Predicting Liquidity Ratio of Mutual Funds via Ensemble Learning

Kun Kong
China Asset Management Co., Ltd.
Beijing, 100033, China
kongk@chinaamc.com

Ruicong Liu
China Asset Management Co., Ltd.
Beijing, 100033, China
liurc@chinaamc.com

Yihui Zhang
China Asset Management Co., Ltd.
Beijing, 100033, China
zhangyh@chinaamc.com

Yixin Chen
China Asset Management Co., Ltd.
Beijing, 100033, China
chenyx@chinaamc.com

Abstract—We are entering a new era of AI, in which the core technology is machine learning. However, many machine learning models are opaque and not intuitive enough, making it difficult for users to understand how AI systems make decisions. In some scenarios, especially in the fields of finance, healthcare, automatic driving, users have a strong demand for the interpretability of the model. Although AI systems provide a lot of benefits, if the decision and behavior cannot be explained to users or regulators, the effectiveness and development of the systems will be restricted. To gain the trust of users, Explainable AI is necessary.

Daily prediction of mutual fund holdings can be a very useful tool. If we can predict the daily holding positions of large mutual funds, we can gain insights into the sentiments of institutional investors which shed lights into the future trend, which is a good index of market sentiment. The funds in the whole market can reflect the investors’ views on the stock market. We leverage on another source of key information - the price of a mutual fund is updated daily, often released a few hours after the market closes. Therefore, we can utilize the daily price fluctuation, combined with quarterly revealed holdings, to make daily predictions of mutual fund holdings. In this paper, we proposed an Ensemble Learning model to predict liquidity ratio of mutual funds. The model has strong interpretability, which is beneficial to users, developers and regulators and all parties involved. Compared with the real fund position data only disclosed once a quarter, our model effectuates timely and efficient high-frequency calculation. In the process of modeling, we creatively apply the framework of Ensemble Learning to portfolio decomposition for the first time. The Ensemble Learning model leverages the diversity of base learners to improve the overall prediction performance. Extensive empirical results on China A-Share market show that our model can achieve superior accuracy, robustness, and generalization ability.

Keywords—XAI, liquidity ratio, mutual funds, Ensemble Learning, fund position

I. INTRODUCTION

Mutual funds have 15 working days after their quarter ends to file their holdings with the SEC. Some funds choose to report their holdings even more frequently, on a monthly basis, but this is less of a norm, as monthly reporting requires a lot of effort and cost on the part of mutual funds. One of the more common items seen in the list of mutual fund holdings reporting is the top 10 holdings that the fund owns. This is usually updated quarterly and is made available on the company’s website quite quickly, within a few weeks. But still, such information is delayed. This means that an investor is not accurately aware of how the money is being invested. This is especially true with funds that have high turnover ratios. Mutual funds argue that the delay is necessary to protect them from strategies such as front-running.

Daily prediction of mutual fund holdings can be a very useful tool for analyzing institutional investors. On the one hand, we can see the judgment of institutional investors with professional knowledge on the current market trend from their shareholding position, on the other hand, we can infer the allocation weight of funds among various assets. If we keep tracking, we can also observe how the capital flows between various types of assets, which is valuable for investors.

The average positions of institutional investors can be an indicator of market sentiments. When the fund manager anticipates that the stock market will usher in a big rise, he will increase his position in stocks. On the contrary, when he is pessimistic about the future market, he will take the initiative to reduce his position. In China, due to regulatory restrictions, the stock position of an actively-managed mutual fund is prohibited to be lower than a certain threshold. Even in the bear market in 2015, many equity funds can only carry through with the minimum position threshold. The average holding position of all funds in the whole market can reflect the investors’ views on the future trend, which is a good index of market sentiment. The relationship between the Shanghai stock index and the average holding position of all aggressive mixed funds is shown in Fig. 1.
Fig. 1 shows the comparison of fund positions and stock market trends from 2013 to early 2020. The blue line is the average real position level of all aggressive mixed funds in the whole market, corresponding to the left coordinate axis. And the orange line is the Wind A-Share Index, corresponding to the right coordinate axis.

As you can see, as long as you buy in the red circle in Fig. 1 when the position of stock fund is low to a certain extent, this simple strategy can easily capture the gains from the rise of the Shanghai Composite Index. Of course, this is just a simple example, and the actual application of fund positions may be much more complicated. From the performance of recent years, the trend of the two lines has been very consistent. This shows that the fund position level can accurately reflect the sentiment of institutional investors. No fund can guarantee that the judgment of the market is accurate at any time. When we conduct statistical analysis on a large number of fund positions, we can see the sentiment of professional institutional investors in the market from the average position level, which is an excellent indicator of investor sentiment.

Through the new model introduced in this paper, we can calculate the average holding position of the fund as an index of investor sentiment, and then we can combine the knowledge of Behavioral Finance to construct the investment strategy. Sentiment affects the rise and fall of the market, and behavior determines the success or failure of investment. Investor sentiment index is an important research object of Behavioral Finance.

Behavioral finance tells us that sentimental changes of irrational investors will magnify their emotional bias and cognitive bias, leading to wrong asset pricing. Many studies [1]-[6] show that the price of assets is affected by investor sentiment, which will lead to a certain deviation on the basis of its real value. In the long run, investor sentiment will gradually return to normal, and the asset price will eventually return to its true value. Therefore, according to these studies, there is a negative correlation between investor sentiment and asset return in the future. In conclusion, as an indicator of investor sentiment, fund holdings position data can indicate the trend of asset prices and provide support for investment research.

In this paper, we propose a novel Ensemble Learning method with strong interpretability for predicting liquidity ratio of mutual funds. Research shows that there is a trade-off between the performance and interpretability of machine learning models[7]. With the increase of the complexity of the model, it is more difficult for users to understand the model, which will reduce the user's trust in the model. The model proposed in this paper can not only restrain the over-fitting effect of the model to a certain extent, but also maintain the basic principles of decision-making, so that users can better understand the AI system. The new model not only guarantees strong predictability, but also has a very logical way of thinking in solving problems. It has achieved a delicate balance between the performance and interpretability of the model.

By analyzing institutional holding positions and their sentiment, we can generate over-buy and over-sell signals which is a proven reliable strategy. Extensive evaluation results show that our proposed method can achieve state-of-the-art performance, achieving superior prediction accuracy than other existing methods. At the same time, the linear model involved in this paper can better explain the formation and decomposition of fund returns. Users fully understand and trust our model, which is crucial for the influential financial forecasting problem.

II. RELATED WORKS

Holdings-based performance attribution, also known as the internal evaluation rule, refers to the analysis of portfolio performance based on real transaction data of the fund in question. Brinson and Fachler [8] and Brinson, Hood, and Beebower [9] introduced the Brinson models as a foundation for investment portfolio performance attribution. These models attribute returns from active management into: security selection (return achieved through selecting different securities than the benchmark), and asset allocation (return achieved through weighting asset classes in a portfolio differently than the benchmark). The Brinson-Fachler methodology underpins many public performance attribution analyses adopted by many rating firms including Morningstar.

Return based performance attribution, also known as external evaluation method, is a type of fund performance attribution methods that make use of the fund net value curves and benchmark performance. It is usually based on the extended form of CAPM model [10] and multi-factor models, such as TM model [11], HM model [12], Fama and French’s three-factor model [13][14] and Carhart's four-factor model [15].

However, none of the above methods can directly tell the daily positions of mutual funds. Sharpe [16] defined the asset class factor model, which linked the portfolio return with the return on assets representing different investment styles, and estimated the allocation weight of various assets through multiple regression statistical analysis. His approach is a pioneering work on decomposition of portfolio net worth.

Fung and Hsieh [17] inherited Sharpe’s idea, applied net value decomposition to CTA hedge funds, and used Principal Component Analysis (PCA) to decompose fund returns into several different investment styles, such as "Systems", "Technical", "Trend Following", etc., so as to analyze the investment style and find the stable dominant investment style of a CTA fund.

Hasanhodzic and Lo [18] decomposed the yield of hedge funds into six risk factors, namely USD, BOND, CREDIT, S&P500, GSCI and DVIX. Based on the factor exposure value calculated by multiple linear model, a cloned hedge fund is constructed by using passive investment instruments. It has lower cost, but can achieve the risk return characteristics similar to the target hedge fund. However, from empirical results, the performance of cloned portfolios is not as good as that of real hedge funds.

Amenc [19] extended the research of Hasanhodzic and Lo and tried to clone hedge funds using nonlinear methods. The results showed that complex nonlinear methods did not enhance the replication effect. Amenc also repeated Hasanhodzic and Lo's experiment, confirming that the performance of the linear replication portfolio is indeed inferior to that of the real fund.

Byrd [20] selected the most likely stock assets of a fund by introducing a new Sequential Oscillatory Selection(SOS) method, and its prediction accuracy has been improved compared with the traditional linear cloning method. The core principle of the SOS algorithm is based on the SFFS [21].
algorithm. Starting from an empty set, it finds the most likely stock in each iteration until the performance is no longer improved. However, as a greedy algorithm, it typically gets stuck a local optima and can leave a lot of room for improvement.

Some recent research has emerged for liquidity ratio prediction. Chan and Li [22] proposed a method to predict fund positions. In their work, the daily return of a Chinese mutual fund is treated as the response variable, while the daily changes of CITIC industry indices are used as explanatory variables. The coefficients of industry indices are estimated by quadratic programming, stepwise regression and Lasso regression, and the coefficients are summed up to obtain the overall position of the fund.

In this paper, we introduce the framework of Ensemble Learning [23] into portfolio analysis for the first time. Its core idea is to construct several "good but different" weak learners, and get a stronger learner through certain synthesis strategies to make the final decision. Ensemble Learning makes use of the independence of the basic model and improves the generalization ability of the system significantly.

III. DATA PREPARATION

In addition to stocks, funds usually hold bonds, inter-bank notes, cash and other assets. However, the changes of these net assets are very small compared with stocks. Therefore, it is beneficial to simplify the model reasonably by considering the fund's position assets as the two parts of stocks and cash [24]. Therefore, the research object of this paper is the common stock funds (stock position > 80%) and aggressive mixed funds (stock position > 60%) in the Wind open-end fund classification standard. The stock position of these funds is relatively high, and the position of other assets such as bonds is relatively low. Ignoring these assets has little impact on the calculation. There are some other specific requirements for the screening of research subjects, which will be explained in detail below.

A. Basic Data

All of our data on funds, stocks, and indices are from Wind financial databases.

- Fund market data: the real stock positions disclosed by 332 general stock funds and 809 aggressive mixed funds in each quarter from 2010 to 2019, as well as the net value sequence of fund restoration units on each day, are used to calculate the daily yield of funds. In order to eliminate some unstable factors of newly established funds, funds less than one year old are excluded. For a fund with multiple share types, such as A, C, E, F, H shares, only class A share fund is studied.

- Fund data: details of the top ten holdings disclosed by the above funds in the quarterly reports from 2010 to 2019 are collected.

- Index market data: daily closing prices of China Securities Exchange scale index, giant tide style index and CITIC primary industry index from 2010 to 2019 are used to calculate daily yield.

- Stock market data: daily stock returns of nearly 3700 A shares and nearly 400 Shanghai-Hong Kong Stock Connect stocks since 2010 to 2019.

B. Derived data: Industry Clustering

In order to make the data suitable for the model, we need to further process the CITIC primary industry index. According to the clustering results of industry classification, 28 CITIC first-level industry indices can be divided into several major industries.

Specifically speaking, except for the comprehensive category, the monthly returns of 28 CITIC first-level industry indices since 2005 have been recorded, and a total of 28 industry samples are used as the input to a hierarchical clustering model.

On the basis of this clustering result, the stocks in the whole market are divided into five categories: Consumer, Midstream manufacturing, TMT, Infrastructure and Cycle, and Finance and Real Estate. The classification results are shown in Table 1.

<table>
<thead>
<tr>
<th>Major Industry</th>
<th>CITIC Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumer</td>
<td>Light Industry Manufacturing</td>
</tr>
<tr>
<td></td>
<td>Textile Clothing</td>
</tr>
<tr>
<td></td>
<td>Agriculture, Forestry, Animal</td>
</tr>
<tr>
<td></td>
<td>Health Care</td>
</tr>
<tr>
<td></td>
<td>Food and Beverage</td>
</tr>
<tr>
<td>Midstream Manufacturing</td>
<td>Building Material</td>
</tr>
<tr>
<td></td>
<td>Machinry</td>
</tr>
<tr>
<td>TMT</td>
<td>Automobile</td>
</tr>
<tr>
<td></td>
<td>Household Electrical Appliances</td>
</tr>
<tr>
<td></td>
<td>Basic Chemical Industry</td>
</tr>
<tr>
<td></td>
<td>Power Equipment</td>
</tr>
<tr>
<td>Infrastructure and Cycle</td>
<td>Defense Industry</td>
</tr>
<tr>
<td></td>
<td>Electronics</td>
</tr>
<tr>
<td></td>
<td>Signal Communication</td>
</tr>
<tr>
<td></td>
<td>Computer</td>
</tr>
<tr>
<td></td>
<td>Medias</td>
</tr>
<tr>
<td>Finance and Real Estates</td>
<td>Petroleum and Petrochemical Industry</td>
</tr>
<tr>
<td></td>
<td>Steel</td>
</tr>
<tr>
<td></td>
<td>Electricity and Utilities</td>
</tr>
<tr>
<td></td>
<td>Transportation</td>
</tr>
<tr>
<td></td>
<td>Architecture</td>
</tr>
<tr>
<td></td>
<td>Coal</td>
</tr>
<tr>
<td></td>
<td>Nonferrous Metals</td>
</tr>
<tr>
<td></td>
<td>Banks</td>
</tr>
<tr>
<td></td>
<td>Non-bank Finance</td>
</tr>
<tr>
<td></td>
<td>Real Estates</td>
</tr>
</tbody>
</table>
Finally, by calculating the market index of these small industries weighted by market value, we can get the new five major industry market indices, which can be retained as derivative data and can be used as a new stock classification standard. The trends of the composite indices are shown in Fig.2.

IV. PROPOSED METHOD

A. Main Idea and Model Assumptions

The main idea of our approach to predict fund positions is the decomposition of fund daily return. The general calculation process is as follows. First of all, the stock pool formed by all stocks in the stock market should be divided into several classes with low correlation, and the daily return of each class of stocks can be directly represented by the ready-made index daily return. Then, the daily return of the fund is decomposed into the daily return of these assets, so that the error of the decomposition of the daily return of the fund is minimized in the time window of the return period, and the position allocation weight of each type of stock assets is obtained. Finally, the positions of these asset classes are summed up, which is the stock position of the fund.

For an equity fund $F$, the liquidity ratio, also known as cash level, $L(F)$ is the percentage of $F$’s total assets (AUM) that are held in cash or cash equivalents.

$$L(F) = 1 - \sum_{k=1}^{n} \beta_k$$

Among them, $\beta_k$ is the position weight of class $k$, which makes the daily return series of funds $R_{f,t}$, approximately satisfy the following conditions:

$$R_{f,t} = \sum_{k=1}^{n} \beta_k R_{k,t}$$

Where $R_{k,t}$ is the average daily yield of stocks in class $k$.

According to the above idea, the model implies at least three assumptions.

Assumption 1: the daily return of each class of stock assets in the fund portfolio should be approximately equal to the daily return of that class in the entire stock pool.

Assumption 2: the allocation weight of the fund on each class of stocks remains constant or changes little in the time window of the return period.

Assumption 3: the positions of bonds and other assets in the fund are low, so the influence of these assets on the measurement of stock positions is negligible.

Assumption 1 is often reasonable when the fund portfolio invest diversely enough in each class. Assumption 2 is required by multiple linear regression. Since the coefficients of independent variables should be constant, it should be assumed that the allocation ratio of various assets is relatively stable. The fluctuation range of daily yield of bonds and other assets held by the fund is generally very small relative to stocks, so it is reasonable to add Assumption 3 in regression analysis.

To further meet Assumption 3, we limit our study to mutual funds that invest heavily in stocks. We select the 332 stock funds (stock position > 80%) and 809 aggressive mixed funds (stock position > 60%) by the Wind classification standard. The positions of these funds' stocks are relatively high, and the positions of other assets are low.

B. Industry decomposition

In order to estimate fund positions accurately, all available public data should be used as much as possible. Each time we estimate the position of a fund, we can first find the latest disclosed holdings of the fund as part of the source data. There are two main types of public fund shareholding information:

- Semi-annual report or annual report. In terms of portfolio information disclosure, public funds must disclose complete stock position information every six months, including all stocks and their positions. However, the limitation is that the deadline for the disclosure of the fund's semi-annual report is 60 days after the end of the first half of the year, while the deadline for the disclosure of the fund's annual report is longer, within 90 days from the end of each year.

- Quarterly report: the quarterly report of public funds must be disclosed within 15 working days after the end of each quarter, which contains the information of top ten stock and their specific positions.

These two kinds of shareholding information have their own advantages and disadvantages. The semi-annual report and annual report have complete shareholding information, but the timeliness is poor. Although we can see all the position information, when it is disclosed two or three months later, the position is likely to have changed a lot. Although the quarterly report can only see the top ten positions, but the disclosure is more timely, and funds that hold heavy positions in stocks are often relatively optimistic, and the probability of continued holding after the disclosure of the top ten positions is relatively higher. After weighing the advantages and disadvantages, we choose to use the holding information disclosed in the quarterly report to supplement the existing model.

After introducing the position information of the fund's top holdings, the daily return of the fund should mainly include two
parts: one is the income from the stocks continuously held in the
top ten positions in the last quarter; the other part is the income
from the remaining stocks after excluding the top ten stocks
from all the stocks and dividing them into different types of
stock assets according to the traditional stock classification
method. It can be written as the following multiple regression
model:

\[ R_{ft} = \sum_{j=1}^{10} (b_{Sj} + \beta_{Sj}) R_{Sj, t} + \sum_{k=1}^{n} \beta_{Ik} R_{Ik, t} \]  

(3)

The position meets the constraint:

- 60% ≤ \( \sum_{j=1}^{10} (b_{Sj} + \beta_{Sj}) + \sum_{k=1}^{n} \beta_{Ik} \) ≤ 95%  
  for aggressive mixed fund
- 80% ≤ \( \sum_{j=1}^{10} (b_{Sj} + \beta_{Sj}) + \sum_{k=1}^{n} \beta_{Ik} \) ≤ 95%  
  for equity funds
- 0 ≤ b_{Sj} + \beta_{Sj} ≤ 95%, ∀ 1 ≤ j ≤ 10
- 0 ≤ \beta_{Ik} ≤ 95%, ∀ 1 ≤ k ≤ n

Among them:

- n is the number of classes of stock assets
- R_{ft} is the daily yield of the fund
- R_{Sj, t} is the daily yield of the j-th heavy position stock
  on day t
- R_{Ik, t} is the overall daily yield of class k stock assets on
day t
- b_{Sj} is the position allocation weight of the j-th heavy
  position stock disclosed in the last quarterly report
- \beta_{Sj} is the change of position of j-th heavy position stock
  relative to the last disclosure position, which is an
  incremental concept
- \beta_{Ik} is the position allocation weight of class k assets

The general form of linear regression is obtained by moving
the constant term to the left of the equal sign

\[ R_{ft} = \sum_{i=1}^{10} b_{Sj} R_{Si, t} + \sum_{j=1}^{10} \beta_{Sj} R_{Sj, t} + \sum_{k=1}^{n} \beta_{Ik} R_{Ik, t} \]  

(4)

The above equation accommodates the special case where
the top 10 holdings in the last quarter have been fully cleared
on the day of estimation. In fact, this has been taken into account
in the model, and \( b_{Sj} + \beta_{Sj} = 0 \) is the case.

In the CITIC industry classification standard, all stocks
are divided into 28 industries. Because there are too many
independent variables in the industry classification model, Lasso
regression is needed to further reduce the impact of
multicollinearity. Lasso regression adopts L1 regularization,
which can produce sparse regression coefficient solutions and
effectively reduce the multicollinearity of regression models.
The optimal objective function of Lasso regression is as follows:

\[
\min_{\beta} \frac{1}{2\tau} \sum_{i=1}^{\tau} (R_{f,i} - \sum_{j=1}^{10} b_{Sj} R_{Si, t} - \sum_{j=1}^{10} \beta_{Sj} R_{Sj, t} - \sum_{k=1}^{n} \beta_{Ik} R_{Ik, t})^2 \\
+ \lambda(\sum_{j=1}^{10} |\beta_{Sj}| + \sum_{j=1}^{10} |\beta_{Sj}|')
\]  

(5)

Among them:

- \( \tau \) is the length of the regression time window (in days),
- \( \lambda \) is the L1 penalty coefficient of lasso regression.

By adding an L1 regularization penalty term to the right of
the least square optimization function, the L1 normal form of the
regression coefficient, that is, the allocation proportion of the
fund on various types of stock assets, is punished. The regression
coefficient is reduced and produced a sparse solution. The final
effect is that the estimated position of the fund in some industries
is 0.

Therefore, we can eliminate some redundant industries by
Lasso regression before entering the final estimation step, and
use L1 regularization to achieve the effect of feature selection,
so as to find some industries that are most likely to be held by
the fund. This is consistent with the actual situation of fund
investment. There is rarely a fund that invests in 28 industries
at the same time, and there should always be some industries
with a zero weight. In this way, the number of categories of stock
assets in the fund portfolio is reduced. To a certain extent, it can
also reduce the multicollinearity between explanatory variables
and improve the accuracy of the model. Finally, these selected
industries are used for the final estimation.

C. Quadratic Programming

Generally speaking, the least-square method can be used to
estimate the position results after finding the most likely asset
class of the fund by feature selection. However, different from
the ordinary linear regression problem, it is necessary to impose
linear inequality constraints on the positions of each type of
assets and the total positions when estimating positions. When
solving linear regression problems with inequality constraints,
the least-square method is no longer applicable. It is necessary
to transform the regression problem into an optimization
problem and then solve it by quadratic programming.

In the process of decomposing fund returns, linear regression
assumes that the position weight of each day in the regression
time window is the same or changes little. In fact, the older
the sample is, the more likely its asset allocation ratio will change
during this period, and the worse the timeliness of the sample;
on the contrary, the newer the sample is, the closer its asset
allocation ratio is to the current real situation, and the more
valuable the sample will be. In other words, the value of the new
sample data is greater than that of the old sample data, and the
difference of sample value should be reflected in the process
of modeling, so that the new sample can play a greater role.

In order to meet the inequality constraints and reflect the
difference of sample value, the model is optimized based on the
quadratic programming method instead of least-square method
when solving the regression equation (4). The error of each
sample is given different weights, and the quadratic
programming is used to solve the optimization problem. The
model pays more attention to the performance of new samples, and enlarges the error weight of new samples, or reduces the error weights of old samples. In this way, each square error is multiplied by a sample weight in the range of $1/4$ to $1$. The weight of the sample starts from close to $1$ and rapidly decays with $\tau/2$ as half-life, that is, the weight of the latest sample is close to $1$, the weight of the oldest sample is $1/4$, and the weight of the middle sample is about $1/2$. The quadratic programming problem can be described as follows:

$$\min_{\beta} \frac{1}{2\tau} \sum_{i=1}^{\tau} w_i \left[ (X^T W X) \beta - \frac{1}{2} \beta^T (X^T W y) + \frac{1}{2\tau} y^T W y \right]$$

s. t.

- $0 \leq \beta_i \leq 95\%$
- $60\% \leq \sum_{i=1}^{p+q} \beta_i \leq 95\%$ for aggressive mixed fund
- $80\% \leq \sum_{i=1}^{p+q} \beta_i \leq 95\%$ for equity funds

Among them:

- $p$ is the number of remaining shares of the top 10 stocks after feature selection
- $q$ is the number of stock classes left after feature selection
- $y$ is a time series composed of the daily return of the fund within $\tau$ days before the estimated date minus the return rate of the top ten positions disclosed in the previous quarter
- $\beta$ is a $(p+q)$ dimensional vector composed of positions of all assets
- $X$ is the matrix formed by the daily return rate of all assets in this $\tau$ day, with the dimension of $\tau \times (p+q)$
- $w_i = \left( \frac{1}{2} \right)^{\frac{\tau}{2}}$ is the weight of sample error
- $W = \text{diag}(w_1, w_2, \ldots, w_\tau)$ is a diagonal matrix with $w_i$ as the main diagonal element

The rest of the letters have the same meaning as the previous formula.

Through the quadratic programming method, the position $\beta$ of each kind of stock assets in the fund portfolio is calculated. Finally, the sum of all weights in $\beta$ is calculated to obtain the estimated value of fund position.

**D. Selection of regression window length**

The market is changeable, but no matter how the market changes, when a fund manager manages a fund, its investment style is usually relatively stable. If the turnover rate of a fund does not change significantly, we can find a regression window length that is most suitable for the fund from the past history. This optimal time window length based on historical data is also likely to be applicable to the current fund position estimation. In order to find the best time window length, we use the real fund position data disclosed by the fund to be tested every quarter. Through grid search, an optimal window length in the range between 10-40 working days, is found to minimize the average error in historical backtesting.

Similarly, generally speaking, the investment fields that a fund focuses on and excels in will not change greatly in a short period of time. Therefore, regardless of the size of a fund, the industry distribution of a same manager tends to remain stable to some extent. Thus, we can use historical data to select the L1 penalty in lasso regression. For each fund product, the best L1 penalty coefficient alpha can be found by backtesting the true historical position data of each fund product.

We can make prediction model better by tuning the hyper-parameters based on historical data, with the assumption that the investment style of a fund to be tested does not change rapidly. More specifically, we assume that the turnover rate and industry-wise concentration of a fund do not change too rapidly and dramatically over time.

Through grid search, each fund can find the appropriate regression time window and L1 penalty coefficient to adapt to the turnover rate and shareholding concentration of the fund. Table 2 shows the industry model’s adaptation to the turnover rate and shareholding concentration of several funds under China Asset Management Company. After adjustment, the model can estimate each fund as accurately as possible. It enhances the universality of the model and reduces the model error systematically.

**Table II. The optimal window length and the number of independent variables of some funds**

<table>
<thead>
<tr>
<th>Fund Code</th>
<th>Window length</th>
<th>Number of explanatory variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>000001.OF</td>
<td>40</td>
<td>13</td>
</tr>
<tr>
<td>000011.OF</td>
<td>37</td>
<td>15</td>
</tr>
<tr>
<td>000021.OF</td>
<td>38</td>
<td>18</td>
</tr>
<tr>
<td>000031.OF</td>
<td>37</td>
<td>14</td>
</tr>
<tr>
<td>000061.OF</td>
<td>37</td>
<td>18</td>
</tr>
<tr>
<td>000945.OF</td>
<td>28</td>
<td>7</td>
</tr>
<tr>
<td>001042.OF</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>002011.OF</td>
<td>40</td>
<td>17</td>
</tr>
<tr>
<td>002229.OF</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>002871.OF</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>002980.OF</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>003567.OF</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>003834.OF</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>004640.OF</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>004686.OF</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>005449.OF</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>288001.OF</td>
<td>29</td>
<td>12</td>
</tr>
<tr>
<td>288002.OF</td>
<td>40</td>
<td>18</td>
</tr>
</tbody>
</table>
E. Ensemble Learning

The stock classification standard of the existing models is very monotonous, which is generally one of size, style or industry. The previous chapters introduce an example of an industry model, which classifies all stocks into 28 CITIC first-level industries. As long as the division standard of stock pool is changed, other traditional models can be generated according to the same modeling process. For example, according to the market value of stocks, stocks can be roughly divided into four categories: CSI 100, CSI 200, CSI 500 and CSI 1000. And according to the style index series released by CNINFO, the stock pool can be divided into six types of assets: Large Cap. Growth, Large Cap. Value, Mid Cap. Growth, Mid Cap. Value, Small Cap. Growth and Small Cap. Value.

However, for funds with different product characteristics in the fund market, their investment concerns are quite different: some funds are called medical funds, focusing on investment in medical related stocks, which are invested from the perspective of the industry; some funds are called growth funds, focusing on investing in high valuation growth stocks, which are invested from the perspective of the style; some funds are called large cap funds, mainly investing in large cap stocks, which are invested from the perspective of the size. For these different types of fund products, Industry model, Style model and Size model all have their own applicable measurement fund, but also have their limitations. It is not appropriate to use a single standard model for prediction.

Although there are three common stock classification methods in traditional models, in practice, one model can only choose one of the three to classify all stocks according to the same standard. However, no matter which classification standard is used to classify stocks, it cannot be applied to all fund products. Once a specific standard of stock classification is determined, it is doomed that this traditional model will be more suitable for measuring some funds than others. From this point of view, if there is a way to take into account the diversification of the stock pool division in the prediction, the model can learn from the strengths of others and achieve greater performance improvement.

Ensemble Learning is an advanced machine learning framework. It uses a large number of learners to learn and integrates the results of all learners according to certain strategies, in order to achieve better prediction effect and stronger generalization ability than any single learner. Generally speaking, most of the base learners of Ensemble Learning are individual learners with relatively weak performance. There are also many strategies to integrate learning outcomes. Taking the classification problem as an example, the simplest strategy is voting. After individual learners produce their own results, we take a simple majority vote to decide the final output.

Stacking is a hierarchical Ensemble Learning framework. The core idea of the algorithm is that in the first layer of the algorithm, a number of heterogeneous primary learners are trained separately, and the primary learners are kept independent of each other. In the second layer of the algorithm, a secondary learner is constructed. The output of the upper layer is taken as the input feature of the second layer, and the results of multiple primary learners are summarized. The final output achieves superior accuracy, robustness, and generalization ability. The pseudo code of algorithm flow is shown in Table 3.

<table>
<thead>
<tr>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input:</td>
</tr>
<tr>
<td>Traing set ( D = {(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)} )</td>
</tr>
<tr>
<td>Base learner algorithm ( L_1, L_2, \ldots, L_m )</td>
</tr>
<tr>
<td>Combination algorithm ( L )</td>
</tr>
<tr>
<td>Output:</td>
</tr>
<tr>
<td>( H(x) = h_1(h_1(x), h_2(x), \ldots, h_7(x)) )</td>
</tr>
</tbody>
</table>

In this study, our model adopts the Stacking Ensemble Learning framework to summarize the results of multiple primary learners.

In the first layer of the model, several base learner predict the position of a fund according to the model process described in the previous chapters. In addition to the industry model, the style model and the size model mentioned above, we can also use the derived data obtained by hierarchical clustering algorithm to divide the stock pool into five major industries: Consumer, Midstream manufacturing, TMT, Infrastructure and Cycle, and Finance and Real Estate, so as to obtain a new base learner.

Such a model architecture has strong scalability. As long as a new stock classification standard is given, a base learner can be added to the first layer of Ensemble Learning to enhance the ability of the whole system.

In the second layer, through some combination strategies, the output results of the first layer are summarized to get the final position calculation results. There are many kinds of combination strategies in the second layer, which can adopt various complex machine learning or deep learning models for regression, or simple linear models.

The final overall model framework is shown in Fig. 3. The first layer is composed of four sub models, from top to bottom are CITIC Industry model, Major Industry model, CNI Style model, and CSI Size model. These sub models all have similar principles, but different standards are used in the classification of stock pools. The measurement results of these models are transferred to the second layer. Due to the small number of base learners, it is easy to over fit and lack of interpretability by using complex combination strategies. Here, the linear model method is used to summarize the results of all models, and the finally result with stronger robustness is produced.
V. INTERPRETABILITY

Our model provides good interpretability in at least three aspects:

Traceable AI: The ability to track the prediction process, from the logic of mathematical algorithms to the nature of data. In our model, the reasoning logic of the sub model is clear. Some position prediction models use complex machine learning algorithms, they lose the basic reasoning logic, or directly feed other irrelevant data features to models to get predicted values. Different from these black box models, as described in Chapter 4, the whole calculation process of each sub model has been strictly supported by logic, which can be understood and trusted by users. The core idea of the sub model is to decompose the return of the fund. This idea can work because the daily return of the fund is calculated by the weighted average of the daily return of each asset in the fund portfolio, which is a positive process. What we do is to use the method of machine learning to deduce the weight of each asset by using the return sequence of funds and various assets. The decomposition of fund return is a process of reverse reduction along the calculation path of fund return. The role of machine learning in it is just a white box tool, and it does not fundamentally invent a prediction method. The model is like a perspective glasses, which can help to mine the deep-seated information contained in the open market information accurately through the method of machine learning.

Reasonable AI: The ability to understand the reasoning behind each individual prediction. Through our model, we can not only know the final total position result, but also get the total position result of each sub model. We can not only know the total position results of each sub model, but also know the holding positions of various assets in each sub model. The integration process is transparent and easy to understand. We adopt the parallel integrated learning architecture Stacking, and the whole model architecture is as simple and clear as shown in Fig. 3, which is clear and trustworthy. In contrast, in Boosting architecture, the results of base learners are transmitted serially, which makes it difficult to see the role of base learners in the whole system. The application of Stacking greatly improves the accuracy of the prediction under the premise of being interpretable.

Intelligible AI: The ability to fully understand the model on which AI decisions are made. We apply machine learning algorithm reasonably. Many machine learning algorithms are used in our model. The model adopts the Ensemble Learning architecture, which greatly improves the prediction accuracy. Hierarchical clustering algorithm is used to divide all stocks into several major industries. And Lasso regression algorithm is used in industry selection. Each machine learning algorithm itself is not complex. These algorithms are applied properly, and can be reasonably explained. They indeed bring improvement to our model.

It is because of these advantages that our method is strongly explanatory. Only after a complete understanding of how the AI system makes decisions can all participants benefit from it. The developers and maintainers of the system can better handle the abnormal conditions, the users of the system can use the AI system more confidently, and the supervisor can understand the function and principle of the system more clearly. All these advantages make it necessary to study Explainable AI.

VI. RESULT ANALYSIS

A. Results of a Single Fund

The upper part of the Fig. 4 is the unit net value chart of a fund product since 2011. After model decomposition, the details of various asset positions shown in the lower part of the Fig. 4 are obtained. Combined with the legend, we can see that the total stock position of the fund is only a summary result. In fact, the model can calculate the allocation proportion of funds in various stock assets and the flow of funds between different categories over time. If the model is accurate enough, then for some users, the detailed position information also means higher value, such as imitating some top funds to do dynamic industry configuration, and so on, which will not be discussed here.
B. Results of All Fund

Due to the low disclosure frequency of real position data, the result analysis method of fund position calculation model is different from the usual situation. Generally, the measurement can be carried out on a daily frequency, and the results should be compared with the real data every day. However, because the real position data can only be disclosed once a quarter, the model can only be verified on the frequency of the quarter. That is to say, in fact, only one day out of every three months can be verified. Fortunately, we have estimated the positions of a large number of common equity funds and aggressive mixed funds in the past 10 years. When the number of samples is large enough, even if the error analysis results are obtained by downsampling, it has certain statistical significance and can verify the performance of the model.

As shown in Fig. 5, the four lines are the average position results of the new model for all 332 stock funds and 809 aggressive mixed funds since 2010. The blue and black lines in Fig. 5 are the real stock positions at the end of the reporting period disclosed in the quarterly report, while the red and green lines are the estimated positions of the new model corresponding to the same time node. The stock position of stock fund is always higher than that of aggressive mixed fund.

It can be seen from Fig. 5 that the estimated results are usually close to the real value, and the trend is similar at many inflection points. For example, in the first half of 2015, when the stock market was in a good situation, the average position level of the whole market was at a high level, especially for the aggressive mixed fund, because the position adjustment was more flexible, the upward trend was more obvious. Generally speaking, the estimation accuracy is good, and the error is analyzed numerically.

C. Performance Comparison

The model error can be obtained by back testing the historical data of all eligible funds. Since one year after the establishment of the fund, the error between the real fund positions disclosed in each quarter and the estimated positions of the model is calculated. Finally, the average absolute error of all funds in all periods is taken as the estimation error of the model. As follows:

\[
\text{mean} (|\text{PosEst} - \text{PosReal}|)
\]

Among them, PosEst is the fund position estimated by the model, PosReal is the real fund position disclosed in the Fund Quarterly Report, and the mean function is the arithmetic average function. The calculation order must be the absolute value of the error first and then the average value, because the average error can only reflect whether all the measurement results are larger or smaller on the whole, and can not reflect the deviation degree of each measurement.

According to the defined error calculation method, the error analysis of the new model using Ensemble Learning and several sub models is carried out to form a comparison. The error results are shown in the Fig. 6.

The second gray column in the Fig. 6 is a benchmark for error. This benchmark is the error calculated from the real position data of one period behind, without any actual prediction. In other words, the real fund positions disclosed in the last quarter of each fund are used as the estimated value of fund positions in this quarter, so as to calculate the absolute error with the newly disclosed real positions.

Among the models, the scale model performed the worst, followed by the style model. The performance of the industry model obtained by hierarchical clustering is better than the model of scale and style similar to the number of its stock classification, which proves the effectiveness of clustering. Among several traditional models, the best performance is the industry model which integrates lasso regression feature selection method, so this traditional model is also the most widely used model in the industry in recent years.
Finally, the average absolute error of the new model in general stock funds is about 2.43 percentage points, and the average absolute error of all aggressive mixed funds is about 4.62 percentage points. It can be seen from Fig.6 that the first red error column is about 20% lower than that of other models in both general stock funds and aggressive mixed funds. Therefore, the idea of Ensemble Learning has brought great improvement to the model.

VII. CONCLUSIONS AND FUTURE WORK

This paper makes a great improvement on the existing fund position measurement method. The core principle of the new model is still to decompose the daily rate of return of the fund. By introducing more effective information of heavy positions and adopting the framework of integrated learning, the new model achieves better results than the existing position measurement models. Also using machine learning algorithm, our model does not lack the support of basic reasoning logic as some complex nonlinear machine learning models. Our model uses an intrinsic interpretable linear model and a transparent Ensemble Learning architecture. Through the rational use of machine learning technology, it has always maintained a good interpretability. After being put into daily use, it is found that there are also some parts of the new model that can be improved in the future:

Firstly, the new model is time-consuming and has a huge amount of computation. The calculation time of a few fund products is acceptable, but with the increase of the number of fund targets, the disadvantages of grid search will be exposed, and the computing time will be increased. At present, there is no good way to deal with it, only to use computers or servers with better performance. In the future, parallel distributed computing may be used to optimize the program.

Secondly, the number of Ensemble Learning sub models is still limited, there are only four. Generally speaking, the more the number of integrated learning base learners, the better the effect of integration. As long as we can find more classification standards and divide the stock pool formed by all the stocks into several types of stock assets that do not overlap each other, a primary learning period can be constructed according to the new classification standard.

Finally, machine learning is suitable for analyzing problems with a large number of samples. If there is a way to increase the sample size by increasing the sample frequency, theoretically, the accuracy of fund position measurement can be further improved.

REFERENCES