Proposal: Inflow and outflow forecasts of Yu 'ebao capital

Group member:
Lai Lin 1801212867
Liu Sheng 1801212891
Lian Di 1801212881
He SongTao 1801212852
Alimujiang 1801212778

1. Background

In the first half of 2019, the size of the Monetary Fund decreased by 356.162 billion from the end of last year. Monetary fund yields also continued to slump, with the latest 7-day annualized yield at around 2.53%. In this context, the growth of the scale of the currency funds accessed by Yu'ebao has also been affected. It is no longer the case that Yu'ebao can have an increase of 10 billion in scale.

There are many reasons why Yuebao's return rate have fallen, including micro and macro factors. At macro level, interest rate and stock market prosperity are the most dominant factors, and GC001 and margin balance can be their best proxy in our research. GC001 is interest rate of 1 day treasury buyback, reflecting the risk-free interest rate for one day. Margin balance reflects the expectations of stock market prosperity, which symbolizing the opportunity cost of Monetary Fund. It is expected that the scale of Yu'ebao is positively related to GC001 and negatively related to margin balance. The relationship between them can be shown in the history.

As for the relationship between monetary fund and interest rate, in the 1970s, a monetary fund was created. Monetary funds first originated in the United States. Before the marketization of US interest rates in the 1970s, the Federal Reserve imposed upper limits on savings deposits and time deposit rates through Q clauses, which resulted in short-term Treasury yields that were significantly higher than deposit rates. The widening spreads made many funds The manager saw the opportunity. So in 1971, the first money market fund in history was established. In the following years, money funds experienced rapid development. By 1982, the scale of assets had reached US $ 235 billion, surpassing stock and bond mutual funds for the first time.

As for the relationship between monetary fund and stock market, during the financial crisis of 2008, due to the stock market crash, market risk appetite declined, and a large amount of funds flowed into the monetary fund. At the end of 2008, the scale of money fund assets increased by 250% year-on-year. But after that, the extremely loose liquidity in 2009-2010 caused the decline in the yield of money funds, the rebound of the stock market, and the competition of bank wealth management products, which led to a contraction in the size of money funds.

Yu'ebao is a kind of money market fund. In 2013-2014, the Internet finance rises, and begins to enter the field of monetary funds. The establishment of Tianhong Zenglibao under the support of Yu'ebao in May 2013 marked that the development of currency funds has entered

a new stage. The shortage of money in June 2013 led to an increase in the yield of money funds, attracting a large amount of capital inflows and a rapid increase in scale. After 15 years, the expansion rate of money funds has gradually slowed down, but the scale is still rising. As of June 17, the net asset value of money funds reached 5.3 trillion yuan, accounting for more than 50% of the fund market size.

2. Significance of this research

The prediction of capital inflow and outflow is of great significance to the liquidity risk management of open-end funds

The research is not only of great significance to yu'ebao itself, but also plays an important role in the liquidity management of open-end funds.

Open-end fund is a kind of investment fund that allows fund share holders to purchase or redeem their own fund shares at any time. Its redemption mechanism not only gives its liquidity advantage, but also brings related risks.

When investors redeem fund units, especially large amount of centralized redemption, fund managers may suffer losses because they cannot realize their assets at a suitable price. In economics, this mismatch between assets and liabilities in liquidity is called liquidity risk. Liquidity risk is a great threat to the safe operation of open-end funds. Redemption risk has the greatest impact on open-end funds: first of all, large-scale redemption reduces the controllable funds of open-end funds. In order to cope with the redemption of investors, fund managers have to realize their assets, which will affect the return and net value of the fund. If fund managers increase the proportion of risk-free investment in response to the redemption of investors, this will increase the liquidity of fund assets, but it can not guarantee the return of the fund. Secondly, large-scale redemption will affect the reputation of fund companies. In the fund market of our country, the idea of investors is not mature. When the fund encounters large-scale redemption, investors often form bad expectation for the fund.

When the open-end fund is redeemed on a large scale, the fund manager will face the dilemma of a large number of selling assets to realize. This situation will not only lead to the further decline of the net value of the fund, affect the operation of the open-end fund, and even threaten its continued existence, but also affect the financial stability of the securities market and even the whole country. Therefore, the goal of open-end fund liquidity management is to maintain the appropriate liquidity of fund assets under the condition of ensuring a certain level of income and net value growth of open-end fund, which is sufficient to meet the redemption requirements under market conditions and reduce the liquidity risk of assets.

To solve the liquidity risk of open-end funds, in addition to strengthening fund supervision, improving laws and regulations, and improving the investment environment, the internal management of funds is also very important. The management methods mainly include:

(1). Optimize the allocation of fund assets. The principle of asset allocation is to divide and sort the assets according to the liquidity, and allocate the capital sources with different liquidity requirements to the assets with different liquidity. The sources of funds of open-end funds include: current savings of residents, idle funds of three types of enterprises, regular savings of residents, institutional funds such as securities companies, insurance funds and social security funds, etc. their demand for liquidity decreases from high to low. The

allocation of assets by the fund manager shall take full account of the proportion of various sources of the fund. Most of the funds from the current savings of residents should be invested in the assets with strong liquidity, such as national bonds. For the long-term funds, such as life insurance funds and social security funds, appropriate amount can be used for the long-term held heavy stocks.

(2). Make reasonable use of liabilities to supplement liquidity. Generally speaking, it is feasible for the fund to solve the short-term shortage of funds by using liabilities, but at present, the financing channels are very narrow. The main ways of using short-term financing include: bank short-term loans, bond repo, national debt repo, etc. when the fund has a huge redemption, the fund manager can actively raise funds from the outside to make up for the lack of liquidity. If fund managers can not alleviate the pressure of fund demand through external financing, internal financing is also a useful means. Internal financing refers to the financing within the fund industry and between different fund companies, mainly including fund lending and exchange.

(3). Effective prediction of liquidity demand. The fund manager shall analyze the holders of the fund units, classify them according to their sources of funds, holding motives, sensitivity to the securities market, sensitivity to the interest rate and other factors, and make a comprehensive analysis with the overall trend of the stock market, fund performance, interest rate, investor preference and other factors. Then, it can realize the prediction of liquidity demand, so as to achieve the management of fund liquidity.

It can be seen that for open-end funds, accurate prediction of capital inflow and outflow plays an important role in reducing capital liquidity risk and satisfying daily business operation.

3. Data Description

(1) Microeconomics Data

1) user_profile_table

We randomly chose 30,000 users, some of who first appears in September, 2014. These users are only contained in test data. Thus, the user_profile_table only contains the basic information of 28,000 users.

Name	Type	Meaning	Examples
user_id	bigint	User ID	1234
Sex	bigint	Gender (1: Male, 0: Female)	0
City	bigint	City location	6081949
constellation	string	Constellation	Sagittarius

2) user_balance_table

There are complete purchase and redeem data from July.1st, 2013 to Aug. 31st, 2014, including all subsets information. All data were masking. The unit of amount is 0.01

Name	Type	Meaning	Examples
user_id	bigint	User ID	1234
report_date	string	Date	20140407
tBalance	bigint	Today's balance	109004
yBalance	bigint	Yesterday's balance	97389

total_purchase_amt	bigint	Total purchase amount = Direct Purchase + Earnings	21876
direct_purchase_amt	bigint	Today's direct purchase amount	21863
purchase_bal_amt	bigint	Today's purchase balance amount	0
purchase_bank_amt	bigint	Today's purchase amount from bank card	21863
total_redeem_amt	bigint	Total redeem amount = consume + transfer	10261
consume_amt	bigint	Today's consume amount	0
transfer_amt	bigint	Today's transfer amount	10261
tftobal_amt	bigint	Today's amount transferd to Yuebao's balance	0
tftocard_amt	bigint	Today's amount transferd to bank card	10261
share_amt	bigint	Earnings amount	13
category1	bigint	Total consumption of category 1	0
category2	bigint	Total consumption of category 2	0
category3	bigint	Total consumption of category 3	0
category4	bigint	Total consumption of category 4	0

3) mfd_day_share_interest

Name	Type	Meaning	Examples
mfd_date	string	Data	20140102
mfd_daily_yield	double	Earnings of ten thousand shares	1.5787
mfd_7daily_yield	double	Annualized rate of return of seven-day （%）	6.307

(2) Macroeconomics Data

Besides, we will consider some macroeconomics elements including Shibor rate, Amount of margin trading and Amount of short selling.

4. Data Processing

There is one type of users who operate in a low frequency, so after once or twice operation, they will not operate anymore and just wait for earnings. It's very hard to predict the operation habit of this type of users in the next month. Another type of users redeem money from Yuebao very randomly. It's difficult to summarize the rules for modeling. When we add up all of these 100,000 users, we can see something that's a little bit more regular. Firstly, it shows a more obvious periodicity based on 7 days as a weekly fluctuation. Secondly, the

amount during working days will be relatively high, while the amount during holidays will be relatively low. By the beginning of 2014, the amount grows rapidly, so the volatility is relatively large. Since March in 2014, there is a stable stage until the end of August.

In conclusion, we will divide the users as four parts according to the quantity of purchase and redemption. The maximum single redemption is used to distinguish high-net-worth users from low-quality users. When we classify data, we not only need to ensure that the number of high-net-worth users should be as much as possible, but also require that the trend between high-net-worth users and low-quality users should be as similar as possible.

5. Modeling

(1) Trend decomposition

We think this is a time-series problem. The trend of total amount of purchase and redeem can be decomposed as long term, midterm and short term to predict separately.

We are going to use singular spectrum analysis to do the trend decomposition.

The core idea of SSA is to derive a series of singular values which includes the information of original series and then to construct different characteristic time series by choosing different singular values. The concrete steps are as follow:

Step 1. Reconstructing the phase space of time series: Given a time series as $X_N = \{x_1, x_2, ...x_N\}$, where N is the effective length. By constructing the phase space of series, we can derive the trajectory matrix：

$$D_m = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ x_2 & x_3 & \cdots & x_{n+1} \\ \cdots & \cdots & \cdots & \cdots \\ x_m & x_{m+1} & \cdots & x_{n+m-1} \end{bmatrix}.$$

Here n is the length of the window, m is the embedding dimension, and $2 \le m \le N/2$, $m \le n$, $N = n + m - 1$.

Step 2. Decomposing the singular values: The SSA is based on a particular transformation known in matrix algebra as singular value decomposition (SVD). Take D_m for its singular value decomposition, $D_m = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T$, $d = \max(i, \lambda_i > 0)$, where $\lambda_1 \ge ... \ge \lambda_m \ge 0$. The collection of $(\sqrt{\lambda_i}, U_i, V_i)$ is called the i^{th} eigentriple of the SVD. Vectors U_i are the left singular vectors of the matrix D_m, numbers $\sqrt{\lambda_i}$ are the singular values and provide the singular spectrum of D_m ; this gives the name to SSA. Vectors $\sqrt{\lambda_i} V_i = D_m{}^T U_i$ are called vectors of principal components (PCs). Let $D_m = USV^T$, if the time series only contains effective information, then the rank of the matrix S is $k < m$； If the time series contains both effective information and noise, then $k = m$.

Step 3. Determining the order of noise reduction based on singular entropy: Introducing the singular entropy [28] is to study the law of the amount of information changing with the order of singular entropy :

$$E_k = \sum_{i}^{k} \Delta E_i (k < d).$$

Here, k is the order of the singular entropy, and ΔE_i represents the increment of the singular entropy in order i. By the following formula we can obtain that:

$$\Delta E_i = -\left(\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \right) \times \log \left(\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \right).$$

When the increment of the singular entropy approaches to the asymptotic value, the effective information of the time series saturates, which means the information is fully included. Afterwards, the increments of the singular entropy are caused by the noise. Therefore, at this point, the number of the i^{th} order can be selected as the one for noise reduction.

Step 4. Reconstructing the time series after noise reduction: After noise reduction, the reconstruction of the time series can be divided as the two processes: first, we can determine the number i as the order of nose reduction from the above analysis. Let all the singular values whose orders are greater than i be 0 to derive a new singular value matrix S', and then get a free noise trajectory matrix with the formula $D_m' = US'V^T$. Secondly, let the reconstructed time series as $G = \{g_0, g_1, g_2, ...g_{N-1}\}$, and $x_{c,b-c+2}$ be one of components of D_m', we will have the following formula (Actually, the formula represents the mean of the sum total of the antidiagonal components of the matrix):

$$g_b = \begin{cases} \frac{1}{b+1} \sum_{c=1}^{b} x_{c,b-c+2} & 0 \le b \le m-1 \\ \frac{1}{m} \sum_{c=1}^{m} x_{c,b-c+2} & m \le b < n \\ \frac{1}{N-b} \sum_{c=N-n+2}^{N-n+1} x_{c,b-c+2} & n \le b < N \end{cases}.$$

According to the above formula, we can obtain a time series containing the effective information after noise reduction.

(2) Feature construction

After decomposing the long-term, middle-term and short-term trend term, we try to construct some factor to capture the trend of each team. As we see, the purchase and redeem amount fluctuates in weekly period and monthly period, it's sensible to use dummy variables while modelling the trend term.

Date based dummy features

1) whether the day is a traditional festival
2) whether the day is weekend
3) whether the day is last day of festival
4) whether the day is the first day for working after the festival
5) whether the day is weekend
6) whether the day is weekday
7) whether the last day is weekday
8) whether the next day is weekday

9) whether the day is the last day of the month
10) whether the day is the middle day of the month
11) whether the day is the first day of the month
12) whether the day is the first week of the month
13) whether the day is the last week of the month

As a monitory fund, the amount of it will be correlated with the macro environment tightly. So, we need to consider the effect of policy and other similar events.

14) whether the date is an IPO day
15) the rest amount of financing funds and bonds
16) rate of national debt repurchase

We also consider the statistic feature in the rolling window of a week and a month.
17) the rolling window median
18) the rolling window max
19) the rolling window min
20) the rolling window std
21) the rolling window skew

(3) Feature selection

As the feature we constructed is more than the best model need, if we regress the model on all the features, overfitting will happen.

1) If the data set of one feature is too small, the prediction will not trustable, we will delete this feature.
2) If the correlation between the feature and explained variable is too small, it's no use to add it in the model.
3) If the correlation between features are high, we will delete the feature with lower variance.
4) If the feature can't separate the data will, any side of the feature is too small, we will delete the feature.
5) Using MV test to retrieve some dependent but unrelated features
6) SHAP test represent the fair score of features depending on their contribution towards the total score in the set of features. SHAP also can visualize how the score changes when the feature test is low/high on each data.
7) The PCA and ICA methods can also minimize the feature and subtract the main feature, we may try this method in our further research.

(4) Prediction model

The loss functions we used in model are not always absolute distance function. But, the competition uses the weighted absolute distance function, it's a tip to use the same function as the competition. However, the absolute function is not differentiable, we my use the $\boldsymbol{\Sigma}\frac{(\mathbf{y_i}-\hat{\boldsymbol{y}})^{\mathbf{2}}}{\boldsymbol{y_i^2}}$ as an alternative method. And, we need to evaluate the model in different time window and different criterions. As a time-series problem, rolling window test is necessary.

Then, we will try different regression model, such as OLS, SVR, random forests, to find the best explaining effect according the above criterions. Moreover, we may combine the different model to construct a better one, but the limitation of data will be a restriction.

6. Main challenge

(1) Sample selection

For this research, we could only obtain small sample dataset on the platform, which is provided by Alibaba. Because of that, we may face with overfitting problem. Beside the situation of capital inflow and outflow shows great difference among different individuals, in order to get an accurate model, we need to be vary carful with processing sample outliers.

Secondly, considering difference among all different days, we can't use all sample set to do the model construction. For example, the capital data always shows abnormal volatility at the end of year or the beginning of new year, which means that for prediction of September, we could only use data from March to August.

Moreover, when the sample difference is big, it is also difficult to divide train and test dataset. Because there exists unique feature for each dataset of different time nodes, we need to be cautious about selecting base for dividing dataset. In general, how to optimally select sample to train our data is the first challenge.

(2) Feature selection

The design of feature engineering is also a challenge. Due to the special composition of this dataset, there are a lot of noise in the data and these noises may be mistakenly assumed as a feature, which increase the difficulty of feature extraction. For example, if we observe this year data, we can find that the capital outflow before specific festival is always greater than outflow after the festival, but this rule is not satisfied with the other year, so we can't include this feature into the model.

Beside in this process, we need to guarantee the following conditions. Firstly, the redundancy rate of selected features is as small as possible. Secondly, the number of dimensions of features should not be too big. Thirdly, each feature should be positively correlated with dependent variables. So, it means that we need to do test to recheck each feature's efficiency.

The process of feature selection also involves time series analysis and decomposition, for the decomposition procedure, it is highly related to the variables we want to include, in our model we decide to include some macroeconomic variables, which needs special treatment in decomposition procedure. The whole process requires us to deeply understand decomposition model and how to evaluate the performance of specific model.

(3) Model construction and adjustment

The first step of model construction is to select appropriate model for solving the problem. Different model would show different performance for same problem. For this research, there are several models in our consideration, such as linear regression, neural network, SVM (Support vector machine), random forest.

For linear regression, it would be easily influenced by outliers, as there are a lot of noises in our dataset it may not be the best choice. As our sample is not big enough, the neural network may also lack of efficiency. For SVM it depends on the selection of kernel function and there is a big issue with parameter adjustment. In previous researcher, some teams choose to use random forest and it shows great efficiency, but our framework is different from them and we need to figure out our own model and prove its efficiency, which is also a challenge.

This process requires us to clearly understand the basic conditions of using each model and how different parameters affect the performance of model. We need to try our best to find appropriate parameters to improve the robustness of model.

(4) Other issues

As this research involves short and long terms feature, we need to fully understand what is the rule behind each feature. Although it is impossible to accurately predict personal capital flow and it is not our goal, there exists basic logical structure to explain the features, if we can't fully understand the framework of capital flow, it would be hard for us to create more efficient model.