

**Introduction to FinTech**

**Data Analysis Project 1**

**Predicting Stock Returns Using Machine Learning**

|  |  |
| --- | --- |
|  | |
| Kai Ren | 2401212437 |
| WeiJie Zhang | 2401212474 |
| Yang Lan | 2401212401 |
| YiMing Jiang |  |
| Rui Hu | 2401212392 |

**Peking University**

2024.11.4

Content

[Abstract 1](#_Toc181539302)

[1. Overall Structure 2](#_Toc181539303)

[2. Data Collection 3](#_Toc181539304)

[3. Factor Engineering 5](#_Toc181539305)

[Daily trading factors 5](#_Toc181539306)

[High frequency factors 7](#_Toc181539307)

[Fundamental factors 9](#_Toc181539308)

[Macroeconomic factors 11](#_Toc181539309)

[4. Model 12](#_Toc181539310)

[Select and Normalize 12](#_Toc181539311)

[In-Sample and Out-Sample Modeling 13](#_Toc181539312)

[5. Backtesting 19](#_Toc181539313)

[5.1 Backtesting Process 19](#_Toc181539314)

[5.2 Backtesing results 20](#_Toc181539315)

[6. Factors Importance analysis 22](#_Toc181539316)

[6.1 Which factors matter? 22](#_Toc181539317)

[6.2 Result Analysis 24](#_Toc181539318)

# Abstract

For Data Analysis Project 1, we built a research framework for stock return forecasting using machine learning models to extract factor features end-to-end. The 500 constituent stocks in the CSI 500 (CSI 500) are modeled, and their 5.5-year trading data from January 1, 2019 to June 1, 2024 are collected, including daily and minute volume and price data, company fundamentals, and macroeconomic data.

Based on these raw data, we study and construct 167 factors, including 26 daily trading factors, 64 high-frequency factors, 66 fundamental factors, and 11 macroeconomic factors. After eliminating the samples with serious missing factors, the factors are standardized using the Z-Score method on the cross-section.

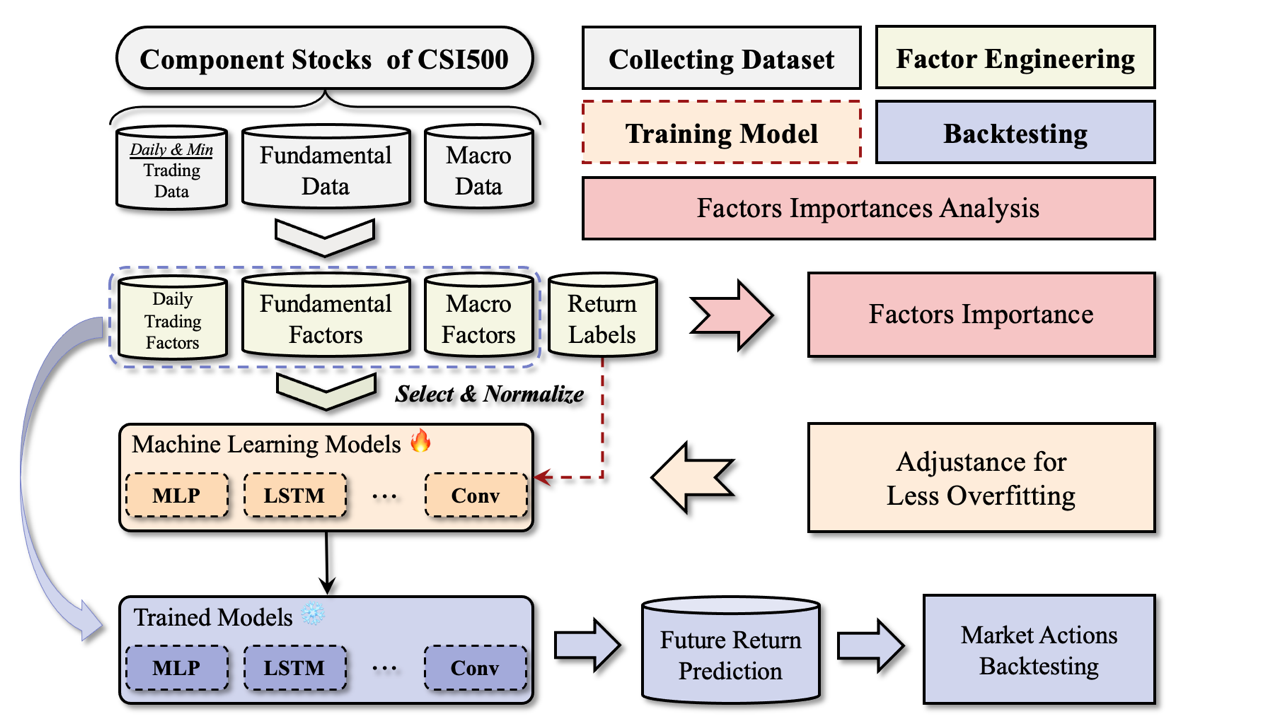
Based on this, we build the classical time-series neural network architectures such as MLP, LSTM, GRU, Transofrmer, and Conv to extract the factor features, which are used to predict the daily return of the opening price on this day. Under the premise that all data are used as training data, both for regression modeling and classification modeling, we obtain 100% of the theoretically derivable ACC and F1 within the sample, which validates the correctness of our modeling framework.

Subsequently, we divided the training samples into three parts: training set, validation set, and test set (January 1, 2024 to June 1, 2024) according to time intervals to simulate real modeling scenarios in the industry. Based on this, we found obvious overfitting phenomenon and tried to adjust it in several aspects, such as adjusting the model architecture and introducing the lag period factor.

Further, we constructed simple strategies and backtested them based on the predicted values over the test interval to obtain detailed returns. Finally, in order to better understand the financial implications of the factors in the modeling process, we used a tree model to provide a detailed analysis of factor importance.

# Overall Structure

For data Analysis Project 1, our framework is shown in Fig 1.1. The process includes data collection, factor engineering, model construction, backtesting and factor importance analysis.



**Fig 1.1.** The Framework of Data Analysis Project 1

# Data Collection

For this project, we use data from January 1, 2019 to June 30, 2024 for the CSI 500 constituent stocks. The CSI 500 consists of 500 stocks from the Shanghai and Shenzhen markets with medium market capitalization and good liquidity. The fact that it covers a wide range of industries and is balanced and diverse enough to represent the overall performance of the Chinese market is the reason why we made such a selection.

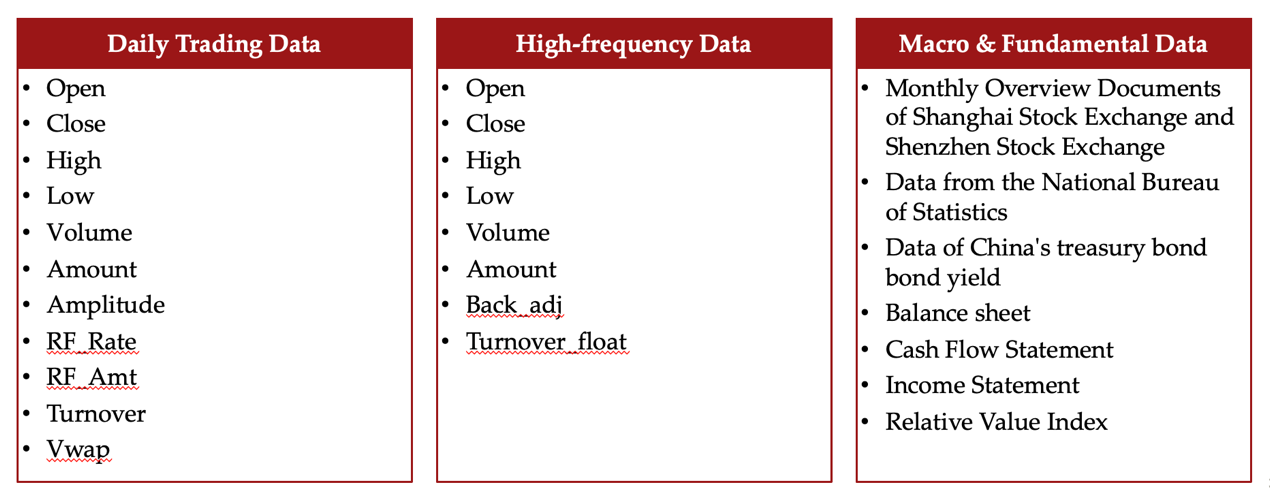
In terms of time dimension, we choose the past five and a half years as the study interval, which is because it is difficult to mine sufficiently valid information with too short a time interval; while if the time span is too large, the accuracy of the results will be challenged by the fluctuation of the data over time, and at the same time too large a volume of data will lead to a larger computational cost. In addition, considering the subsequent modeling needs, i.e., through the rolling training and forecasting methodology, the factor model is trained using time series of 1 year or a few months, and then forecasting the return of 1 month or a few weeks in the future. Taking the above considerations into account, we believe that the time horizon of January 1, 2019 to June 1, 2024 is very reasonable and can meet the demand for studying current market changes, and the conclusions obtained are highly current and accurate. Finally, considering that the CSI 500 constituents will be adjusted periodically, we chose the 500 constituents of the CSI 500 index on September 30, 2024 as the modeling object.

For data collection, Fig 2.1 shows the overall data included as source data. We first collected the daily frequency trading data of all stocks in the above time interval based on the AkShare packet, which totaled 606,475 entries, smaller than the product of the number of stock samples and the number of trading days, 656,000 entries, due to the fact that some stocks were suspended from trading on some of the trading days, which resulted in missing data. Among them, China Baoan (000009.sz) has the highest number of trading days with 1,312 days, and Kodak Manufacturing (600499.sh) has the lowest number of trading days with only 277 days.

Second, we collect trading data at the minute frequency within the time interval for the subsequent calculation of the high-frequency factor. Finally, we collect macro data including SSE SZSE monthly profile files, NBS data, and China government bond yields in the CSMAR database during the time interval to calculate the macro factor, and collect annual financial data of companies such as income statements, cash flow statements, balance sheets, and relative value indicators for the subsequent calculation of the fundamentals factor.

For the label construction, we use the daily return under the next day's opening price as the forecasting target, i.e.

Which aims to predict future stock price movements using factor data up to the present.

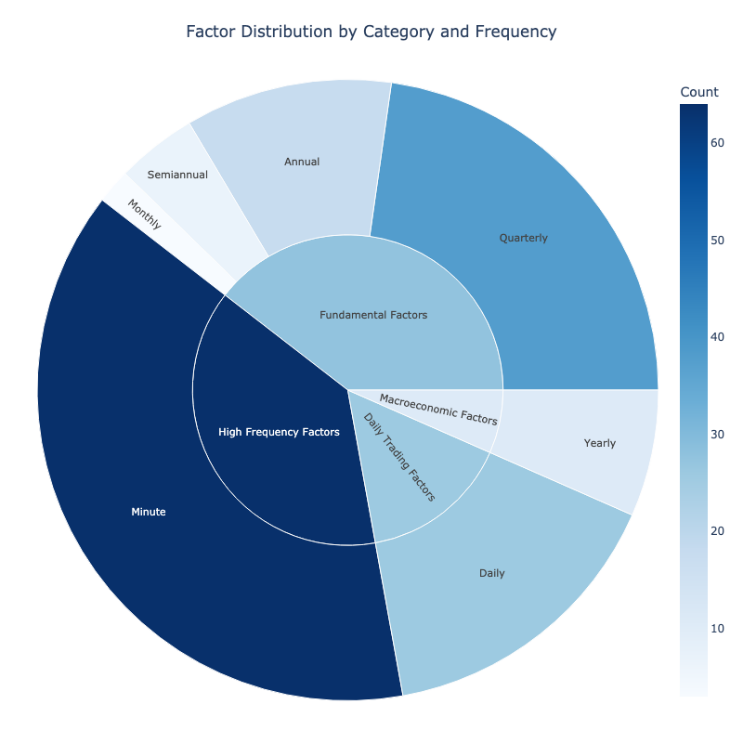


**Fig 2.1.** Overall data included as source data.

# Factor Engineering

Based on the collected data, we build a large collection of stock-level predictive characteristics with an extension of the variable definitions in the original papers listed in Leippoldet al. (2021). Fig 3.1 shows factor distribution. Our collection includes 167 characteristics in total, among which 26 are daily trading factors, 64 are high frequency factors, 66 are fundamental factors which have been documented in Leippoldet al. (2021), while 11 are macroeconomic factors identified macroeconomic situation. It is noteworthy that our data frequency is higher and factors categories are more complex compared with that in Leippoldet al. (2021), which may improve our prediction performance.

Table 1 in the Appendix provides a summary of all stock-level characteristics.



**Fig 2.1** Factor Distribution by Category and Frequency

## 

## Daily trading factors

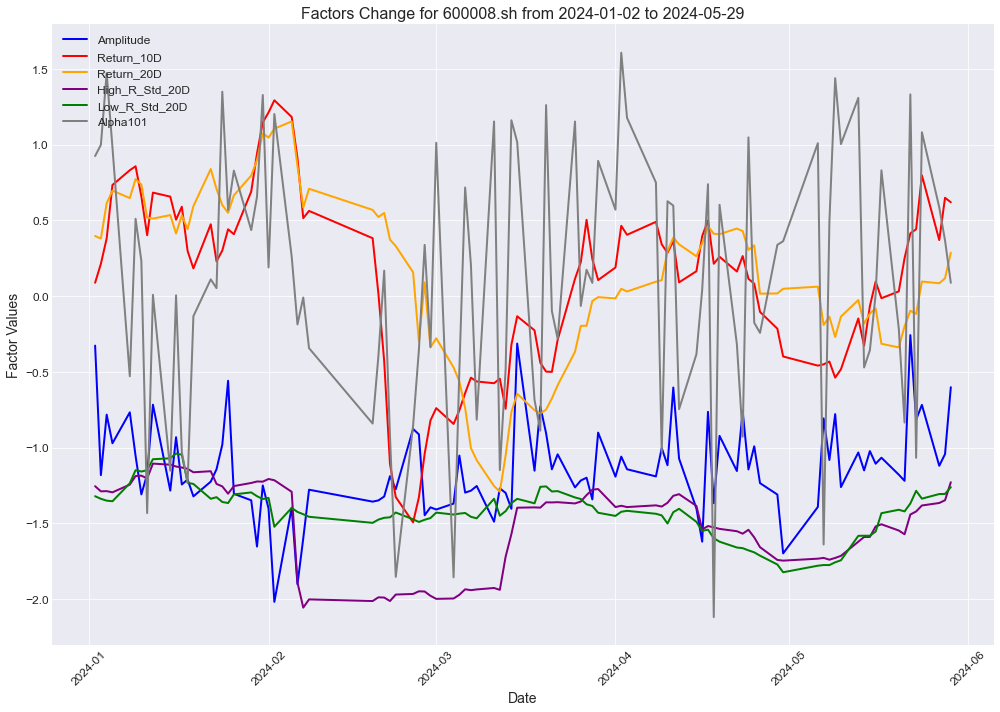
In this study, we developed a series of daily trading factors based on the primary stock price and volume data for the Chinese stock market. These factors, termed daily trading factors, capture essential market dynamics at a daily frequency and serve as the core building blocks for our machine learning model's input features.

The rationale behind constructing these factors lies in their ability to encapsulate market conditions and investor sentiment effectively. Factors derived from high-frequency price and volume data, such as daily returns, volatility measures, and turnover rates, often reflect the nuanced behavior of market participants. By including these factors in the model, we aim to capture not only the overall market direction but also subtle shifts that might indicate future price changes.

Here we list the details of some daily trading factors:

* **:** This factor represents the daily price range as a percentage of the previous close price, calculated as . Amplitude is a measure of volatility within the provides insights into the price fluctuations, which can be indicative of uncertainty or market momentum.
* and: These are moving averages of stock returns over the past 10 and 20 trading days. By averaging returns over these periods, we smooth out short-term fluctuations and capture medium-term trends.
* Standard Deviation Factors (e.g.,, ): These factors measure the 20-day rolling standard deviations of the high and low prices, respectively. By tracking the variability in daily highs and lows, these factors indicate the consistency (or lack thereof) in daily price ranges.
* : This factor, inspired by popular quant finance literature, incorporates daily HLOC prices to create a composite indicator that serves as a potential predictor of future returns. When Alpha101 approaches, it indicates strong buying pressure with closing prices near daily highs.

Fig2.2 shows an example of above factors change for stock 600008.SH from January 2024 to May 2024, revealing several notable patterns that align with the actual market performance. Most significantly, both the 10-day and 20-day return factors exhibited strong predictive power during the February 2024 rally, where they reached peak values (approximately 1.0-1.2) coinciding with the stock's breakthrough above the 5 RMB threshold. demonstrated superior sensitivity to short-term price movements, validating its utility for short-term trading signals.



**Fig 3.2** Daily Trading Factor Change for 600008.sh

## High frequency factors

High-frequency factors, derived from intraday trading data recorded at 5-minute intervals, provide insights into short-term market dynamics and investor behaviors that are often obscured in daily or lower-frequency data. The motivation for constructing these factors lies in capturing intra-day patterns, volatility, and liquidity fluctuations that are critical for short-term prediction and risk management.

The process of constructing high-frequency factors follows a systematic pipeline, consisting of: Raw Data → Basic Factor Construction → Operator Application. This pipeline is illustrated in Fig 3.3, and each step is discussed in detail below:

**Step 1: Basic Factor Construction**

Starting with raw 5-minute trading data, we define several basic factors that capture key aspects of price movement, spread, and liquidity at each interval. These basic factors include:

* **Imbalance**: Defined as , this metric reflects the relative price range and is used as a proxy for intraday volatility.
* **Price Spread**: Calculated as , the spread shows the variation within each 5-minute interval, offering insights into liquidity and price sensitivity.
* **Price Ratio**: Given by , this ratio provides an alternative view of the price spread, normalized by the high price.
* **VWAP (Volume-Weighted Average Price)**: Calculated as , VWAP represents the average price at which trading occurred and is often used as a benchmark for institutional trading performance.

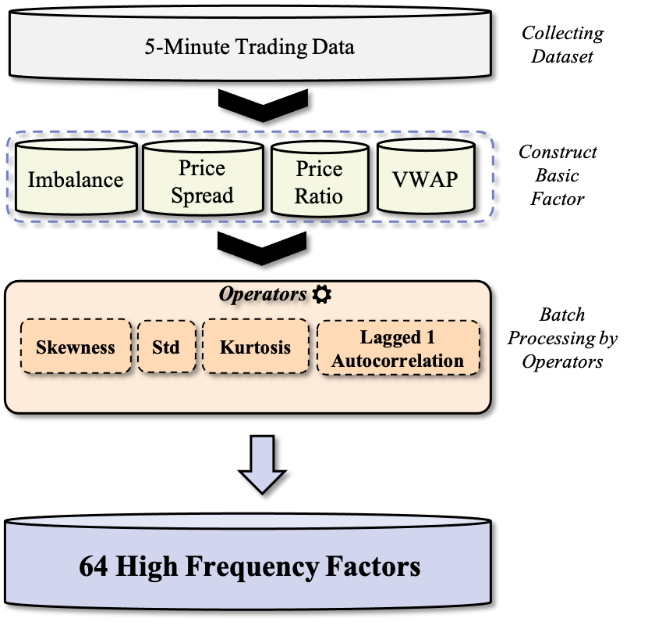
**Step 2: Application of Operators to Basic Factors**

Once the basic factors are constructed, we apply a series of operators to each of them. These operators are statistical functions and transformations designed to capture various distributional properties and temporal dependencies over the course of the trading day. Specifically, we apply the following operators:

* **Standard Deviation (std):** Measures the volatility of each basic factor over a specified time window, providing insights into the factor's variability.
* **Skewness:** Indicates the asymmetry in the distribution of values for each factor, highlighting any potential biases in intraday trading patterns.
* **Kurtosis:** Reflects the "tailedness" of the factor’s distribution, useful for identifying periods with extreme values or outliers.
* **Lagged 1 Autocorrelation:** Calculated across the whole day, morning, afternoon, and the difference between the morning and afternoon sessions, this operator captures the persistence or reversal tendencies in factor values.

By systematically applying each operator to the basic factors across these four intervals (whole day, morning, afternoon, and morning-afternoon difference), we generate a set of metrics that can describe each factor’s behavior throughout the day.

The combination of the basic factors and the statistical operators yields a robust set of 64 high-frequency factors.



**Fig 3.3** Pipeline for Constructing High Frequency Factors

## Fundamental factors

In our study, fundamental factors constitute a significant portion of the model's input features, accounting for around 40% of the total factors. These factors are derived from company financials, representing various aspects of a firm’s profitability, growth, leverage, and operational efficiency. By incorporating these factors, we aim to capture the impact of company-specific fundamentals on stock prices, allowing the model to distinguish between stocks based on their inherent value and growth potential. Moreover, since these factors often exhibit stability over time compared to market-based factors, they provide a steady baseline for analysis across economic cycles.

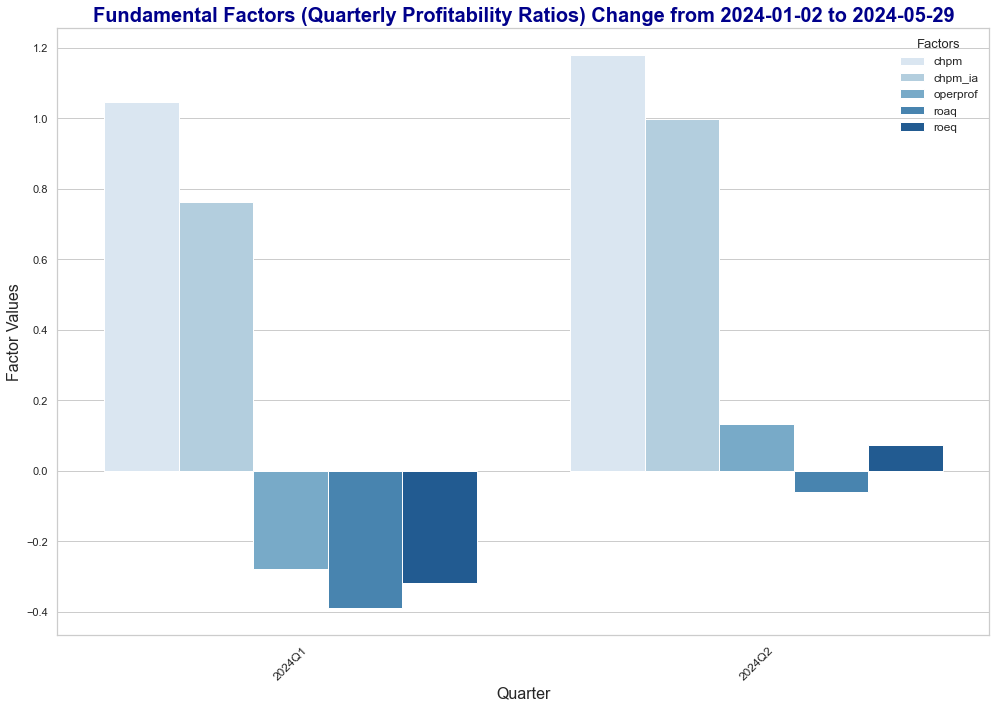
Here we list the details of some fundamental factors:

* (Accruals)*:* Based on Sloan (1996), this factor represents the quality of a company's earnings by measuring the extent to which cash flows differ from reported earnings. to construct acc, i.e.,

where represents the difference between two consecutive periods, denote current assets, cash/cash equivalents, current liabilities, debt included in current liabilities, income tax payable, depreciation and amortization expense, respectively. High accruals may indicate that earnings are inflated by non-cash items, which could lead to future corrections. This factor is useful in detecting earnings manipulation and assessing earnings sustainability.

* : Capitalized management expenses. This characteristic is constructed according to the definition in Eisfeldt (2013). It measures the value of a company's intangible assets, specifically those related to its organizational knowledge, processes, and capabilities that aren't directly reflected on the balance sheet. This factor is typically estimated using financial proxies, like historical SG&A (selling, general, and administrative) expenses. Firms with high organizational capital are often better equipped to generate future cash flows, making this factor particularly relevant for assessing long-term value and growth potential.

Fig3.4 shows a group of fundamental factors which focus on financial performance on a quarterly basis. The figure supports the effectiveness of our constructed quarterly profitability factors in capturing fundamental changes and predicting stock performance. Most notably, the sequential improvement in CHPM (from 1.05 to 1.18) and CHPM\_IA (from 0.75 to 1.0) between Q1 and Q2 2024 successfully presaged the stock's upward momentum in February. The rising trend in these asset utilization efficiency metrics effectively captured the company's improving operational fundamentals, which subsequently manifested in positive market performance. The synchronized positive movement across multiple profitability factors in Q2 2024 provided a robust signal that preceded the stock's price appreciation.



**Fig 3.4** Fundamental Factor Change for 600008.sh

## Macroeconomic factors

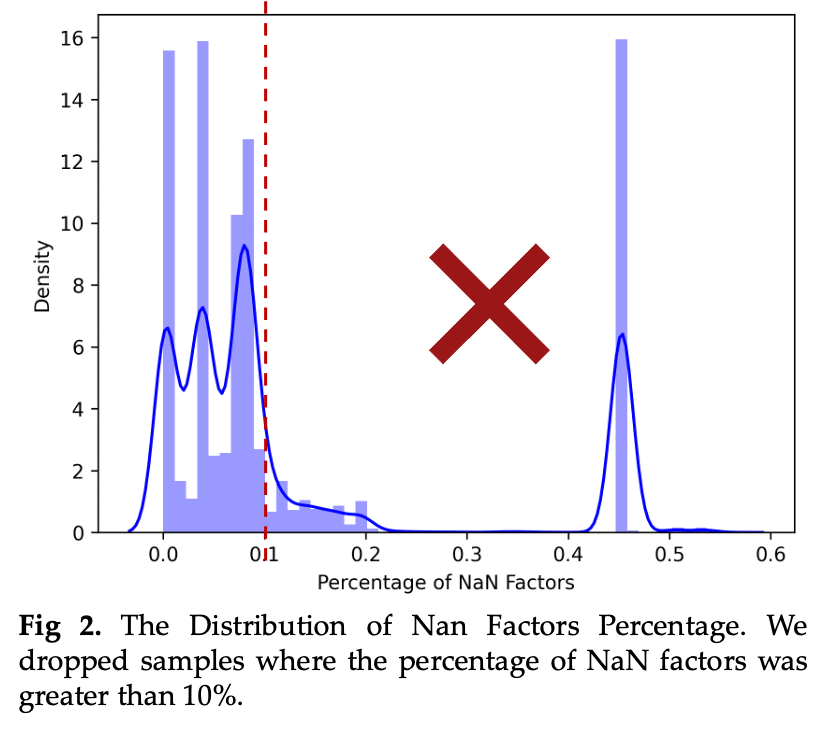
In addition to the above characteristics, we also construct 11 macroeconomic predictors based on the data downloaded from CSMAR and the National Bureau of Statistics websites. These macroeconomic factors provide context on market valuation, liquidity, economic growth, inflation, and monetary policy trends, which are crucial drivers of stock performance over longer horizons.

* Eight of those variables are based on the variable definitions in Welch (2008), including dividend price ratio (), dividend payout ratio (), earnings price ratio (), book-to-market ratio (), net equity expansion (), stock variance (), term spread (), and inflation (). The remaining three include monthly turnover (), M2 growth rate (), and international trade volume growth rate (), which are identified in previous studies as effective macroeconomic predictors. Here we list the details of 3 macroeconomic factors:
* (Earnings Yield): The earnings yield, which is the inverse of the Price-to-Earnings (P/E) ratio, is calculated as total market earnings divided by total market capitalization. In essence, reflects the potential returns available to investors relative to the cost of equity, making it a critical measure for understanding market attractiveness and investment potential.
* (Term Spread): The term spread is the yield difference between 10-year and 1-year government bonds and is a well-known indicator of economic expectations. A positive term spread, where long-term yields are higher than short-term yields, typically suggests expectations of economic growth.
* (Dividend Premium): This factor is the difference between the log of total dividends and the log of the weighted average price in the Chinese A-share market. A higher indicates that dividends are valued relative to stock prices, possibly reflecting investor preference for income-generating stocks in uncertain times.

# Model

## 4.1 Select and Normalize

The above factor data are designed to involve 500 stocks, with a total of 606,475 daily frequency samples. We first screen the samples: on the one hand, some samples with missing labels due to trading suspension and marginal conditions are directly excluded; on the other hand, samples with more missing factors are also directly deleted. In Fig. 4.1, we count the distribution of the proportion of factor nulls in each sample, and we can see that the factor values of some stock sample data are indeed more serious, which is mainly due to the failure to collect complete data information from public platforms when calculating the macro and fundamental factors. Here, a threshold of 10% is used to remove the sample data with the proportion of missing factors larger than the threshold. After screening, we obtained 449765 daily frequency sample data for 360 stocks. Finally, we fill all the remaining null factor values to 0 directly.

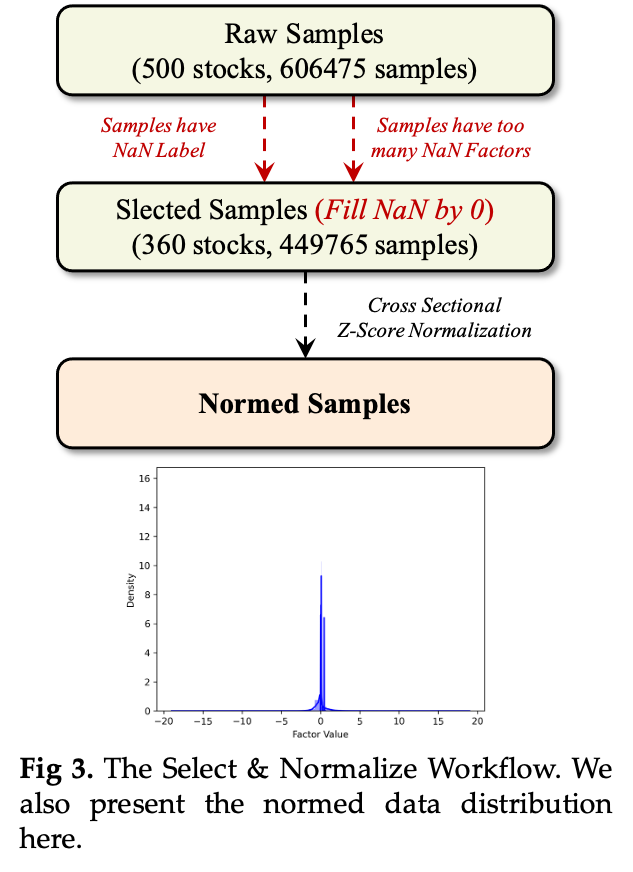


**Fig 4.1** The Distribution of Nan Factors Percentage.

Based on the screened data, we performed the Z-Score Normalization method on the daily frequency stock panel dimension:

where  denotes the factor value of factor i of stock c after normalization at date t, denotes its original factor value, mean (·) denotes the mean operation, and std (·) denotes the variance.

In Fig. 4.2, we summarize the specific workflow of this section and show the distribution of all the factor values after normalization, which can be seen to be in the normal range.



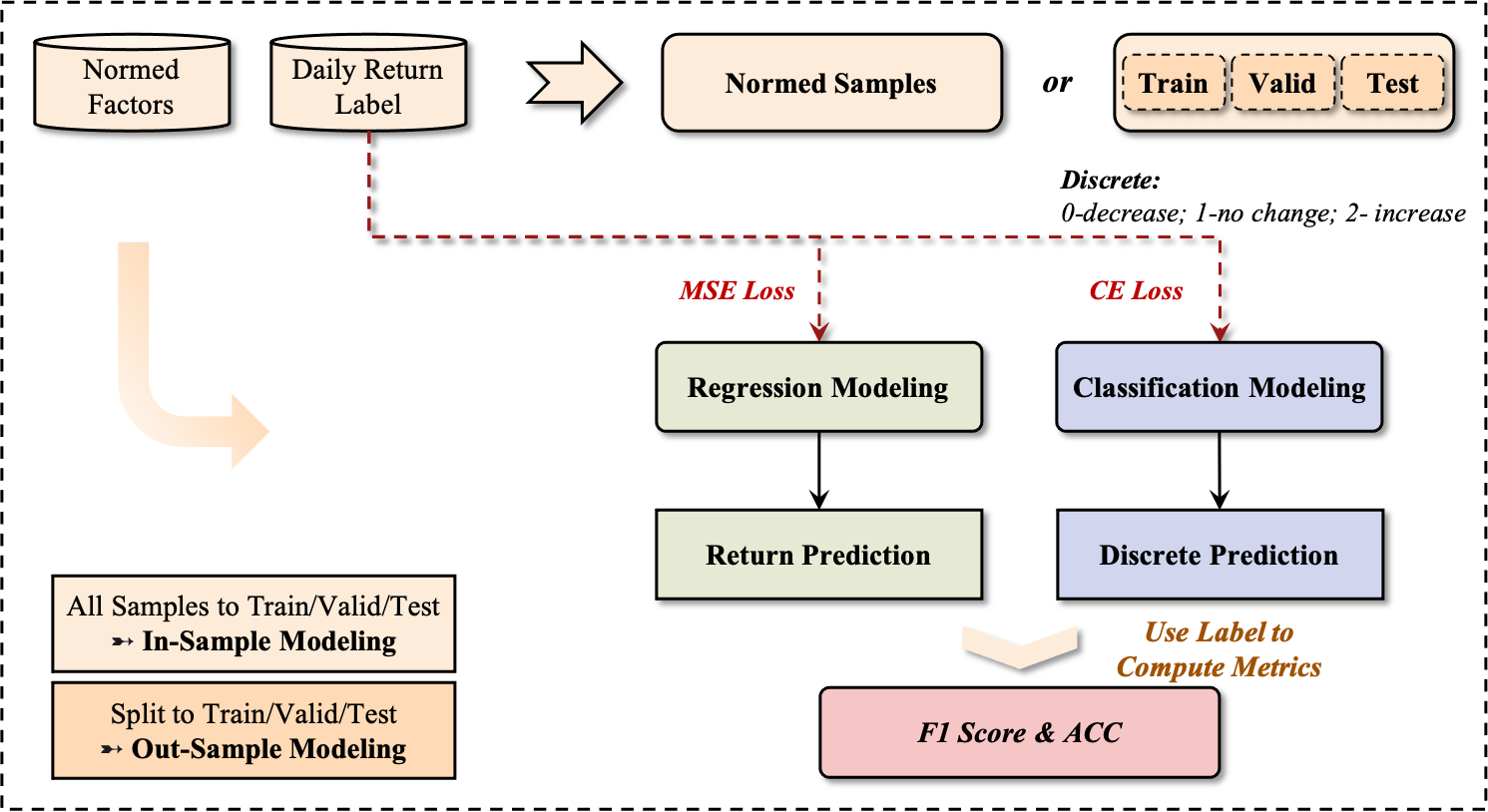
**Fig 4.2** The Select & Normalize Workflow.

## 4.2 In-Sample and Out-Sample Modeling

Using the above filtered and normalized data, the subsequent modeling operations can be carried out. According to the requirements of Project 1, the F1 and ACC classifiers are used to determine the goodness of the modeling.

Naturally, two modeling ideas can be constructed: (1) directly model the regression problem on the yield labels, optimize the model parameters using Mean Square Error as the loss function, and then discretize the continuous values to obtain the F1 and ACC classification indicators; (2) first discretize the yield labels, with 0 indicating down, 1 indicating unchanged, and 2 indicating up, and then perform the classification problem on the discretized labels. problem modeling, and optimize the model parameters with Cross Entropy as the loss function to directly obtain the discrete F1 and ACC classification metrics. In addition, we have also tried: In-Sample modeling, where all the sample data are used as the training set, and Out-Sample modeling, where the data sets are divided into training, validation, and testing according to time.

The framework design of this part is shown in Fig. 4.3.



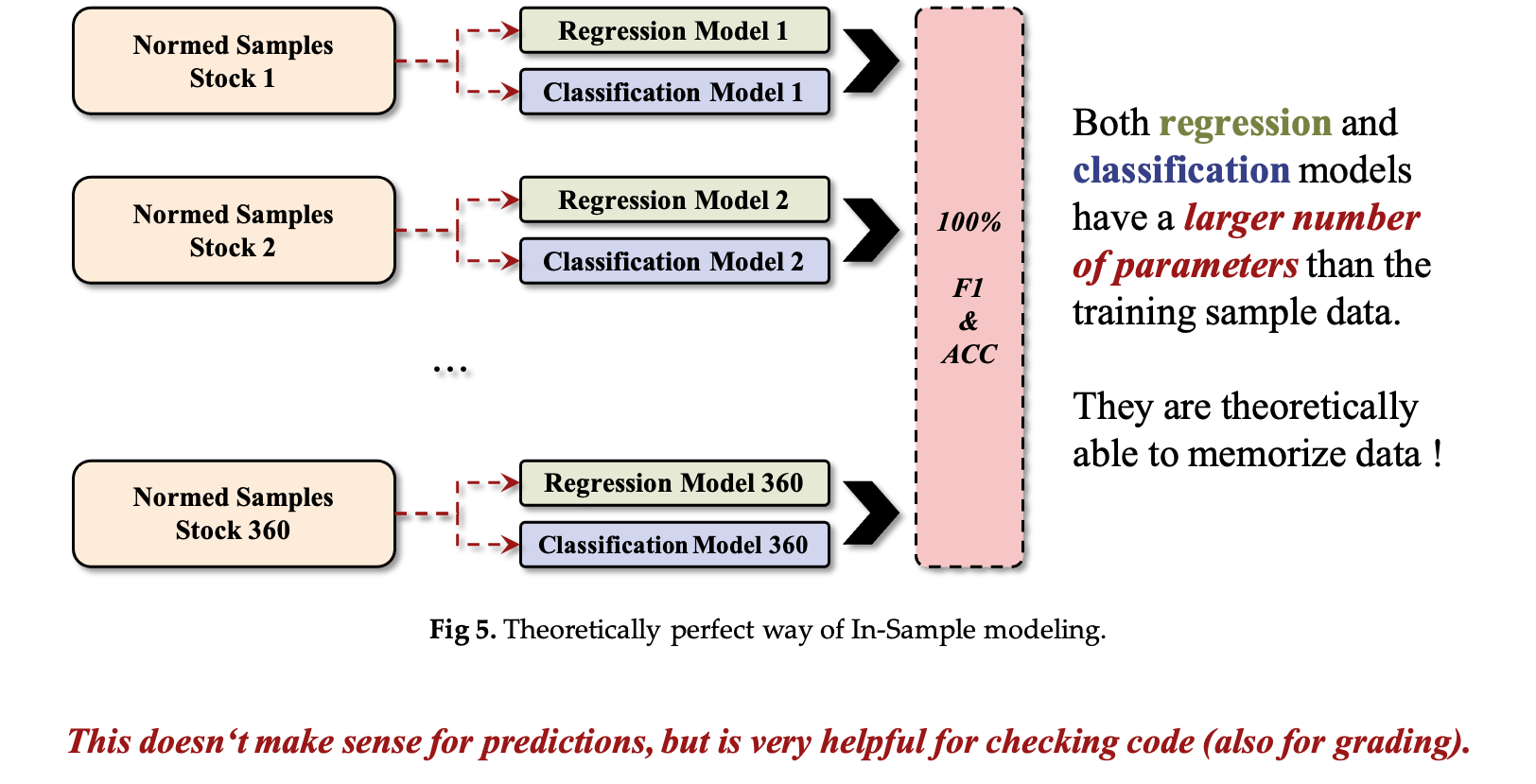
**Fig 4.3** In-Sample and Out-Sample Modeling Framework.

***In-Sample Modeling***

We first perform in-sample modeling, training, validation and testing on the overall 449765 full-sample data. It can be found that each stock has at most more than 1300 days of trading data, and the sample covariance is extremely weak due to the presence of minute-frequency aggregation factors. Therefore, theoretically there exists a way for the in-sample modeling to achieve a perfect fit and obtain predictions with no deviation from the labels, i.e., to achieve 100% ACC and F1 on the full sample.

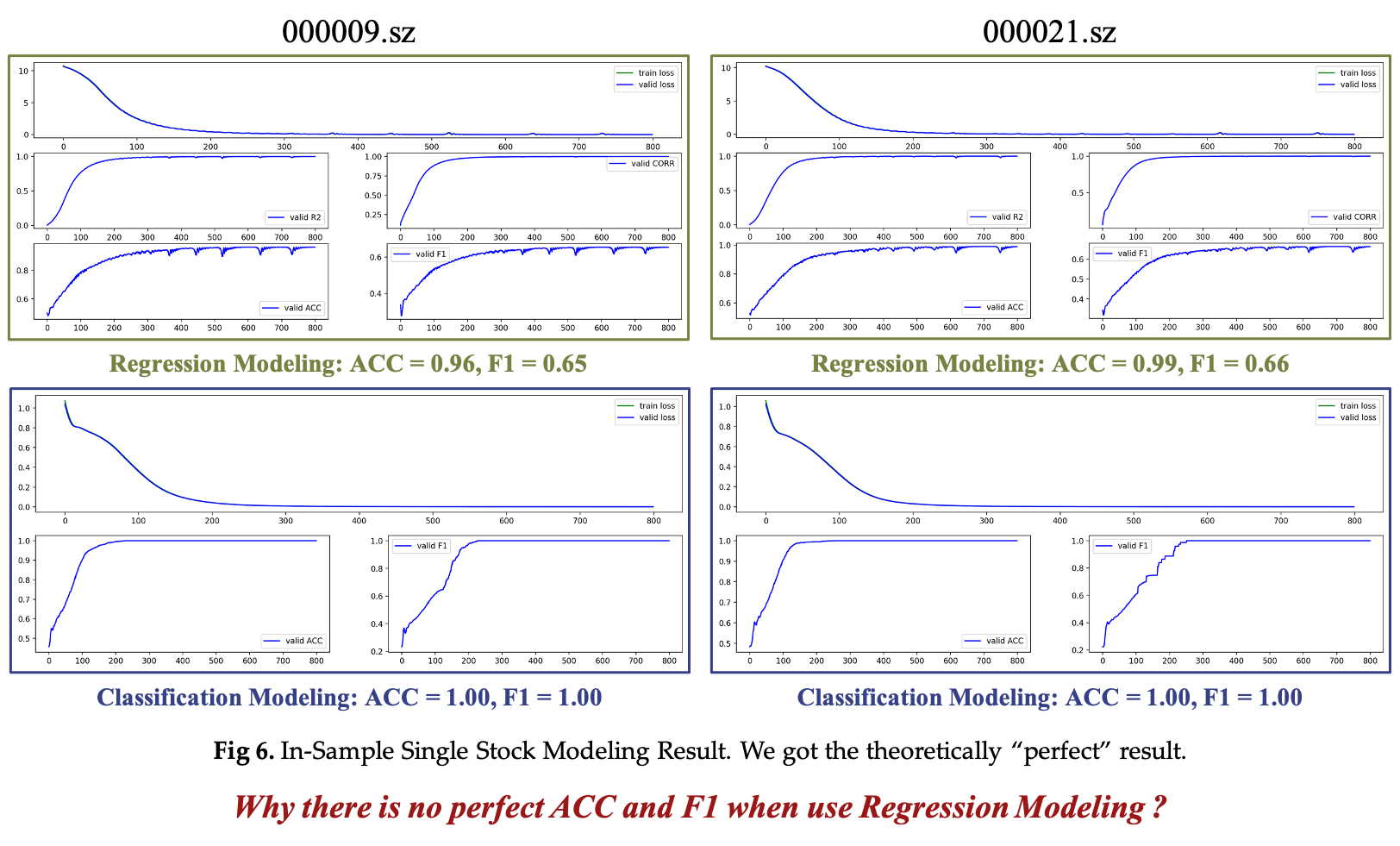
In this way, as shown in Fig. 4.4, we train a model with more parameters than the sample data for each stock, and train it with a smaller learning rate and more epochs to achieve the result that the model remembers the data completely, and then achieves a “perfect” fit within the sample.

Here, we choose a two-layer MLP (with nearly 30,000 parameters) and train 800 Epochs at lr=0.001. The results are shown in Fig. 4.5., which indeed achieve a perfect fit. Interestingly, if regression modeling is performed, the ACC and F1 calculated after discretization do not reach 100%, even though R2 and Corr have reached 1. This is due to the fact that the discretization algorithm here uses symbolic computation, and the output of the neural network is hardly strictly equal to 0 numerically, which leads to a bias in the discretized values. Because of the category imbalance, there is a large gap between F1 and ACC.



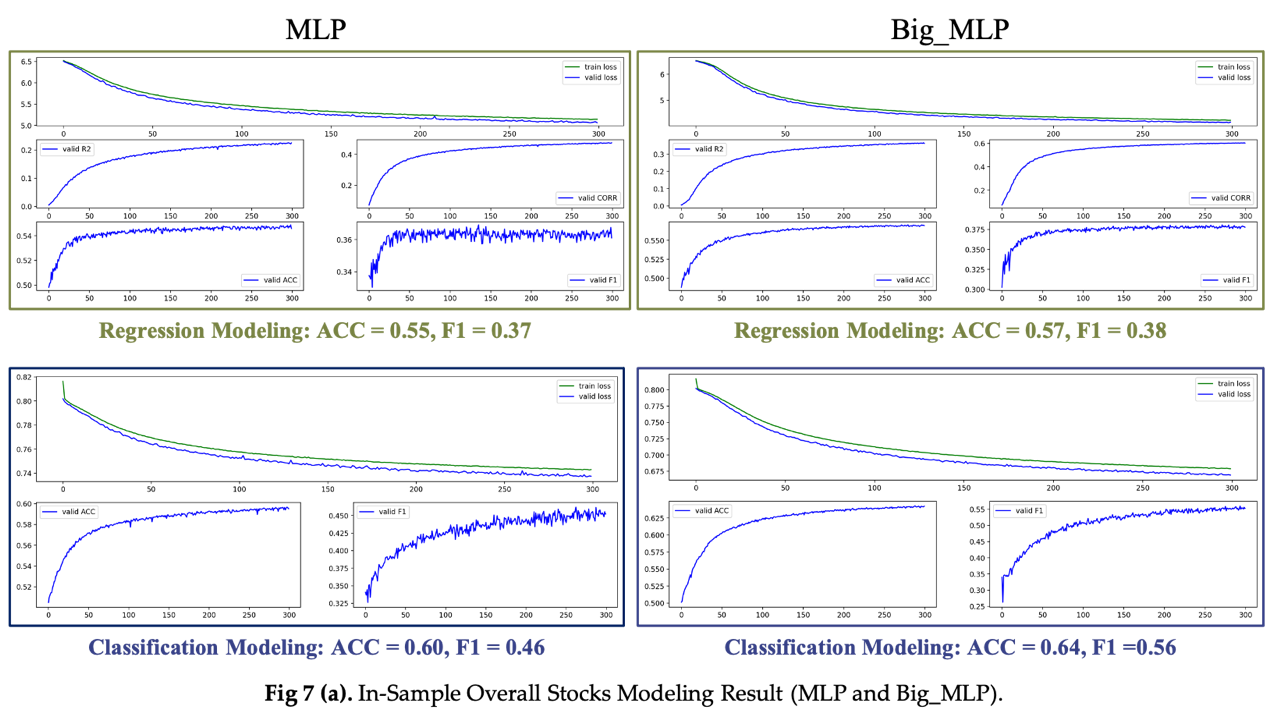
**Fig 4.4** Theoretically perfect way of In-Sample modeling.

In summary, we are here artificially overfitting the model on the data, which does not make sense in a real-world application because the predictive power of a model trained in this way for out-of-sample data is unknowable! However, there are two important implications from a coding perspective: first, a systematic modeling process was set up, and second, the presence of a perfect fit proves that there are no major bugs in the coding, and in fact we validated each model using this process to ensure correctness. Therefore, it does not make much sense to continue tweaking the models and making other attempts from this direction, and it is to be expected: as the complexity of the model decreases, the fit becomes worse.

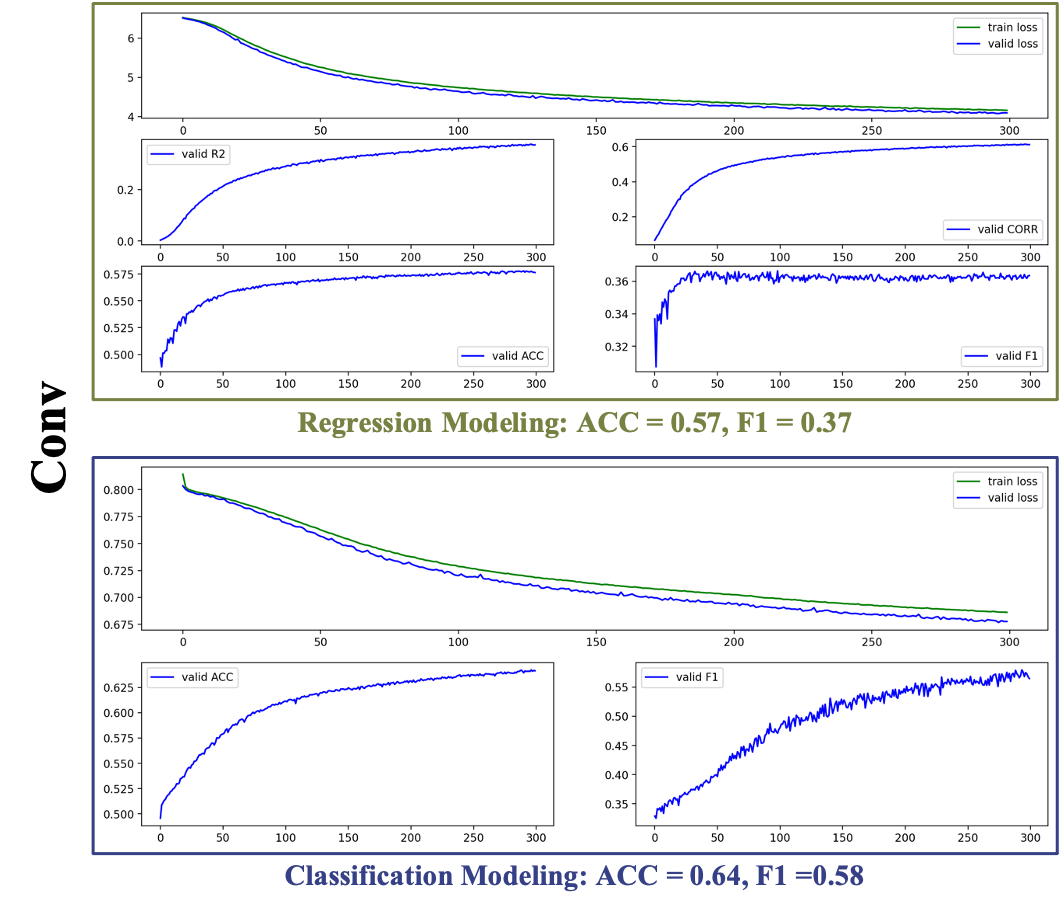


**Fig 4.5** In-Sample Single Stock Modeling Result.

In fact, in order to better explore the factor characteristics of a stock, each stock is generally not modeled individually, but rather the model is constructed using data from the entire stock. Here we directly use the full sample data for modeling, which is also divided into two ideas: classification model and regression model, and the specific experimental results are shown in Fig. 4.6. Here we use three models, MLP\_Small (30,000 parameters), MLP\_Big (50,000 parameters), and Conv (10,000 parameters), to explore the effect of the number of parameters and the model architecture factor feature mining.



**Fig 4.6(a)** In-Sample Overall Stocks Modeling Result (MLP and Big\_MLP).

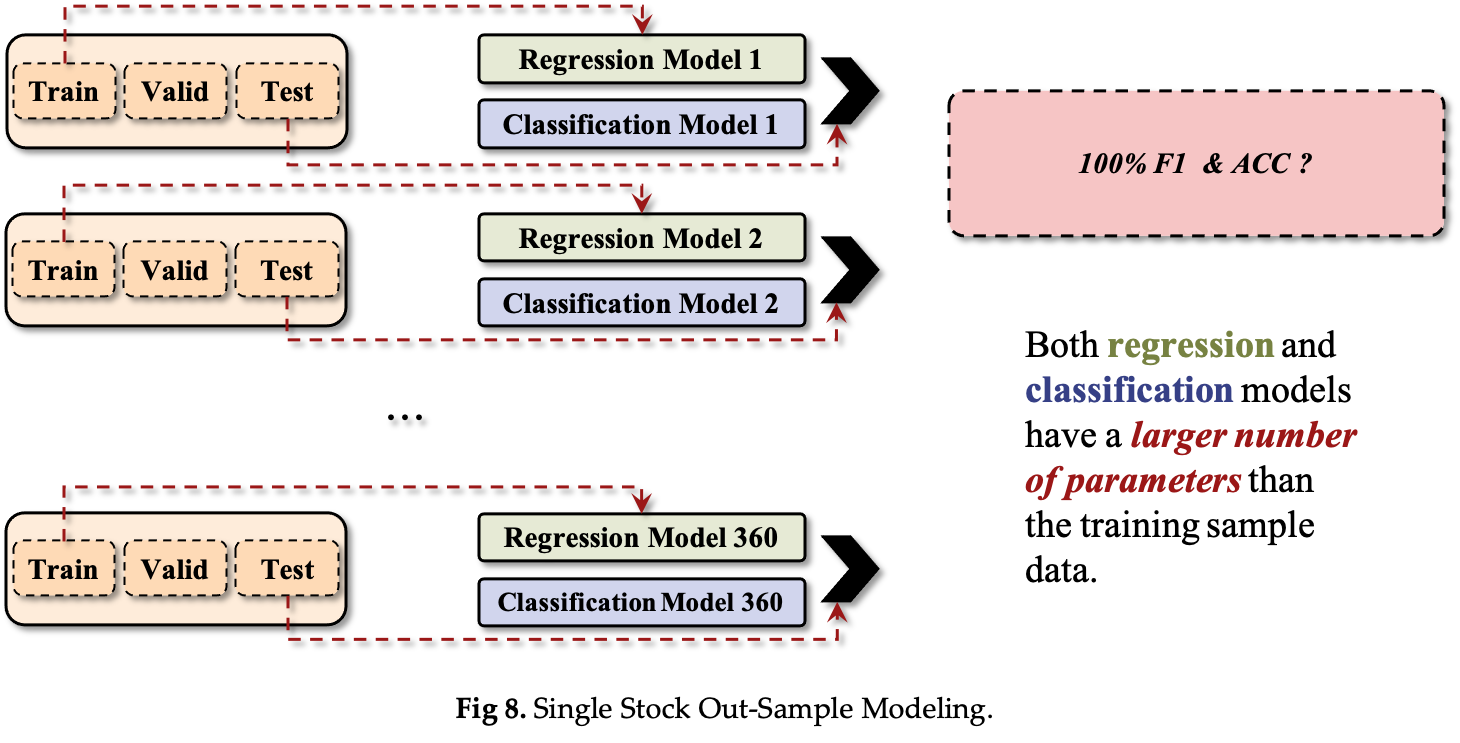


**Fig 4.6(b)** In-Sample Overall Stocks Modeling Result (Conv).

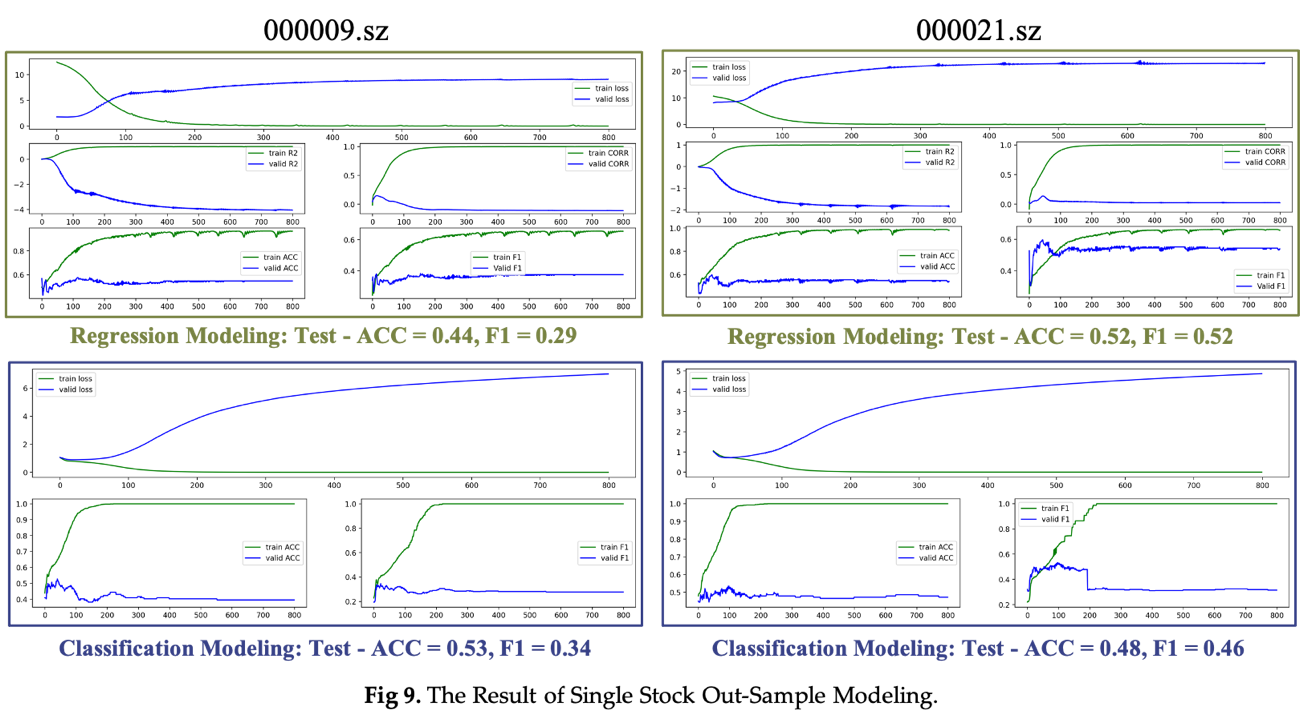
***Out-Sample Modeling***

In order to more realistically simulate the modeling scenarios in practice, we further divided the full sample data into training set (January 1, 2019 to June 1, 2023), validation set (June 1, 2023 to January 1, 2024), and test set (January 1, 2024 to June 1, 2024) according to the time to conduct out-sample modeling. We tried the individual stock modeling approach again to see if we can get better modeling results, the process is shown in Fig 4.7. and the results are presented in Fig 4.8. The overfitting phenomenon can be clearly observed, where the ACC and F1 on the training set are not ideal. As a matter of fact, in real quantitative data modeling scenarios, overfitting is always a gap in front of quantitative researchers. Common methods to overcome overfitting can be categorized as data expansion, model simplification, and constrained training.

Next, we start from these perspectives to explore the optimal modeling approach, and the related experimental results are shown in Table 1. Lag\_1, Lag\_2, and Lag\_3 denote the number of lagged periods of data, Lag\_1 only applies to the current period of data, and Lag\_3 uses the current period and the previous two periods of data for a total of three periods of data.Overall Stock(os) denotes the use of the whole-market stock modeling.Classification(cls) and Regression(reg) denote the use of classification or regression modeling. Classification(cls) and Regression (reg) indicate the use of classification or regression modeling.



**Fig 4.7** Single Stock Out-Sample Modeling).



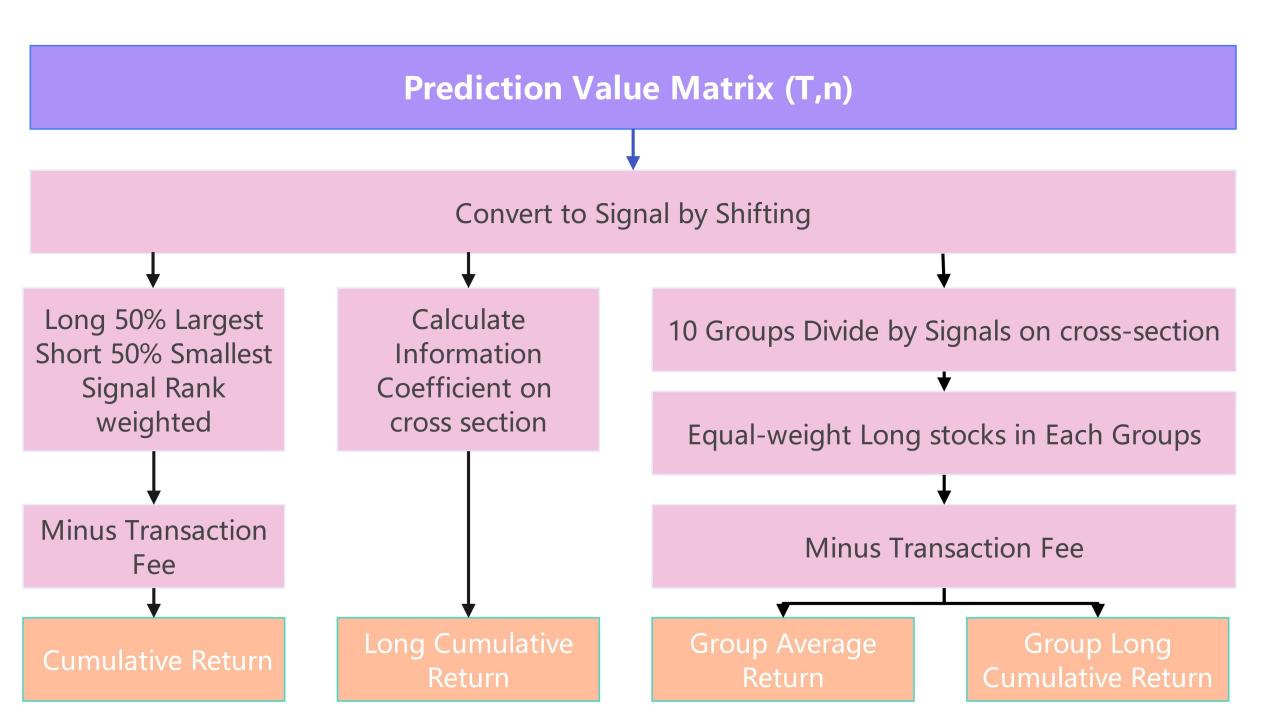
**Fig 4.8** The Result of Single Stock Out-Sample Modeling.

# Backtesting

## 5.1 Backtesting Process

As is known to all, the metrics of model prediction, such as accuracy or F1, can estimate the performance of prediction tasks, but not the actual profitability and stability in transaction. Therefore, we propose backtesting method to evaluate model’s performance in transaction through simple long-short strategy and trading simulation.

The backtesting is conducted with model prediction in test set January 1, 2024 to June 1, 2024).



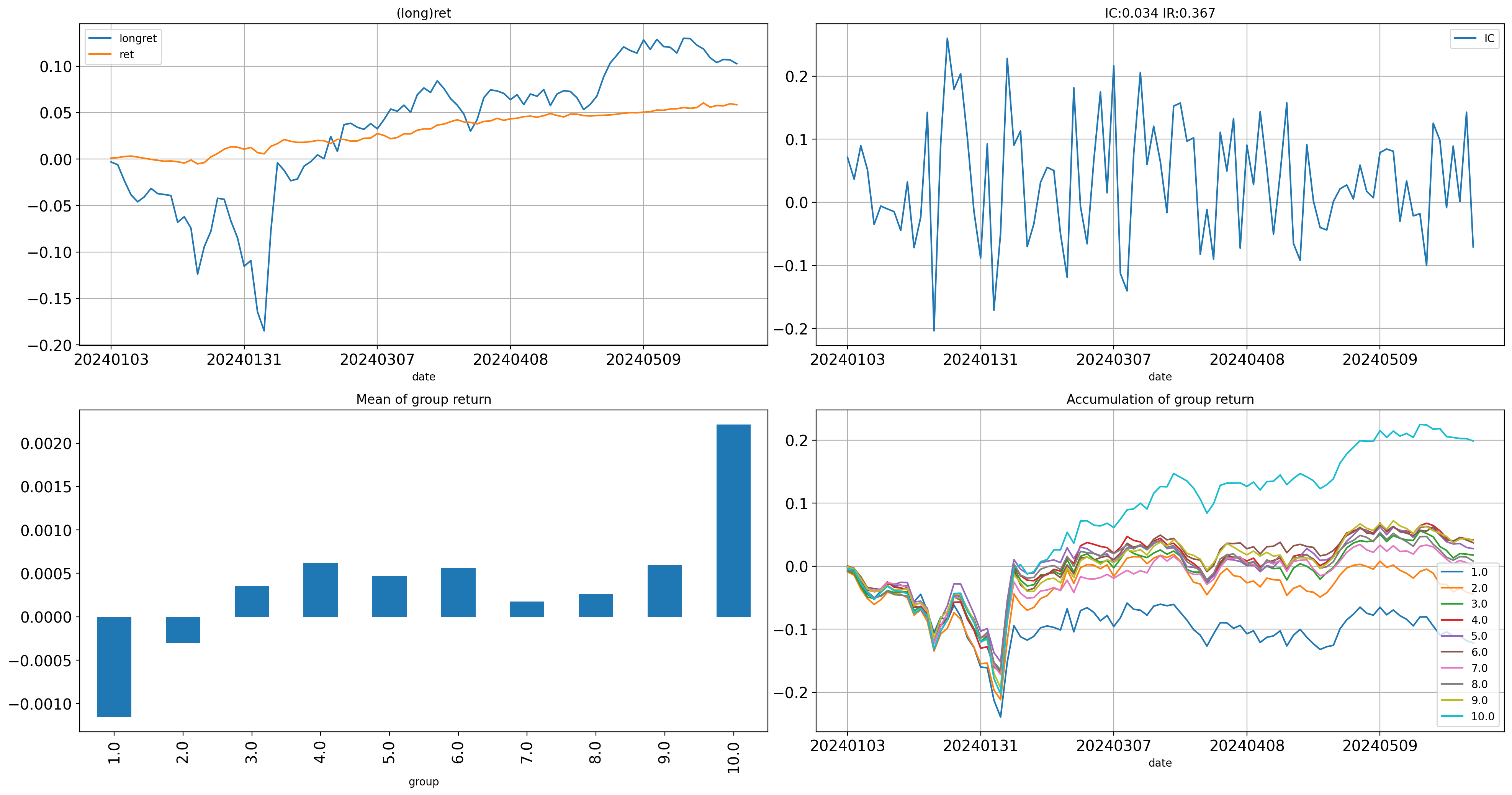
**Fig 5.1** Pipeline for Backtesting.

We define 1 day holding period and 0.01% transaction fee rate. Detail process as below:

1. Since the factor is calculated with the data before t and t, the factor is shifted 1 day to generate the signal of t+1 morning trade.
2. (long) cumulative return: the stock pool is long and short, weighted by percentage ranking, and calculate cumulative return after excluding the turnover cost.
3. Information Coefficient (IC) sequence: The spearman correlation coefficient of signal and excess return cross section is calculated.
4. The average return of ten groups: First, tickets is divided into 10 groups [1-10] on cross section according to their signals, and the averag return of each group is calculated.
5. Ten group cumulative return rate: calculate the accumulative return based on the average return for each group.

## 5.2 Backtesing results

According to the backtesting result shown, the mean of group return shows linear change and group 1’s accumulation return curve is always below group 10. The model can divide stocks with different return to 10 groups and it can distinguish high return stocks from low return ones on cross section.



**Fig 5.2** Indicators performance.

According to the performance indicators resulting from backtesting in Fig 5.2, the conclusion are as follows:

1. The return is 6%, indicating a moderate profitability. But the return of CSI500 is -2% during the same period of time, which means the excess return is 8%.
2. The 10% long return and -4% short return is acceptable in Chinese market because the short is hard to implement.
3. The Sharpe ratio of 4.14 is excellent, as a ratio above 2 is generally considered good in finance, indicating strong risk-adjusted returns.
4. The turnover is 95%, which is high but acceptable when the upper limit of turnover is 200% (whole short to whole long or reverse) .
5. The win rate suggests 60% possibility to profit in trades, which is above average.
6. The IC of 0.03 and IR of 0.37 is good enough to imply the prediction capacity and profitability.
7. A drawdown of only 1% is very low, indicating it is balanced on profitability and risk resistance.

**Table 5.1** Indicators result

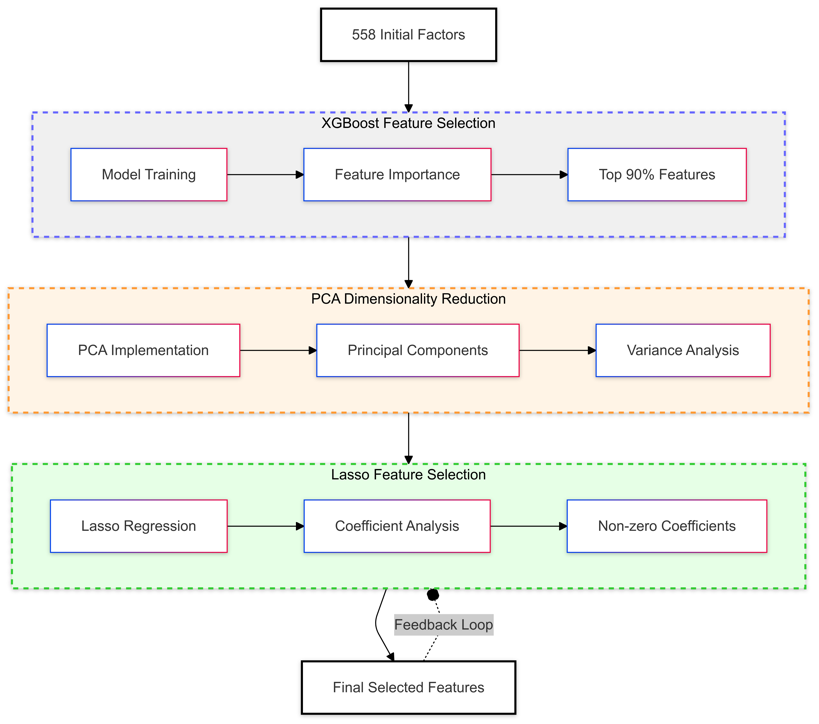
|  |  |  |  |
| --- | --- | --- | --- |
| **return** | **long\_return** | **short\_return** | **sharpe** |
| 0.06 | 0.10 | -0.04 | 4.14 |
| **turnover** | **win\_rate** | **ic** | **ir** |
| 0.60 | 0.03 | 0.37 | 0.01 |

# Factors Importance analysis

Based on the factors, we construct more lagged factors which named as “Factor\_original\_name(t-n)”, where n represents the lag period, to capture delay effects and autocorrelation in time series data. In the end, we possess 558 factors in total. To evaluate these factors and figure out their contribution to the final result, we did factor importance analysis and figured out why they are important.

## 6.1 Which factors matter?

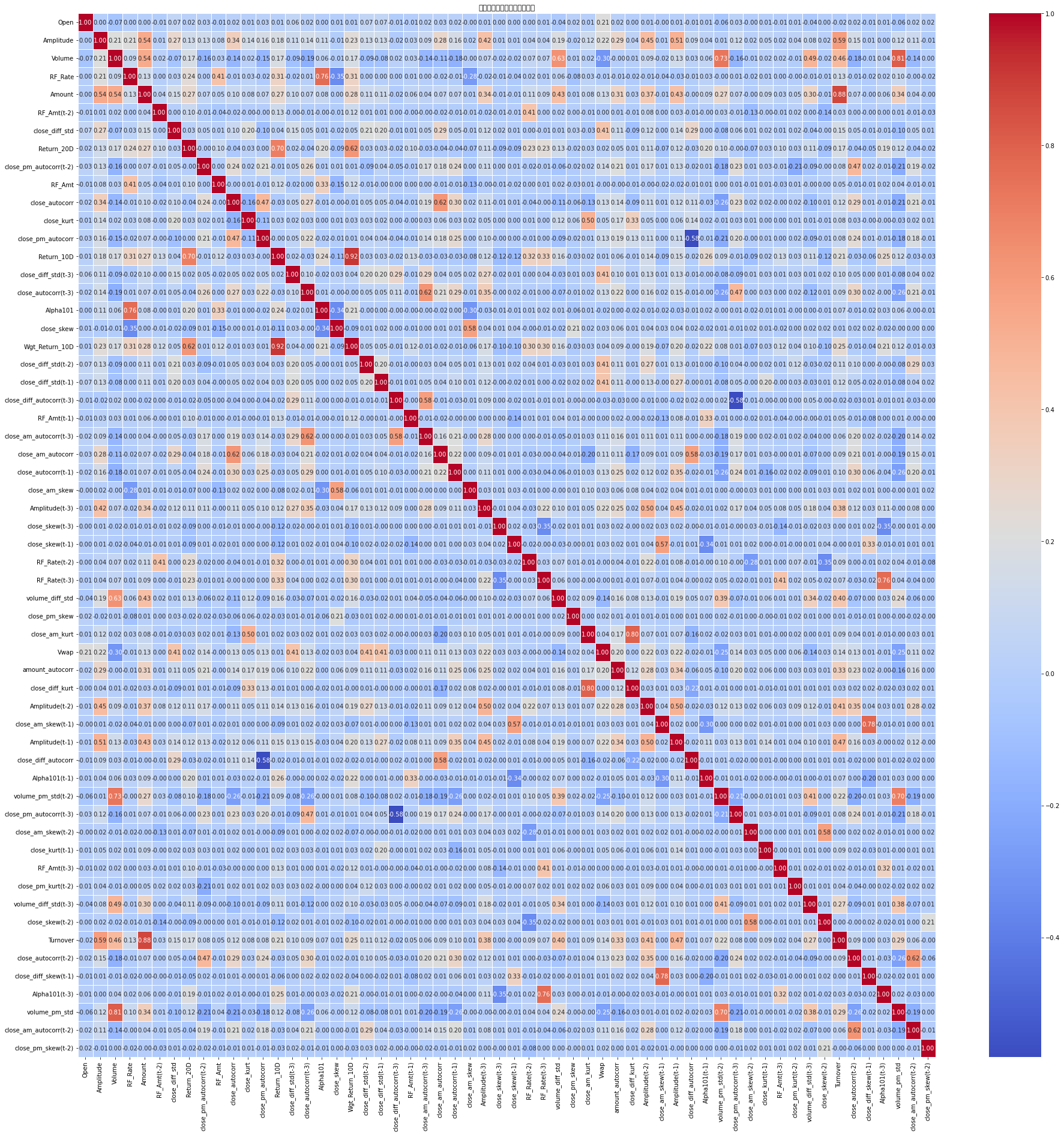
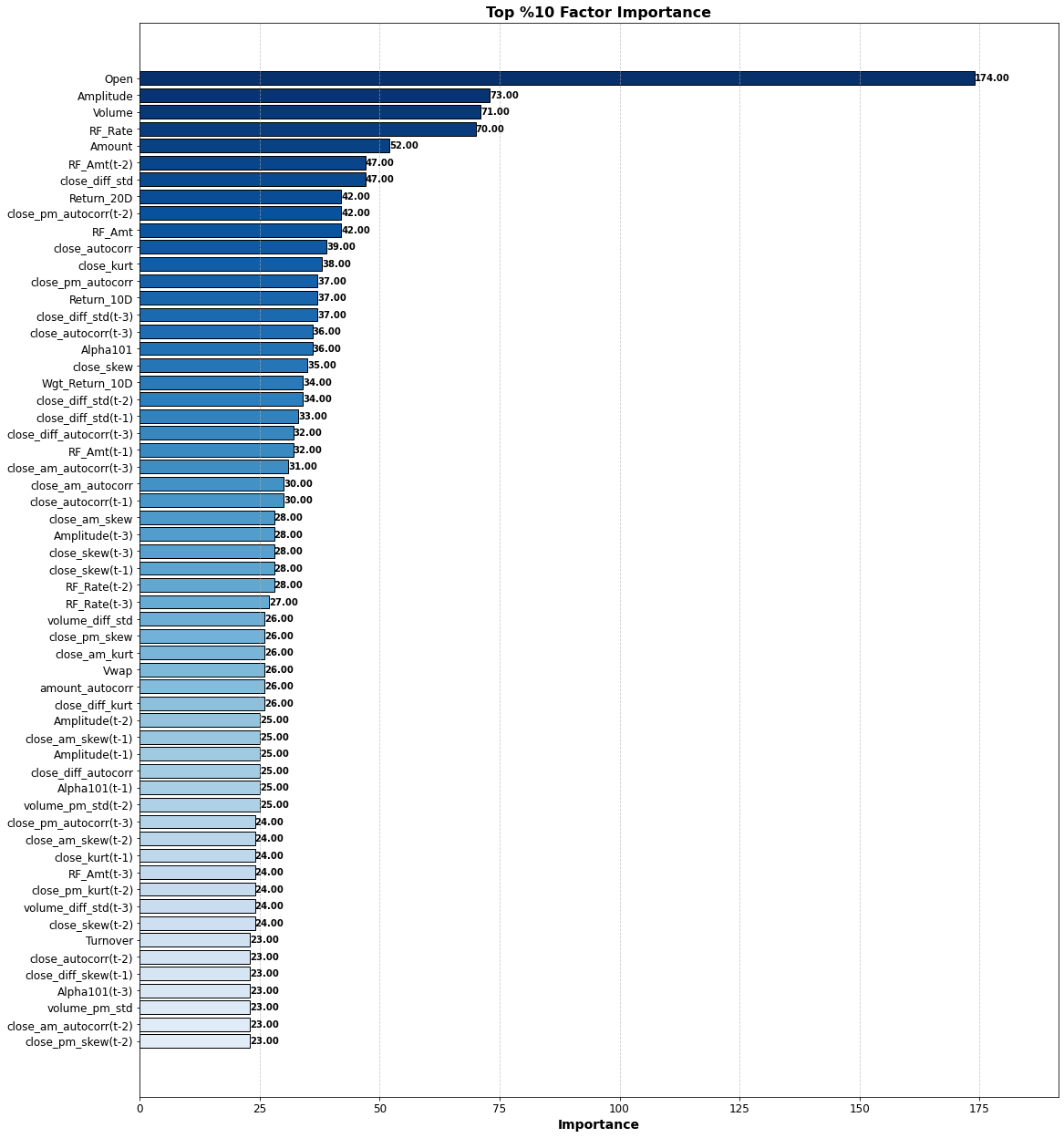
There are many methods to evaluate feature importance, among which, we use tree model, PCA and Lasso regression, the process is presented in Fig 6.1.



**Fig 6.1** Flowchart for Feature Importance

***Feature Importance Assessment Using XGBoost***

We first utilized the XGBoost algorithm. XGBoost inherently provides a measure of feature importance, which quantifies the contribution of each factor to the model’s predictive power. By ranking features based on their importance scores, we were able to focus on those that contribute most to model performance. Fig 6.2 presents the 10% important factors selected by XGBoost(Including the lagged ones) and Fig 6.3 shows their correlation.

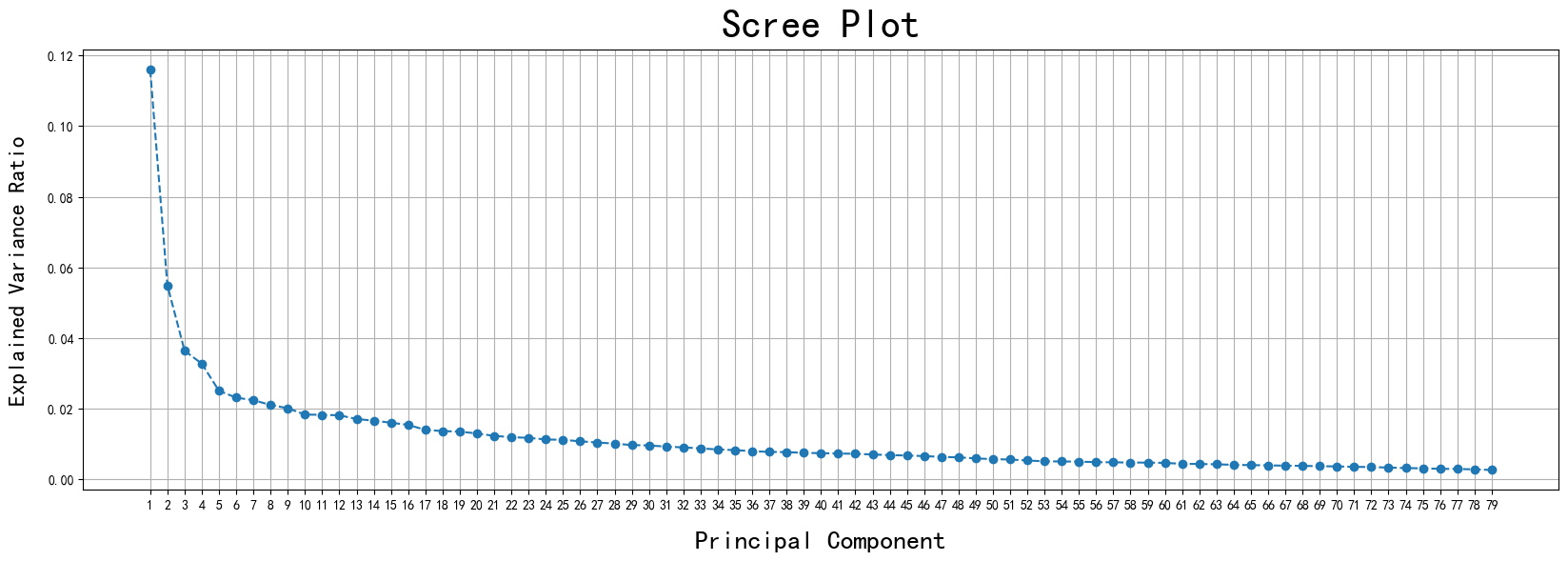


**Fig 6.2** Top10% Feature Importance  **Fig 6.3** Top10% Feature Correlation

***Dimensionality Reduction via PCA***

Following the assessment of feature importance, we selected the last 90% of factors based on their scores. This threshold ensures that we retain the most impactful variables while not directly discarding less relevant ones, which may contain useful information.

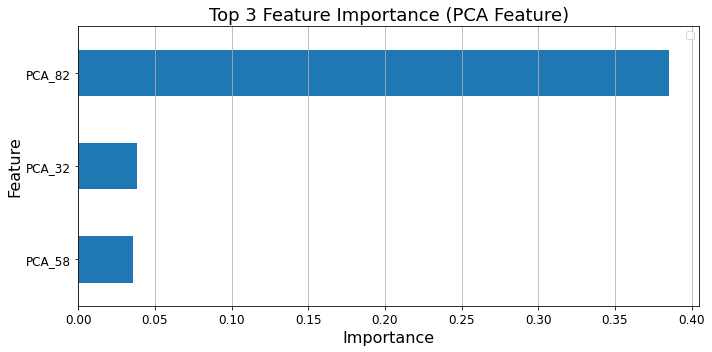
With a refined set of factors in hand, we applied Principal Component Analysis (PCA) to further reduce dimensionality. By applying PCA, we compressed 500 features into 79 components, and the explained variance ratio is 0.95.



**Fig 6.4** Scree Plot for PCA

***Evaluation of New Factors Using Lasso Regression***

The final step in our process involved employing Lasso regression to evaluate the new factors generated by PCA. Lasso, or Least Absolute Shrinkage and Selection Operator, introduces a penalty for the absolute size of the coefficients in the regression model. Through Lasso regression, we were able to identify the most influential principal components that warrant inclusion in our final model, selecting 3 components from PCA, and their importance is shown in Fig6.5.



**Fig 6.5** Feature Importance

## 6.2 Result Analysis

With regards to the ordering of overall variable importance, we find that factors whose importance ranking top 10% all belong to daily trading factors, high frequency factors and their related lagged data. It is reasonable since daily trading factors and high frequency factors are often more sensitive to immediate market conditions. These factors can capture short-term price movements, volatility, and liquidity, which are crucial for predicting price changes in the short run. In financial markets, recent data tends to carry more weight. High-frequency and daily factors reflect the latest market behaviors, making them more relevant for immediate decision-making.

***Fundamental factors***

For fundamental factors, the ranking of Factor\_18 () as the highest importance among fundamental factors, followed by Factor\_89 (), with a sharp decline (>20%) in importance between them.

The factor is a momentum indicator that captures the tendency of assets with strong past performance to continue outperforming. And the monthly frequency of provides a more stable signal over time compared to yearly or quarterly data. Meanwhile, , the largest holder rate reflects the proportion of shares owned by the largest shareholders. This factor can be significant as high ownership concentration may signal confidence in the company, affecting market sentiment and stock performance.

As for the sharp decline, while both factors are important, captures a broader trend of momentum that might overshadow specific ownership insights provided by . If reflects broader market trends that encompass the effects of , it may render the latter less important. In other words, the relationship between ownership concentration and stock performance may be mediated by momentum, making it less independently valuable.

The second influential group contains fundamental signals and valuation ratios, such as Capitalized management expenses (), change in employees (), sales by market capitalization (), number of recent earning increases (), change in profit margin (), and book-to-market (). Here we also observe that lagged fundamental factors exhibit relatively low importance, since most fundamental factors are updated quarterly, lagged versions of these factors are inherently outdated. Markets can experience significant changes within a quarter, influenced by new economic data, geopolitical events, or corporate developments. Lagged factors may fail to reflect these changes, as they are based on past data that may no longer apply.

***Macroeconomic factors***

For Macroeconomic factors, , which measures the level of issuance activity, has the largest variable importance. China has been adopting an approval-based IPO system ever since its stock market opened, and it is well-known that the China Securities Regulatory Commission often suspends or reduces the volume of IPOs when the market is down, making it reasonable for to play an important role in predicting monthly returns.

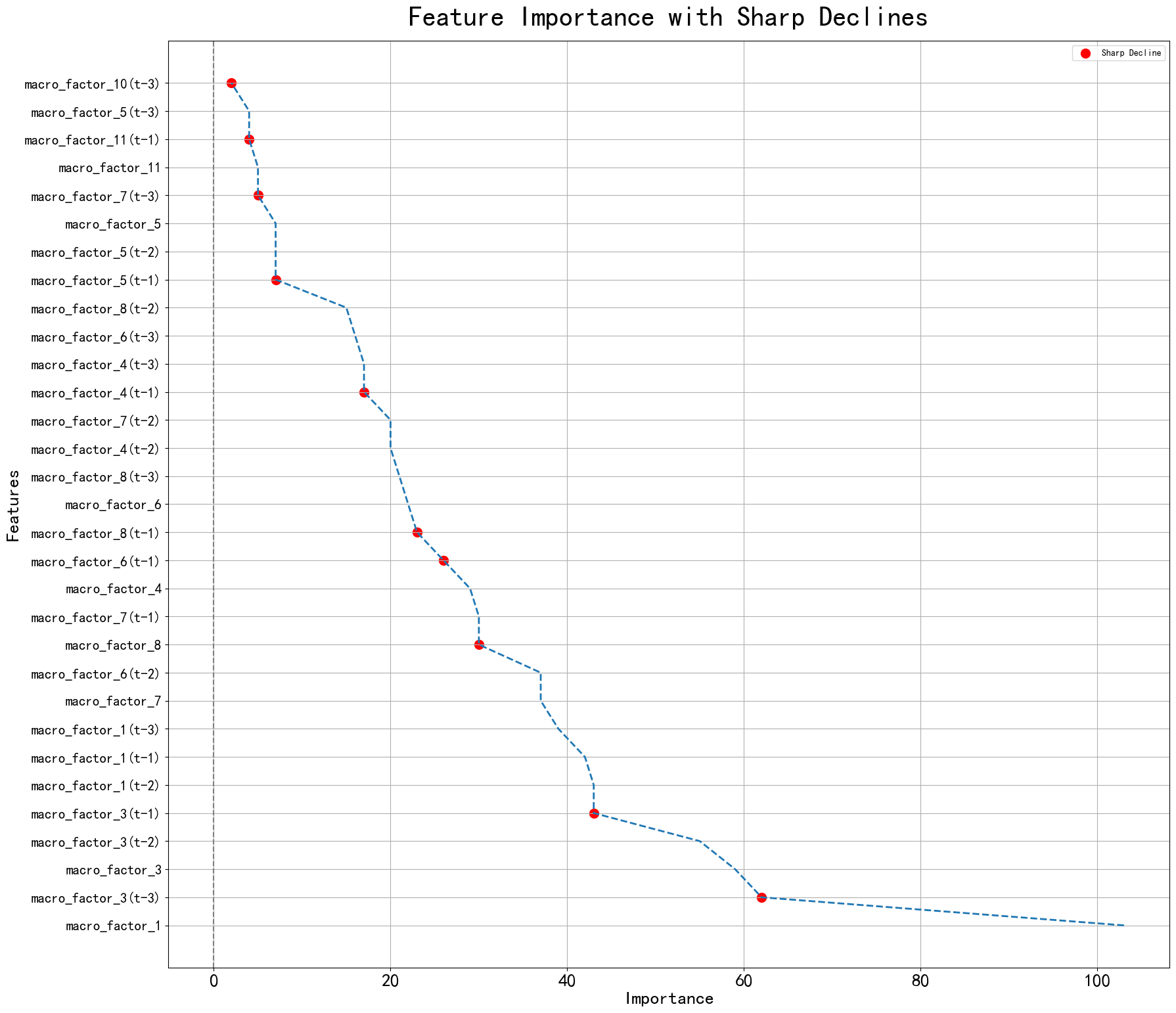
Moreover, XGBoost also puts substantial weight on and their lagged features. A high signifies active trading and elevated investor participation, often associated with a speculative environment or heightened volatility. In the Chinese stock market, characterized by a high proportion of retail investors and a tendency for rapid sentiment-driven movements, is particularly relevant. As for the over 10% drop in importance between the top-ranked factor, , and the second-ranked, , highlights potential distinctions in their predictive relevance.

As for the decline between and , we consider following aspects:

Differential Impact on Market Phases: In the Chinese market, can serve as a proxy for broader economic trends, which are less prone to short-term fluctuations compared to liquidity measures like .

Factor Redundancy and Complementarity: 's distinct economic implications make it less redundant when compared to . The high importance of reflects its unique value, while , although impactful, may partially overlap with other factors in capturing sentiment and volatility.

Structural Market Characteristics: The sharp drop may also reflect inherent structural differences in how capital flow and turnover dynamics influence asset pricing in China. , capturing net capital flow, aligns well with the market’s policy-sensitive nature, where capital availability can shift rapidly with regulatory changes.



**Fig 6.6** Macroeconomic features change

# References

1. Preethi G, Santhi B, et al. Stock Market Forecasting Techniques: A survey. Journal of Theoretical &Applied Information Technology, 2012, 46(1):52-69.
2. Leippold, M., Wang, Q., & Zhou, W. (2020). Machine-Learning in the Chinese Stock Market. Wealth Management eJournal.
3. Gu, S, Kelly, B., Xiu, D. (2020.) Empirical asset pricing via machine learning. Rev. Financ. Stud. 33 (5), 2223–2273.
4. Sloan, R. (1996), Do stock prices fully reflect information in accruals and cash flows about future earnings? (Digest summary), Accounting Review 71(3), 289-315.
5. Eisfeldt, A. L. and Papanikolaou, D.(2013), Organization capital and the cross-section of expected returns, The Journal of Finance 68(4), 1365-1406.

# Appendix

Table 1