Proceedings of the 2023 International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/21/20230238

Quantitative Factor Exploration Based on Insider Trading Detection

Liyu Yang[1,a,*], Tongyang Wang[2,b], Ciyu Cai[3,c]

[1]*School of Information Science and Technology, ShanghaiTech University, Shanghai, 201210, China*
[2]*School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China*
[3]*International School, Beijing University of Posts and Telecommunications, Beijing, 100876, China*
a. yangly2@shanghaitech.edu.cn, b. wtydzd@alumni.sjtu.edu.cn,
c. caiciyu2021@bupt.edu.cn
**corresponding author*

Abstract: Factor research has always been the focus of financial quantitative forecasting research. In the existing multi-factor strategies, we have noticed that the combination modes of factors are different and arbitrary. We hope to develop a more accurate and effective multi-factor model by selecting the most common and most interpretable multi-factors and combining them with equal weight method and assigned weights developed by Hidden Markov Model after some optimizations applied to selected multi-factors. At the same time, we noticed that in the existing data information, there are reports or information that reveal the insider trading of related companies. The existing reports show that the abnormal data volume caused by insider trading will make the prediction of the model inaccurate. Therefore, we added insider trading as a factor into our model training through the order imbalance algorithm to obtain more accurate prediction results. The results show that the multi-factor model is interpretable and effective, and its effect is better than the predicted value than that of the single factor model. After adding the related factors of insider trading into the forecast, it has a certain normalization effect on the original predicted value with large deviation, but has little influence on the effect of the original value with small deviation, which proves the effectiveness of our factor based on insider trading.

Keywords: inside trading, combination forecasting method, hidden Markov model

1. Introduction

In this project, we will gradually explore the multi-factor model construction and try to construct a new factor.

In exploring the multi-factor combination strategy, we first integrated and cleaned the CSI 500 data set, removed the missing values and extreme values, centralized and standardized the corresponding data, and obtained the expected factors through the corresponding calculation method.

Since the combination of multiple factors is composed of single factors, we first test the single factors one by one. We use cross-sectional regression analysis to t-value test to ensure that the factor returns are statistically important. Then we apply a quantile test to find out the monotonicity [1] of

DOI: 10.54254/2754-1169/21/20230238

these single factors. Finally, from the single factor that t<0.05 and is monotonous, we choose six single factors with the best effect, which means their IR is the highest.

After successfully selecting the single factors, their correlation might be high, which will lead to serious consequences if not operated properly. So we apply stepwise regression analysis to test whether there exists multicollinearity, and we will use Schmidt Orthogonalization to reduce the correlation between factors. Now that factors have little correlation, we combined the single factors with equal weight to obtain a multi-factor model [2]. And we tested the effect of the model. We also built a prediction model based on HMM to assign weights instead of applying an equal-weight strategy, and substituted the obtained multi-factors into the stock value prediction to obtain more accurate prediction results. And they both achieve better performance than the benchmark.

However, we learned that the existing multi-factor model could not well explain the expected return of the stock in the future in a short time before the annual or quarterly report was published in accordance with regulations, which requires a new factor to explain.

So we constructed a new factor related to informed trading. Informed trading refers to the act of buying or selling the securities, insider trading, or disclosing information to advise others to buy or sell the securities before the issuance, trading or other information that has a significant impact on the securities trading price is disclosed by the person who knows the internal information of securities trading or the person who illegally obtains the internal information of securities trading. We crawled the published data of legal insider trading in the past two years for analysis. Before the first quarter, half a year, third quarter, and the deadline for the annual report (4.31, 8.31, 10.31), each stock backdates the past 30 trading days, calculates the absolute order imbalance during this period, and weights it according to the number of days from the time when the report is published to obtain a new factor for insider trading. We put the new factors into the model for prediction and got the results. The results show that in some cases where the difference between the original forecast value and the stock is large, the results after adding the insider factor are closer to the real value, while in the case where the original deviation is small, the results after adding the insider trading factor change little [3].

2. Related Works

2.1. Combination Forecasting Method

The article "Asset class correlation affects portfolio volatility and returns" by Phillip Brzenk illustrates that utilizing a combination forecasting approach involves applying multiple forecasting methods to a single problem. This strategy combines not only several quantitative methods but also integrates various qualitative methods. In practice, the mix of qualitative and quantitative methods is more common. The primary goal of this combined approach is to leverage information from different methods to enhance prediction accuracy as much as possible.

For instance, during periods of economic transition, it is challenging to find a single prediction model that accurately reflects the reality of frequent macroeconomic fluctuations and consistently explains the reasons behind these changes. Both theoretical and practical research indicate that when various single prediction models differ, and data sources are distinct, the combined prediction model may result in a superior predictive value compared to any independent prediction value. The combined prediction model can reduce systematic prediction errors and significantly improve prediction effectiveness.

Two basic forms of portfolio forecasting exist:

Equal weight combination, where each prediction method's predicted value is combined into a new predicted value with equal weight.

DOI: 10.54254/2754-1169/21/20230238

Unequal weight combination, where different prediction methods are assigned varying weights. Stock Pricing Method and Model

Linear optimization is one of the mainstream ways to construct factor combinations, that is, to maximize the value function of a different factor linearly weighted by their respective weights. We hope that under the monthly position adjustment and some restrictions, by adjusting the weights of different factors [4], we can maximize the stock value function, that is, maximize the return, which is the best portfolio that the linear model can achieve [5]. Factors are represented by multidimensional vectors, and each stock is affected by many factors. It has the advantages of simplicity and intuition, low complexity of optimization calculation and little time consumption. But correspondingly, its disadvantage is that it may discard the correlation information between different stocks, and at the same time [6], it will make the final optimization result more concentrated.

The optimization method of quadratic programming under the "factor return" linear assumption can integrate the correlation information between individual stocks. We calculate the expected return rate of each stock through hidden Markov chain, and then use Markov mean variance theory to optimize a value function to achieve such a goal: to ensure the minimum volatility of the stock portfolio while ensuring the fixed total return rate of the stock. However, its disadvantage is that the factor is not necessarily proportional to the expected rate of return, and its weight matching result is highly sensitive to parameters, so it requires high accuracy of training data.

The quadratic programming combination of the covariance matrix of style factor data estimation is an improvement of the above method [7]. The covariance of the above method is the covariance of the expected return rate of each stock. Here, it is updated as the covariance of the factors of each stock multiplied by the number of factors. Because the optimization function contains both multi-factor information and expected return information, it may produce better results.

2.2. Detection of Abnormal Fluctuations and Insider Trading

The most effective way to detect insider trading is absolute order imbalance [8]. Order imbalance is defined by the difference between the bid and ask volume, when its absolute value reflects the intensity of market emotion. However, in the Chinese security market, the bid and ask volume is not directly accessible. Lee and Ready proposed an algorithm that classifies trades as either buyer-initiated or seller initiated, which is defined by: Volume Order Imbalance(VOI) [9]:

$$OI_t = \delta V_t^B - \delta V_t^A, \tag{1}$$

where

$$\delta V_t^B = \begin{cases} 0 & P_t^B < P_{t-1}^B \\ V_t^B - V_{t-1}^B & P_t^B = P_{t-1}^B \\ V_t^B & P_t^B > P_{t-1}^B \end{cases} \tag{2}$$

$$\delta V_t^A = \begin{cases} V_t^A & P_t^A < P_{t-1}^A \\ V_t^A - V_{t-1}^A & P_t^A = P_{t-1}^A \\ 0 & P_t^A > P_{t-1}^A \end{cases} \tag{3}$$

At time t, V_t^B and δV_t^A represent the bid and ask volumes, while P_t^B and P_t^A denote the best bid and ask prices, respectively. If the current bid price is lower than the previous bid price, it could indicate that a trader either canceled their buy limit order or an order was filled at P_t^B−1. However, since we lack a more detailed order or message book, we cannot definitively determine the trader's

DOI: 10.54254/2754-1169/21/20230238

intent. Therefore, we conservatively set $\delta V_t^B = 0$ [10]. If the current bid price remains the same as the previous price, the difference in bid volume represents incremental buying pressure from the last five periods. On the other hand, if the current bid price is higher than the previous price, it can be interpreted as upward price momentum due to a trader's willingness to buy at a higher price. Downward price momentum and selling pressure can be similarly inferred from the current and previous ask prices.

For the majority of days, the order imbalance autocorrelation is significant up to lag 15. Its first difference exhibits a significant negative autocorrelation at lag-1, consistent with Chordia's results [11]. This finding suggests that positive imbalances often lead to periods of persistent positive imbalances, as traders tend to split their orders across multiple periods, as explained by Chordia. The absolute order imbalance has been employed as a proxy for the probability of informed trading measure by Easley, Kiefer, O'Hara, and Paperman. Multivariate Regressions Analysis on Informed Trading

Ahern (2018) follows this direction and shows absolute order imbalance is the only positive and significant proxy related to informed trading among several empirical proxies from previous literature when the information is short-lived through empirical analysis. Specifically, the indicator has the most significant effect when combining order imbalance and urgency (defined by the reciprocal of the days before the event).

The research presents OLS regression coefficients where the dependent variable in columns 1–4 is a dummy equal to one if any insider trading occurred on a given day, and in columns 5–8 is the logged volume of shares traded by insiders on a given day. Observations are from a panel of 410 events with trading days t = -120 to t = -2, relative to the announcement date of t = 0. p-values from standard errors clustered at the event level are presented in parentheses. Significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *. Dependent variable: Informed trading dummy Dependent variable: Ln(1+informed trading volume)

3. Method

3.1. Single Factor Detection

The single-factor test is an important step before building a multi-factor model. By testing the effectiveness and importance of a single factor, high-quality single factors can be screened. The implementation steps are as follows:

3.2. Define Backtesting Time Range T and Stock Set S

Exclude companies that have been listed for less than half a year and ST shares.

3.3. Factor Construction

For the existing factor, we extract the factor value of stocks in all stock sets S within the time range T from the DataYes Platform. For the newly defined factor, the new factor is constructed by definition and algorithm from the basic data on the platforms such as Flush and tushare.

3.4. Data Processing

We delete the missing values in the data, and delete the extremely abnormal corresponding data. After calculating the corresponding maximum and minimum values in the dataset, we standardize the data to (0, 1) to facilitate the next step of factor processing.

DOI: 10.54254/2754-1169/21/20230238

3.5. Regression Method

We use cross-sectional regression to test the validity of factors, because it can easily deal with the unknown time series of factor returns. In all cross-section periods, we conducted regression tests on the specific factor, and we can obtain the factor return rate series of the factor (that is, the series formed by the regression coefficient X_{T_d} in all cross section periods) and the corresponding t value series. The t-value refers to the t-test statistic of a single regression coefficient X_{T_d}, describing the significance of a single variable. The regression weight we use is the least squares regression OLS, and the factor we choose has the property that its factor return is statistically significant according to the t-value test.

3.6. Monotony Detection

In order to test the monotonicity of factors, the layered back testing method is used here. This is the quintile. The monotony of a factor means that the effect of the equal weight strategy of the stocks with the highest factor value ranking interval will be better than that of the stocks with the highest factor value ranking interval at any time. From the perspective of images, the factor layered back test map with good monotony will produce five parallel and disjoint curves as far as possible. We will choose the factor with good monotony.

3.7. Multi-Factor Combination

Introduction: After completing the single-factor test, the first six factors with the best IR among the factors with statistical significance (t<0.05) and monotonicity (manually screened according to the hierarchical back testing effect) are selected as the primary multiple factors in the factor library.

The back test time range T and stock set S remain unchanged. Transaction frequency is 1d.

3.8. Integration of Multiple Factors

We hope that the smaller the correlation between the selected factors, the better. Too large a correlation between several factors may cause a high degree of multicollinearity, leading to a large error in the model's ability to explain the stock price and the market. The simplest way is to delete a factor with a small IR value from the factors with large correlation, but this will reduce the explanatory power of the multi-factor model. Therefore, we need a better method to solve the problem of multicollinearity, and reduce the correlation between factors while retaining the explanatory power of factors [12].

I. Test: VIF (Variance Expansion Test) method

VIF is used to check whether there is multicollinearity between factors. If there is, step 2.2.2 is required. Otherwise, step 2.2.2 is not required. For each factor, we use other N factors for OLS regression interpretation. The size of R^2 determines the explanatory power of explanatory variables to dependent variables. The higher the VIF, the stronger the linear correlation between explanatory variables and dependent variables. If VIF is more than 5, then there is a multicollinearity problem [13].

II. Integration approach

The problem of multicollinearity is solved by stepwise regression, that is

a. Start from the mean model.

b. Add one variable at a time from the potential list of variables. Based on the largest partial F-test statistics, say F_0^*, add variable if $F_0^* > F_{in}$

c. Check if any of the existing variable should be eliminate. Using partial F-test, i.e., eliminate the variable if $F_0^* < F_{out}$ ($F_{out} < F_{in}$ otherwise looping)

DOI: 10.54254/2754-1169/21/20230238

d. Repeat b and c until no more variable should be added or eliminate.

After we find the corresponding available factor, we choose Schmidt orthogonalization to orthogonal the factor. Factor orthogonalization, we think, is essentially to rotate the original factor (through a series of linear transformations) to obtain a group of new factors that are orthogonal in two after rotation. The correlation between them is zero and the explanatory degree of earnings (i.e., the overall variance) remains unchanged. The zero correlation ensures that there is no collinearity between the rotated factors [8], while the degree of interpretation remains unchanged to ensure that the information contained in the original factors can be preserved.

Schmidt orthogonality is a sequential orthogonality method, so it is necessary to determine the order of factor orthogonality. The main idea of Schmidt orthogonality is to give a set of vectors, and orthogonal and normalize each vector with all previous vectors in the given order [14].

Another idea of orthogonality is to reduce the modification of the original factor matrix as much as possible to obtain a set of orthogonal bases. In this way, the similarity between the post-orthogonal factors and the causes can be maintained to the greatest extent. In addition, we hope to treat each factor equally and avoid factors that are biased to the top of the orthogonal order in Schmidt's orthogonal method [15]. Symmetric orthogonality treats each factor equally. They rotate at the same angle to obtain a set of orthogonal bases. The corresponding relationship between factors before and after orthogonality is best maintained. The correlation between factors before and after orthogonality is also higher than Schmidt orthogonality, so the degree of interpretation is stronger.

Finally, all the factors obtained will be re-standardized.

3.9. Factor Combination Mode

Given the combination of factors, we will get the comprehensive factor.

I. Equal weight combination.

After Schmidt orthogonalization, we get a new multi-factor, and then we add or subtract a new comprehensive factor according to the equal weight of the positive and negative correlation between the factor and the yield. Equal weight method is not a perfect combination method, but it is simple and interpretable [13]. We carry out equal weight combinations for the most effective factors in each category, and then select stocks and build positions to obtain our equal weight factor combination strategy.

The specific strategies are as follows:

a. The time span is set from January 1, 2013 to December 01, 2022, and the stock pool is CSI 500 stocks.

b. Our position adjustment date is the first trading day of each month.

c. Rank 300 components according to indicators, with each 50 stocks as one file, and build positions for each file.

d. The buying and selling price of stocks during position adjustment is the closing price of the day.

e. Equal weight reconstruction and combination.

f. The portfolio will not be adjusted during the portfolio holding period.

g. During the combination construction, the transaction cost is 0.1%, the transaction commission is 0.1%, and the stamp duty is 0.1%.

II. HMM.

Hidden Markov Model (HMM) is a machine learning model, which is widely used in language recognition, natural language processing, returns prediction and other fields. We build a hidden Markov model to predict the potential returns of the stock generated by the factors we build and assign weights according to the prediction by different factors. The input of the model is the factor we have processed before, and the output is the predicted stock value [2].

DOI: 10.54254/2754-1169/21/20230238

3.10. Multi-factor Backtest

The regression test of the multifactor model can be simply summarized as the following three steps:

a. Calculate the exposure of each asset on all factors β_i

b. Estimation of multi-factor model through regression analysis.

c. Joint inspection of asset pricing error α_i and the expected return of each factor k.

The core issue of multi-factor model research is why the expected return on assets is different across different assets. We use cross-section regression to test the multi-factor model, because it can easily deal with the unknown time series of factor returns. In all cross-section periods, we conducted regression tests on factor d, and we can obtain the factor return rate series of the factor (that is, the series formed by the regression coefficient X_{T_d} in all cross-section periods) and the corresponding t value series. The t-value refers to the t-test statistics of the single regression coefficient X_{T_d}, which describes the significance of a single variable. The absolute value of the t-value is greater than the critical value, indicating that the variable is significant, that is, the explanatory variable (the exposure of individual stocks in factor d of the T period) is a factor that affects the dependent variable (the income rate of individual stocks in the T+1 period) [10].

I. We first determine the factor exposure of the selected factor d through the following time series linear regression model:

$$R_{it}^{e} = \alpha_i + \beta_i^{`} f_t + \epsilon_{it}$$

II. Use the estimate of factor exposure obtained in the first step as the explanatory variable, and take the time series average of asset return R^e in all T periods as the explanatory variable β The linear regression model satisfied on the section is:

$$E_T[R_i^e] = \beta_i \lambda + \alpha_i$$

Get the estimate of the expected rate of return of the factor and the estimate of the pricing error of each asset α_i.

3.11. Insider Trading Cases

3.11.1. Regulation Rules in Chinese Security Market.

Insider trading refers to the act that a person who knows the internal information of securities trading or who illegally obtains the internal information of securities trading buys or sells the securities for insider trading, or divulges the information, and advises others to buy or sell the securities before the issuance, trading or other information that has a significant impact on the securities trading price is disclosed.

Common insider trading behaviors include:

a. Insiders use insider information to buy or sell securities or suggest others to buy or sell securities based on insider information.

b. Insiders of insider information disclose insider information to others, so that others can use the information for insider trading.

c. Those who illegally obtain insider information use insider information to buy or sell securities or suggest others to buy or sell securities.

Contrary to everyone's idea, not all insider trading in China's securities market is illegal. The directors, supervisors and senior executives of listed companies may legally trade part of the company's shares under strict restrictions [11].

DOI: 10.54254/2754-1169/21/20230238

In accordance with Article 44 of the Securities Law of the People's Republic of China, Paragraph 2 of Article 141 of the Company Law of the People's Republic of China, and Article 189 of the Securities Law of the People's Republic of China, the relevant provisions of the Shanghai Stock Exchange and Shenzhen Stock Exchange trading management measures.

Directors, supervisors, senior executives and shareholders holding more than 5% of the shares of a listed company shall not trade the shares of the company during the following periods.

(1) From 30 days before the announcement of the Company's annual report to the final announcement date.

(2) 10 days before the announcement of the Company's performance forecast and performance express to the final announcement date.

(3) The period from the decision of a major transaction or major event to two trading days after the announcement of the event.

(4) Other major events that may have a significant impact on the trading prices of the Company's shares and derivatives occur within two trading days after the announcement.

(5) Where a director, supervisor, senior manager of a listed company, or a shareholder holding more than 5% of the shares of a listed company sells the shares of the company he holds within six months of buying them, or buys them within six months of selling them, the proceeds therefrom shall belong to the company, and the board of directors of the company shall recover the proceeds therefrom [9].

However, if a securities company holds more than 5% of the shares due to the purchase of the remaining after-sales stocks through exclusive sale, this shall not exclude.

(6) During the term of office of a director, supervisor or senior manager of a listed company, the number of shares transferred each year shall not exceed 25% of the total number of shares held by him.

(7) The directors, supervisors and senior executives of the listed company shall report to the listed company within two trading days from the date of the occurrence of the fact. The listed company shall fill in the report online through the website of the stock exchange within two working days after receiving the report. The exchange will publicly display the information filled in by the listed company on the website the next day.

As the all trading of directors, supervisors and senior executives should be published in the website, we crawled the above-mentioned published data on legal insider trading in the past two years for further analysis.

3.11.2. Example of Insider Trading.

In this subsection, we provide a latest illegal insider trading case and show how it is correlated with our legal insider trading data. According to the news "Insider trading again! ST security control real controller was punished for leaking, friends, comrades were fined more than 25 million yuan" We found one instance of insider trading.

On August 17, 2021, Yu Ling, the controlling shareholder and actual controller of ST Security Holding Company, received the Notice of Case Filing from China Securities Regulatory Commission (CSRC). Due to his suspected insider trading (disclosure of inside information) in the securities market, CSRC decided to file a case against him in accordance with relevant laws and regulations.

On January 26, 2022, the Xinjiang Bureau finally released the administrative penalty decision, and also released the specific details of the case.

On December 12, 2022, Yu Li, another independent director of the company, was registered by the China Securities Regulatory Commission on suspicion of insider trading of the company's shares.

DOI: 10.54254/2754-1169/21/20230238

Zhang Lei, general manager of the company, Li Chunfu, vice Chairman, and Zhang Bin, chief financial officer, made millions of yuan in profits during this period (non-sensitive period for insider trading), among which Zhang Lei and Li Chunfu were listed on the insider buying list.

3.11.3. Overall Data of Legal Insider Trading in the Market.

Since we have provided a case of insider trading which make a profit for insider, we are also curious about how the overall data of legal insider trading in the market is. If it has already provided an enormous return for overall legal insider traders, we cannot imagine how big "the iceberg under the sea level" is, which is the profit made by all illegal informed trading.

We calculate the overall return of published insider traders in the whole equity market. The result is shown in Appendix due to the page limits. Just simply apply the “follow the published insider trading” strategy, we can get an incredible excess earning! However, we doubt whether using these kinds of strategies is legal, as it utilizes information that has not arrived at the market.

Thus, we strictly follow the regulation rules mentioned above and conduct our new factor exploration only by retrospecting the possibility of insider trading after the information arrived instead of just detecting the insider trading and following it.

3.11.4. New Factor Construction.

In the above cases, we have shown the existence and manifestation of insider trading. Now we develop a new factor to characterize it.

Following Holden and Jacobsen (2014), we calculate order imbalance where the number of buys and sells per day are identified using the algorithm of Lee and Ready (1991) (shown in 2.4).

We backtrack each stock for the past 30 trading days, calculate the absolute order imbalance during this period, and weight it according to the number of days from the report release time.

Order imbalance × daily urgency

$$\sum_{i=1}^{30} oib\big(d(t-i)\big) \times \frac{1}{i+10} \tag{4}$$

Here is the economic meaning of our new factor:

Aktas, de Bodt, Declerck, and Oppens shows that if the probability of information arrival is constant, the probability of informed trading is equivalent to the order imbalance [7].

However, in a period of time before the release of major events, the rate of information attainment was not balanced. Therefore, we revised the order imbalance according to the number of days before the announcement to better measure the probability of insider trading in stocks and build our new factor.

4. Experiment

4.1. Expected Results

Based on our field research and observations in the Chinese stock market, we expect that the existence of informed trading might reflect the deficiency of management ability of a public company, which may lower the long-term expected earning rate of the stock.

In the short term, we expect that the manifestation pattern of informed trading might be similar to the reverse momentum factor, as the insider may make a transaction in the reversed direction to leave the market after the information of major events is released.

Since big-cap stocks is relatively hard to manipulate and hardly involved with insider trading, we also expect that our factor performs much better in small-cap stocks than big-cap stocks.

DOI: 10.54254/2754-1169/21/20230238

4.2. Experiment Setting

We first conduct the single-factor backtest to show the validity, stability, and monotonicity of our novel factor in short-term and long-term stock selection respectively. Then we combine our factor with numerous existing valid factors on the multi-factor model to gain a even better performance.

We conduct our experiment mainly in the stock pool of the CSI Smallcap 500 Index. The CSI500 Index contains high-quality stocks with small caps, which are easier to be involved with insider trading than the stocks in HS300. In the past twenty years, the SCI500 perform much better than HS300.

We also compare the effectiveness of our factor in two different stock pools (HS300 and CSI500) to provide insight into which kinds of stocks that our factor suits better.

Our backtesting period is set from January 1, 2013, to December 01, 2022.

4.3. Single Factor Test

As we mentioned in the previous section, the order imbalance indicator has two major effects, expressed as the negative effect in the long term and the reverse momentum effect in the short term. We discuss these two effects separately. For short-term effect, we calculate the momentum in 30 days period (which is the same as OIB), assign the OIB with a negative sign if the momentum is positive (short position), assign the OIB with a positive sign if the momentum is negative (long position). Then we calculate the return from both the short position and long position with top 10 stocks every 30 days. For long-term effect, we simply select the top 100 stocks (1/5 of stock pool) with the lowest OIB indicators.

We construct a backtest using our novel factor during January 1, 2013, to December 01, 2022 in HS300 stock pools in the DataYes Platform.

The results are shown below.

To sum up the result, the short-term OIB factor achieve a 16.1% average earning rates for 10 years, the long-term OIB factor achieve a 12.2% average earning rates compared to the 7.0% benchmark of CSI500. Both of them also achieve a way better result on indicators such as IC, IR and sharp ratio (See figure 1-4). At the same time, the information ratio IR can be calculated (See figure 5).

Figure 1: Backtest result on long position in 2013-2017.

DOI: 10.54254/2754-1169/21/20230238

Figure 2: Backtest result on long position in 2018-2022.

Figure 3: Backtest result on short position in 2018-2022.

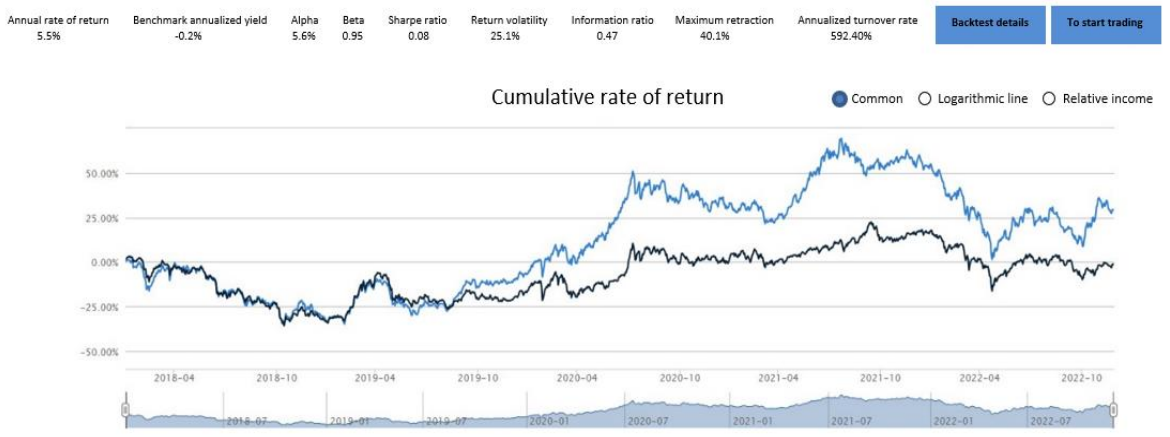

Figure 4: Backtest result with long-term effect in 2018-2022.

DOI: 10.54254/2754-1169/21/20230238

Figure 5: Visual display of layered back testing of PB factor.

4.4. Test Monotonicity of Single Factor

The layered back testing method is used to test the factor monotonicity. This is the quintile. The monotony of a factor means that the effect of the equal weight strategy of the stocks with the highest factor value ranking interval will be better than that of the stocks with the highest factor value ranking interval at any time. In the following tables, we rank the stocks in CSI500 based on short-term OIB and backtest to show the monotonicity of our factors (See table 1).

Table 1: Quantile test of our factor among CSI500 Index stocks.

stocks' rank	Annualized rate of return	Benchmark annualized rate of return	α	β	Sharpe ratio	Volatility	Information ratio	Maximum fallback	Annual turnover rate
1-10	0.0796	-0.0022	0.0708	0.7	0.22	0.2064	0.5	0.3682	17.18
11-20	0.0915	-0.0022	0.0848	0.76	0.26	0.2134	0.63	0.3474	21.16
21-30	0.0232	-0.0022	0.0156	0.74	-0.06	0.2059	0.15	0.3802	20.49
31-40	0.0264	-0.0022	0.023	0.85	-0.04	0.2282	0.22	0.3991	22.42
41-50	0.0549	-0.0022	0.0518	0.86	0.09	0.2322	0.43	0.4116	23.1
51-60	0.0351	-0.0022	0.0335	0.9	0	0.2382	0.31	0.3615	22.97

4.5. Results in Big-cap Stocks

We apply our novel factor in the stock pool of HS300, the result shows that our factor does not work in big-cap stocks (Overall -0.8% average learning rate compared to -0.2% benchmark). It indicates that our factor suits small-cap stocks better, which is accordant to our expected results.

4.6. Combination Mode of Multiple Factors

The first ten factors with the best IC among the factors with statistical significance (t<. 05) and monotonicity (selected manually according to the hierarchical back test effect) were selected from the factor pool as the multiple factors for preliminary selection. Screen out one factor with small IC value from the factors with large correlation, solve the problem of multicollinearity, and reduce the correlation between factors while retaining the explanatory power of factors (See figure 1).

4.7. Multifactor Backtesting

In all cross-section periods, we conducted regression tests on factor d, and we can obtain the factor return rate series of the factor (that is, the series formed by the regression coefficient X_{T_d} in all cross

DOI: 10.54254/2754-1169/21/20230238

section periods) and the corresponding t value series. The t-value refers to the t-test statistics of the single regression coefficient X_{T_d}, which describes the significance of a single variable. The absolute value of the t-value is greater than the critical value, indicating that the variable is significant. That is, the explanatory variable (the exposure of individual stocks in the factor d of the T period) is a factor that really affects the dependent variable (the income rate of individual stocks in the T+1 period).

5. Conclusion

We build a multi-factor stock portfolio forecasting model. We first find a novel factor based on insider trading detection with a solid effect. Then we use a quantile test to ensure that the selected single factor is monotonous. Finally, we combine them by the equal weight method and HMM model to obtain the corresponding multi-factor model. We use the model to estimate the stock price and confirm the feasibility of our model through backtesting.

To build up our novel factor based on insider trading, we use the order imbalance(OIB) algorithm to integrate the possibility of insider trading into a factor, and then add it to the model for prediction and comparison. We found that in stocks, the prediction results of the model after adding insider trading factors are closer to the real value, while in the points with more accurate prediction, our prediction results are almost unchanged. From this, we can see that the construction of the insider trading factor is reasonable and effective. We also show that our factor works better in small-cap stocks than big-cap stocks, as it has a higher possibility and magnitude involved with insider trading.

References

[1] *George HK Wang and Jot Yau. 2000. Trading volume, bid–ask spread, and price volatility in futures markets. Journal of Futures Markets: Futures, Options, and Other Derivative Products 20, 10 (2000), 943–970.*

[2] *Darryl Shen. 2015. Order imbalance based strategy in high frequency trading. Ph. D. Dissertation. oxford university.*

[3] *Jonathan M Karpoff. 1987. The relation between price changes and trading volume: A survey. Journal of Financial and quantitative Analysis 22, 1 (1987), 109–126.*

[4] *Shai Fine, Yoram Singer, and Naftali Tishby. 1998. The hierarchical hidden Markov model: Analysis and applications. Machine learning 32, 1 (1998), 41–62.*

[5] *Ananth Madhavan, Matthew Richardson, and Mark Roomans. 1997. Why do security prices change? A transaction-level analysis of NYSE stocks. The Review of Financial Studies 10, 4 (1997), 1035–1064.*

[6] *Kenneth R Ahern. 2020. Do proxies for informed trading measure informed trading? Evidence from illegal insider trades. The Review of Asset Pricing Studies 10, 3 (2020), 397–440.*

[7] *Nihat Aktas, Eric De Bodt, Fany Declerck, and Herve Van Oppens. 2007. The PIN anomaly around M&A announcements. Journal of Financial*

[8] *Jonathan M Karpoff. 1987. The relation between price changes and trading volume: A survey. Journal of Financial and quantitative Analysis 22, 1 (1987), 109–126.*

[9] *Rama Cont, Arseniy Kukanov, and Sasha Stoikov. 2014. The price impact of order book events. Journal of financial econometrics 12, 1 (2014), 47– 88.*

[10] *Ruey S Tsay. 2005. Analysis of financial time series. John Wiley & sons.*

[11] *Pierre Collin-Dufresne and Vyacheslav Fos. 2015. Do prices reveal the presence of informed trading? The Journal of Finance 70, 4 (2015), 1555– 1582.*

[12] *Md Rafiul Hassan. 2009. A combination of hidden Markov model and fuzzy model for stock market forecasting. Neurocomputing 72, 16-18 (2009), 3439–3446.*

[13] *Md Rafiul Hassan and Baikunth Nath. 2005. Stock market forecasting using hidden Markov model: a new approach. In 5th International Conference on Intelligent Systems Design and Applications (ISDA'05). IEEE, 192–196.*

[14] *Lu Chao, and Zhang Siyu. 2022. Insider Trading by Non-Executive Directors and Listed Companies: Empirical Evidence from China's A-share Market. Journal of Central University of Finance and Economics 5 (2022), 72–83.*

[15] *ZHANG Xudong, HUANG Yufang, DU Jiahao, and MIAO Yongwei. 2020. Stock price prediction based on discrete hidden Markov model. Journal of Zhejiang University of Technology 48, 2 (2020), 148–153.*