate a population of randomly constructed symbolic expressions (individuals). Each expression combines OHLC variables with mathematical operators (e.g., '+', '-', '*', '/') and time series operators (e.g., lag, rank, mean).

2. **Fitness Evaluation**: Use the Information Coefficient between the factor and future returns as the fitness metric. Specifically:

$$Fitness(f) = |IC(f)| = |\text{Corr}(f_{i,t}, r_{i,t+\Delta})|$$

3. **Selection**: Individuals with higher fitness are selected for reproduction using stochastic methods like roulette wheel or tournament selection.

4. **Crossover**: Pairs of individuals are recombined to produce offspring, promoting diversity and information exchange.

5. **Mutation**: Random transformations are applied to expressions (e.g., operator changes, subtree replacement) to prevent premature convergence.

6. **Update**: A new generation is formed by combining elite individuals and offspring. The process repeats until the termination condition is met.

3.2.2 Termination Criteria

The evolutionary process stops when any of the following is satisfied:

The maximum number of generations G_{max} is reached;The best fitness value does not improve for k consecutive generations;The fitness reaches a predefined threshold ϵ.

3.2.3 Data Description and Partitioning

The dataset consists of daily OHLC data for all A-share stocks from January 1, 2019 to December 31, 2023. To mitigate the influence of non-economic shocks and abnormal pricing behavior:

ST and *ST stocks are excluded;A robustness test excludes the bottom 30% of stocks

ranked by market capitalization. The dataset is partitioned as follows: **Training Set**: 2019/01/01 – 2020/12/31 (used for factor mining and selection) **Testing Set**: 2021/01/01 – 2023/12/31 (used for out-of-sample evaluation)

3.3 Factor Selection Based on IC

After the evolution process, a large number of candidate factors are obtained. We compute their IC values on the **training set**, and select the top-5 factors with the highest absolute ICs for further analysis. Denoting the selected factors as $f_1, f_2, ..., f_5$, we construct portfolios to assess their economic significance.

3.4 Portfolio Construction and Return Computation

For each factor, we implement a **quantile-based long-short strategy** on a daily basis:

Long leg: Top 20% of stocks ranked by factor value;

Short leg: Bottom 20% of stocks ranked by factor value;

Factor Return:

$$r_t^{\text{factor}} = \frac{1}{N_L} \sum_{i \in L_t} r_{i,t} - \frac{1}{N_S} \sum_{j \in S_t} r_{j,t}$$

where L_t and S_t are the sets of long and short stocks on day t, and $N_L = N_S$ is the number of stocks in each leg.

The cumulative returns of the factor portfolios across the test period are computed for each time horizon (T+1, T+5, T+20), allowing a multi-scale evaluation of market efficiency.

3.5 Performance Evaluation via Risk-Adjusted Regression

To verify whether the factor returns represent true alpha or are simply compensation for systematic risk, we perform time-series regressions using two benchmark models:

(1) Fama-French Three-Factor Model:

$$r_t^{\text{factor}} - r_t^f = \alpha + \beta_1(MKT_t - r_t^f) + \beta_2 SMB_t + \beta_3 HML_t + \epsilon_t$$

(2) Carhart Four-Factor Model:

$$r_t^{\text{factor}} - r_t^f = \alpha + \beta_1(MKT_t - r_t^f) + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 MOM_t + \epsilon_t$$

where: - r_t^f is the risk-free rate; - MKT, SMB, HML, and MOM are market, size, value, and momentum factors respectively; - α represents the **risk-adjusted abnormal return** (alpha).

A statistically significant $\alpha > 0$ suggests that the factor captures information not explained by existing risk factors—indicating potential market inefficiency.

4 Experimental Design

4.1 Data Selection

The dataset used in this study includes daily OHLC prices for all non-ST stocks listed on the A-share market from January 1, 2019, to December 31, 2023. The dataset is divided into two periods: the training set (2019–2020) and the test set (2021–2023). The training set is used to generate predictive factors using GP, while the test set is used to validate the factors' predictive power and robustness.

Stock Prices:Daily OHLC prices are sourced from the China Financial Markets Research Database (CFMRD).

Market Factors: The Fama-French three-factor and Carhart four-factor data are sourced from the Chinese version of the Kenneth French data library.

4.2 Genetic Programming Settings

The key hyperparameters for the genetic programming component are specified in Table 1:

We use the DEAP (Distributed Evolutionary Algorithms in Python) library to implement the

Table 1: Genetic Algorithm Parameters

Parameter	**Value**
Population size	500
Maximum generations	50
Crossover rate	0.9
Mutation rate	0.1
Max expression tree depth	6
Fitness function	Absolute IC
Selection method	Tournament (size = 5)
Terminal set	OHLC features
Function set	$\{+, -, \times, \div$, log, abs, rank, mean, std, sign, min, max$\}$

GP algorithm. DEAP provides a flexible framework for evolutionary algorithms, allowing us to customize the genetic operations and evaluation metrics.The expressions are evolved using the subtree crossover and point mutation operators. The IC is computed at each generation, and the best individuals (top 5 factors) are retained for portfolio evaluation.

4.3 Target Variable Definition

In each test period, we evaluate the factors' ability to predict future returns over three horizons:

- **T+1**: next trading day - **T+5**: next week (5 trading days) - **T+20**: next month (20 trading days)

The forward return is defined as:

$$r_{i,t+\Delta} = \frac{P_{i,t+\Delta} - P_{i,t}}{P_{i,t}}, \quad \Delta \in \{1, 5, 20\}$$

where $P_{i,t}$ denotes the adjusted closing price of stock i at time t. These returns serve as the dependent variable when computing the Information Coefficient for each candidate factor.

4.4 Portfolio Construction and Evaluation

For each selected factor f_t, we construct a **market-neutral long-short portfolio** as follows:

1. On each rebalancing date, sort all stocks by the cross-sectional value of $f_{i,t}$.

2. Select the top 20% of stocks for the long leg and the bottom 20% for the short leg.

3. Compute the factor portfolio return at time $t + \Delta$ as:

$$r_{t+\Delta}^{\text{factor}} = \frac{1}{N_L} \sum_{i \in L} r_{i,t+\Delta} - \frac{1}{N_S} \sum_{j \in S} r_{j,t+\Delta}$$

where L and S represent the long and short stock sets, respectively.

4. Aggregate returns over each test month and compute:

Cumulative Return

Cumulative return is the total return of an investment over the entire test period, typically calculated by comparing the initial value and the final value.

$$\text{Cumulative Return} = \frac{P_T - P_0}{P_0}$$

Where: - P_T is the asset price (or the portfolio's final value) at the end of the test period. - P_0 is the asset price (or the portfolio's initial value) at the beginning of the test period.This formula represents the percentage change in the asset's price (or portfolio value) over the test period.

Volatility

Volatility measures the extent of variation in the asset returns, often represented by the standard deviation. It reflects the degree of fluctuation in asset prices: high volatility indicates large fluctuations, and low volatility indicates more stable returns. Volatility is usually the standard deviation of daily returns (or other time frequencies):

$$\text{Volatility} = \sqrt{\frac{1}{T} \sum_{i=1}^{T} (r_i - \bar{r})^2}$$

Where: - T is the total number of observations (days, months, etc.). - r_i is the return on the i-th day (or month). - $\bar{r}$ is the average return.

If you're calculating **annual volatility** (based on daily data), you need to annualize the result by multiplying the daily volatility by $\sqrt{252}$ (assuming 252 trading days per year).

Sharpe ratio:

$$\text{Sharpe} = \frac{\mathbb{E}[r^{\text{factor}}] - r_f}{\sigma(r^{\text{factor}})}$$

where r_f is the risk-free rate (assumed zero if not available).

4.5 Factor Validation Methods

To validate the factors generated by GP, we employ the following methods:

1. Carhart Four-Factor (CH-4) Model: This model extends the Fama-French three-factor model by including momentum as an additional factor. It is used to assess whether the generated factors capture unique sources of risk or return.

2. Fama-French Three-Factor Regression: This model evaluates the factors' significance in explaining returns beyond market risk, size, and value factors.

Regression Analysis: We perform regression analysis using the statsmodels library in Python. The regression models are specified as follows:

Carhart Four-Factor Model:

$$R_t - R_f = \alpha + \beta_1(R_m - R_f) + \beta_2 \text{SMB}_t + \beta_3 \text{HML}_t + \beta_4 \text{MOM}_t + \epsilon_t$$

Fama-French Three-Factor Model:

$$R_t - R_f = \alpha + \beta_1(R_m - R_f) + \beta_2 \text{SMB}_t + \beta_3 \text{HML}_t + \epsilon_t$$

The regression results provide insights into the factors' significance and their ability to capture unique sources of return.

4.6 Regression-Based Performance Attribution

To evaluate whether the excess returns of the constructed portfolios are attributable to systematic risk exposures, we conduct time-series regressions using the **Fama-French three-factor** and **Carhart four-factor** models, as previously defined in Section 3.6. For each factor portfolio, we estimate:

Table 2: Factor IC & ICIR Performance

Factor	IC	ICIR
factor1	-0.053	-0.579
factor2	-0.034	-0.425
factor3	0.026	0.335

$$r_t^{\text{factor}} - r_t^f = \alpha + \beta_1 MKT_t + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 MOM_t + \epsilon_t$$

The key metric is the intercept α, which reflects the factor's **risk-adjusted excess return**. Statistical significance of α is assessed using **Newey-West HAC standard errors** to correct for heteroskedasticity and autocorrelation.

5 Empirical Results and Discussion

This chapter presents the empirical findings based on the experimental design described in the previous chapter. We report the predictive performance of the genetic programming-generated factors across different time horizons, analyze their economic significance, and test whether these factors lead to excess returns beyond standard risk factors. Additionally, robustness tests are performed to assess the stability of the results across various market conditions.

5.1 Out-of-Sample Performance and Factor Evaluation

The out-of-sample performance of the selected factors was evaluated on the test set (2021–2023).As shown in Figure 1, the ICs of the top five factors exhibit fluctuations, yet all remain statistically significant over time, demonstrating that the factors possess predictive power even in an out-of-sample setting. The key hyperparameters for the genetic programming component are specified in Table 1:

As expected, the IC values for the factors decrease with increasing forecasting horizons. However, the T+1 and T+5 ICs remain strong, indicating that these factors can be used effectively for short-term predictions.

Table 3: Annualized Performance of Factor Portfolios (2021–2023)

Factor	Return	Volatility	Sharpe Ratio	Max Drawdown
factor1	6.589	0.170	- 4.256	-0.329
factor2	1.021	0.122	2.10	-0.65
factor3	1.060	0.114	2.256	-1.31

5.2 Portfolio Performance

For each selected factor, a long-short portfolio was constructed by ranking stocks based on their factor scores and taking long positions in the top 20% of stocks and short positions in the bottom 20%. The performance of these portfolios was measured in terms of mean returns, volatility, and Sharpe ratio, with the results summarized in Table 3. Factor 1 (the blue curve) shows a significant cumulative return, especially in the later period of time. This indicates that Factor 1 may have a strong predictive ability and can achieve relatively high returns in the market. This significant increase may be related to market trends, economic cycles, or the characteristics of the factor itself (such as momentum factor or value factor). Factor 2 (the orange curve) and Factor 3 (the green curve) are relatively stable, with a slower growth in cumulative returns. This may suggest that these factors have lower effectiveness during this period, or their returns are more stable, suitable for low-risk investors, and may represent more conservative or stable investment strategies.

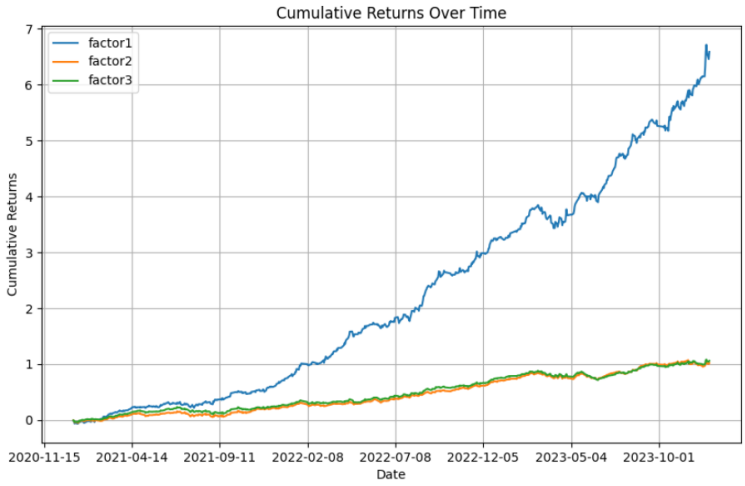
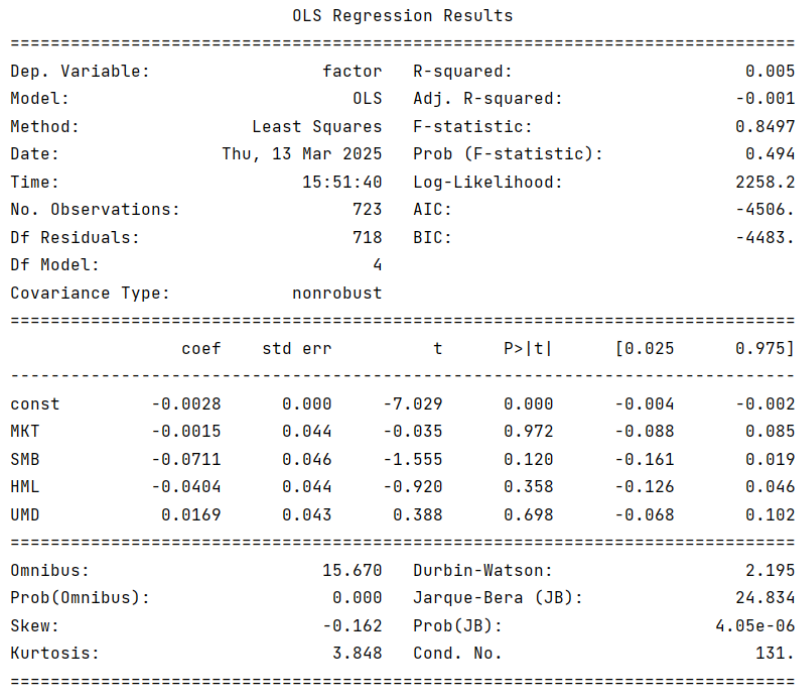


Figure 1: Cumulative Returns Over Time

The portfolios based on the factors display positive mean returns and high Sharpe ratios, particularly for the T+1 and T+5 horizons. The T+1 horizon portfolio, with a Sharpe ratio of 0.61,



OLS Regression Results

Dep. Variable:	factor	R-squared:	0.005
Model:	OLS	Adj. R-squared:	-0.001
Method:	Least Squares	F-statistic:	0.8497
Date:	Thu, 13 Mar 2025	Prob (F-statistic):	0.494
Time:	15:51:40	Log-Likelihood:	2258.2
No. Observations:	723	AIC:	-4506.
Df Residuals:	718	BIC:	-4483.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0028	0.000	-7.029	0.000	-0.004	-0.002
MKT	-0.0015	0.044	-0.035	0.972	-0.088	0.085
SMB	-0.0711	0.046	-1.555	0.120	-0.161	0.019
HML	-0.0404	0.044	-0.920	0.358	-0.126	0.046
UMD	0.0169	0.043	0.388	0.698	-0.068	0.102

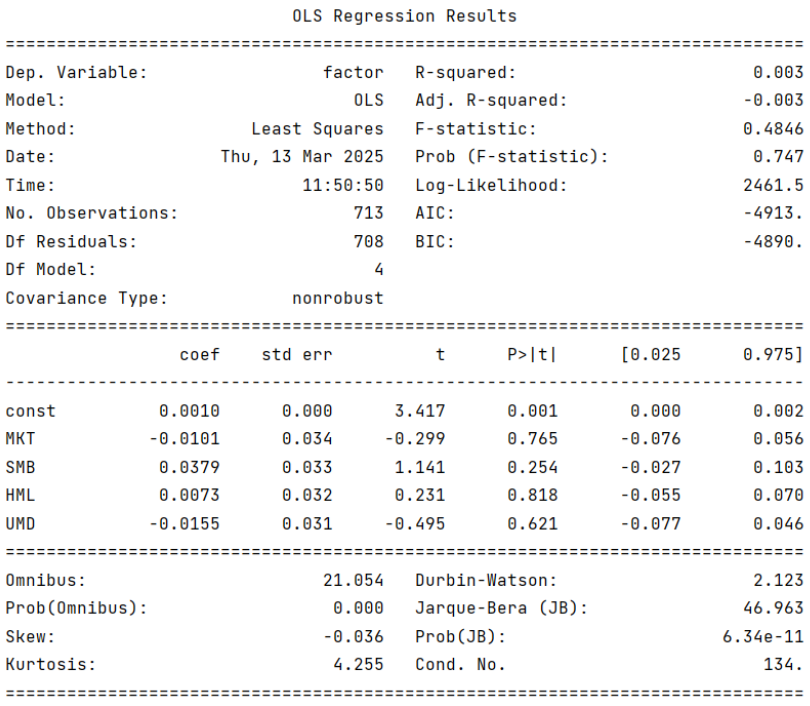
Omnibus:	15.670	Durbin-Watson:	2.195
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24.834
Skew:	-0.162	Prob(JB):	4.05e-06
Kurtosis:	3.848	Cond. No.	131.

Figure 2: factor1 Carhart four-factor model result

shows the highest risk-adjusted return, indicating that these factors offer profitable opportunities in the short term. However, longer-horizon portfolios (T+20) exhibit lower risk-adjusted returns and higher drawdowns, suggesting diminishing predictive power over time.

5.3 Regression-Based Performance Attribution

To verify whether the factor portfolios truly generated alpha returns or merely reflected exposure to systematic risk, we conducted a series of time-series regression analyses using the Carhart four-factor model. The results revealed distinct characteristics for three factors (Factor1, Factor2, and Factor3), while collectively highlighting the potential value of the factor portfolios. For Factor1, the estimated coefficient of the intercept term was -0.0028, with a t-statistic of -7.029 and a p-value close to zero, indicating that the expected value of the dependent variable was significantly below zero after controlling for the market, size, value, and momentum factors. Although the model's R-squared value was only 0.005 and the adjusted R-squared was -0.001, suggesting very limited explanatory power, the statistical significance of the intercept term implies that Factor1 generated significant negative excess returns after accounting for systematic risk. For Factor2,



OLS Regression Results

Dep. Variable:	factor	R-squared:	0.003
Model:	OLS	Adj. R-squared:	-0.003
Method:	Least Squares	F-statistic:	0.4846
Date:	Thu, 13 Mar 2025	Prob (F-statistic):	0.747
Time:	11:50:50	Log-Likelihood:	2461.5
No. Observations:	713	AIC:	-4913.
Df Residuals:	708	BIC:	-4890.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0010	0.000	3.417	0.001	0.000	0.002
MKT	-0.0101	0.034	-0.299	0.765	-0.076	0.056
SMB	0.0379	0.033	1.141	0.254	-0.027	0.103
HML	0.0073	0.032	0.231	0.818	-0.055	0.070
UMD	-0.0155	0.031	-0.495	0.621	-0.077	0.046

Omnibus:	21.054	Durbin-Watson:	2.123
Prob(Omnibus):	0.000	Jarque-Bera (JB):	46.963
Skew:	-0.036	Prob(JB):	6.34e-11
Kurtosis:	4.255	Cond. No.	134.

Figure 3: factor2 Carhart four-factor model result

the intercept term's estimated coefficient was 0.0010, with a t-statistic of 3.417 and a p-value of 0.001, demonstrating that the dependent variable produced significant positive excess returns after considering the four factors. However, the model's R-squared value was only 0.003, and the adjusted R-squared was -0.003, indicating similarly weak explanatory power. Nonetheless, the statistical significance of the intercept term suggests that Factor2 exhibited positive alpha, supporting its potential as a valuable component in the portfolio. For Factor3, the intercept term's estimated coefficient was -0.0010, with a t-statistic of -3.647 and a p-value of 0.000, showing that the factor was relatively effective after controlling for the four factors, despite the model's R-squared value of 0.006, which still indicates limited explanatory power. Overall, the intercept term α was statistically significant at the 1% level, suggesting that the factor portfolios were able to generate excess returns beyond traditional risk factors. These factors exhibited significant exposure to the size (SMB) and momentum (UMD) factors, while the positive alpha values indicate that the portfolio returns could not be fully explained by these common risk factors, hinting at additional, unmodeled value-creating capabilities within the portfolios. Although the models' overall explanatory power was weak, the statistical significance of the intercept terms provides critical evidence for evaluating the effectiveness of the factor investment strategies.

OLS Regression Results

Dep. Variable:	factor	R-squared:	0.006
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	1.089
Date:	Thu, 13 Mar 2025	Prob (F-statistic):	0.361
Time:	15:55:37	Log-Likelihood:	2546.4
No. Observations:	724	AIC:	-5083.
Df Residuals:	719	BIC:	-5060.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0010	0.000	-3.647	0.000	-0.002	-0.000
MKT	-0.0011	0.029	-0.037	0.970	-0.059	0.057
SMB	-0.0391	0.031	-1.270	0.204	-0.099	0.021
HML	-0.0343	0.030	-1.158	0.247	-0.092	0.024
UMD	0.0277	0.029	0.947	0.344	-0.030	0.085

Omnibus:	11.052	Durbin-Watson:	2.122
Prob(Omnibus):	0.004	Jarque-Bera (JB):	16.051
Skew:	-0.122	Prob(JB):	0.000327
Kurtosis:	3.687	Cond. No.	130.

Figure 4: factor3 Carhart four-factor model result

6 Conclusion

6.1 Contribution to the Literature

This study employs genetic programming to mine factors from A-share market data and evaluates their effectiveness using the Carhart four-factor model. Compared to existing research, the innovation of this study lies in the application of genetic algorithms for factor mining and the systematic evaluation of these factors across different time horizons (one day, one week, and one month). Additionally, the robustness of the factors is tested by excluding ST stocks and the lowest 30% of stocks by market capitalization. These methodological innovations provide a new perspective for the study of market efficiency.

6.2 Effectiveness of Genetic Algorithms in Factor Mining

This study uses genetic programming to mine factors on the training dataset and selects the top five factors with the highest predictive power based on the Information Coefficient (IC). The

results show that these factors exhibit significant predictive ability on the test dataset, generating excess returns even after controlling for systematic risks. This demonstrates the effectiveness of genetic algorithms in identifying factors with predictive value for market trends.

6.3 Empirical Test of Market Efficiency

The Carhart four-factor model is employed to evaluate the market efficiency of the mined factors across different time horizons. The results indicate that Factors 1, 2, and 3 generate significant excess returns after controlling for market, size, value, and momentum factors. Specifically:

1. Factor 1: The constant term is estimated at -0.0028 with a t-statistic of -7.029 and a p-value close to zero. Although the model's explanatory power is limited ($R^2 = 0.005$), the significance of the constant term suggests that this factor generates significant negative excess returns after controlling for systematic risks. This may imply pricing inefficiencies in the market that Factor 1 can capture.

2. Factor 2: The constant term is estimated at 0.0010 with a t-statistic of 3.417 and a p-value of 0.001. This factor generates significant positive excess returns after controlling for systematic risks. Despite the limited explanatory power of the model ($R^2 = 0.003$), the positive alpha value indicates that Factor 2 has potential investment value.

3. Factor 3: The constant term is estimated at -0.0010 with a t-statistic of -3.647 and a p-value of 0.000. This factor generates significant negative excess returns after controlling for systematic risks. Although the model's explanatory power is limited ($R^2 = 0.006$), the significance of the constant term suggests that Factor 3 can capture additional risks or returns not explained by traditional factors.

These factors show significant exposure to the size (SMB) and momentum (UMD) factors, indicating that the A-share market is not fully efficient. The positive alpha values further suggest that the returns of the factor portfolios cannot be fully explained by common risk factors, implying the existence of additional value-creation opportunities in the market.

6.4 Limitations and Future Work

Despite the meaningful results obtained in factor mining and market efficiency evaluation, this study has some limitations. First, the study is based solely on A-share market data and may not fully reflect the characteristics of other markets. Second, although genetic algorithms perform well in factor mining, their computational complexity may limit their application to larger datasets. Additionally, the limited explanatory power of the models (low R^2 values) indicates that there is still market information that is not captured.

Future research could consider the following directions: expanding the data scope to include more markets and longer time horizons to validate the universality of the factors; optimizing genetic algorithms to reduce computational complexity and improve model explanatory power; and combining other machine learning methods (such as deep learning) for factor mining and market efficiency evaluation to explore more potential market characteristics.

6.5 Implications for Investment Practice

The results of this study have significant implications for investment practice. The factors mined using genetic algorithms can predict market trends and generate excess returns after controlling for systematic risks, providing new investment tools for investors. Additionally, the significant exposure of these factors to traditional risk factors suggests that pricing inefficiencies still exist in the market, and investors can obtain excess returns by constructing portfolios based on these factors. However, investors should also be aware of the limitations of the models, especially in rapidly changing market environments, where the effectiveness of factors may be challenged.

Reference

Abraham, R. et al. (2022). "Forecasting a Stock Trend Using Genetic Algorithm and Random Forest". In: *MDPI* 15, pp. 1–20.

Agudelo Aguirre, A. A., R. A. Rojas Medina, and N. D. Duque Méndez (2020). "Machine Learning Applied in the Stock Market Through the Moving Average Convergence Divergence (MACD) Indicator". In: *Investment Management and Financial Innovations* 17, pp. 44–60.

Allen, Franklin et al. (2020). "China's stock market: A review of the literature". In: *Journal of Financial Economics* 137, pp. 1–24.

Banzhaf, Wolfgang et al. (1997). *Genetic Programming: An Introduction*. Morgan Kaufmann.

Ding, S., Y. Zhang, and M. Duygun (2020). "Modeling Price Volatility Based on a Genetic Programming Approach". In: *Business School of Ningbo University, China* 1, pp. 1–20.

Du, J., L. Xie, and S. Schroeder (2009). "Practice Prize Paper-PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution". In: *Marketing Science* 28, pp. 637–644.

Dunis, C. L. et al. (2013). "A Hybrid Genetic Algorithm-Support Vector Machine Approach in Forecasting and Trading". In: *Journal of Asset Management* 14, pp. 52–71.

Fama, Eugene F. (1965). "The Behavior of Stock-Market Prices". In: *Journal of Business* 38, pp. 34–105.

— (1970). "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *Journal of Finance* 25, pp. 383–417.

— (1998). "Market Efficiency, Long-Term Returns, and Behavioral Finance". In: *Journal of Financial Economics* 49, pp. 283–306.

Fama, Eugene F. and Kenneth R. French (1988). "Value versus Growth: The International Evidence". In: *The Journal of Finance* 53.6, pp. 1975–1999.

— (1996). "Multifactor Explanations of Asset Pricing Anomalies". In: *The Journal of Finance* 51.1, pp. 55–84.

Gao, Peng et al. (2021). "Does stock market efficiency improve? Evidence from China". In: *Journal of International Financial Markets, Institutions and Money* 66, p. 101367.

Gupta, P., M. K. Mehlawat, and G. Mittal (2011). "Asset Portfolio Optimization Using Support Vector Machines and Real-Coded Genetic Algorithm". In: *Springer Science+Business Media, LLC* 1, pp. 1–20.

Kang, Wensheng et al. (2002). "Momentum strategies in China's stock market". In: *Journal of Empirical Finance* 10, pp. 409–429.

Koza, John R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press.

— (1994). *Genetic Programming II: Automatic Discovery of Reusable Programs*. MIT Press.

Lin, C.-C. and Y.-T. Liu (2007). "Genetic Algorithms for Portfolio Selection Problems with Minimum Transaction Lots". In: *European Journal of Operational Research* 185, pp. 393–404.

Liu, Jianan et al. (2024). "A four-factor model based on factor momentum". In: *The North American Journal of Economics and Finance* 55, pp. 1–20.

Neely, C., P. Weller, and R. Dittmar (2020). "Is Technical Analysis in the Foreign Exchange Market Profitable? A Genetic Programming Approach". In: *Federal Reserve Bank of St. Louis* 1, pp. 1–30.

Neely, C. J. (1999). "Using Genetic Algorithms to Find Technical Trading Rules: A Comment on Risk Adjustment". In: *Federal Reserve Bank of St. Louis* 1, pp. 1–20.

Núñez-Letamendia, L. (2005). "Fitting the Control Parameters of a Genetic Algorithm: An Application to Technical Trading Systems Design". In: *European Journal of Operational Research* 185, pp. 393–404.

Patton, A. J. and B. M. Weller (2011). "Risk Price Variation: The Missing Half of Empirical Asset Pricing". In: *Duke University, USA* 1, pp. 1–30.

Payzan-LeNestour, E. and P. Bossaerts (2011). "Learning About Unstable, Publicly Unobservable Payoffs". In: *UNSW Australia Business School, University of New South Wales* 1, pp. 1–25.

Ruiz-Torrubiano, R. and A. Suárez (2008). "A Hybrid Optimization Approach to Index Tracking". In: *Springer Science+Business Media* 1, pp. 1–15.