Final Report: Inflow and outflow forecasts of Yuebao Monetary Fund

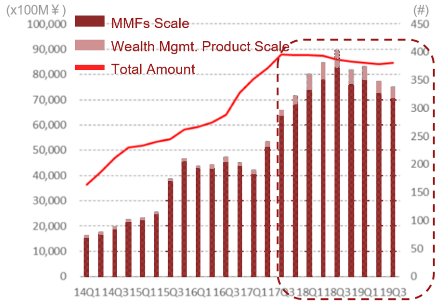
Lai Lin 1801212867, Lian Di 1801212881, Liu Sheng 1801212891, Alimujiang 1801212778, He SongTao 1801212852

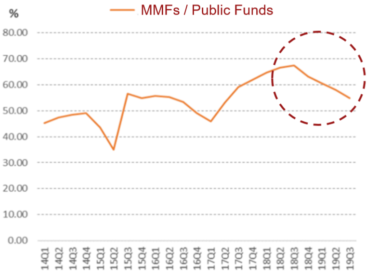
1. Introduction to Yu’e Bao

Yu’e Bao, launched by Ant Financial Services Group in 2013, is a kind of monetary market fund (MMF) invested in short-term securities. the underlying of MMFs are basically safe and liquid securities, including government bonds, central bank bills and high rating corporate bonds. The average maturity of them are about 120 days. Because of these features, MMF has many advantages, including high security, high liquidity and stable returns, which make it very like the bank deposit, but with higher returns. Yu’e bao is a MMF platform providing online MMF purchasing and redemption service. At the end of first half of 2019, Yu’e Bao’s user reaches 600 million, which accounts for half of China’s population.

2. Significance of fund inflow and outflow of Yu’e Bao

No matter the scale or the percentage of MMF continue to shrink in China, but the mechanism why Yu’e Bao fund scale change is still unknown. The scale can be changed in two directions, inflow and outflow. For the perspective of fund managers, he needs to accurately predict the inflow and outflow. The reasons are twofold. Firstly, funds manager needs to accrue appropriate risk reserve. Provided that market yields rise, the bond prices will fall. If investors want to redeem their MMFs but the fund doesn’t have enough reserve, the fund has to sell bonds before maturity at a very low price, which will decrease the net worth of the fund. The other investors see such situation will also redeem their funds, which will enter a vicious circle. Also, according to CSRC’s regulation, the total net worth of MMF at month-end must not exceed 200 times the monthly balance of the risk reserve. So fund managers should set appropriate risk reserves based on the inflow and outflow to meet regulatory requirements. Secondly, from the right graph we can see that the maturity of underlying asset in MMF is not fixed, and the proportion of long-term asset becomes larger. Normally, Fund managers should allocate assets of appropriate duration according to the inflow and outflow to reduce the liquidity risk.



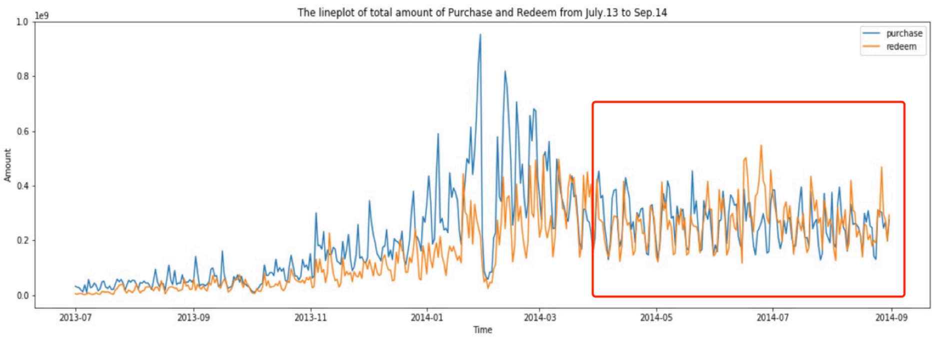


3. Research Question

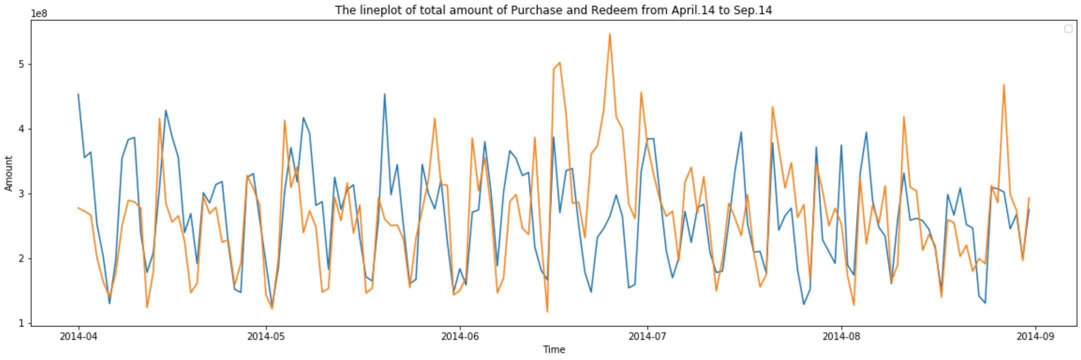
We want to build a model to forecast the future fund inflows and outflows of Yu’e Bao. This can help fund managers manage positions of funds and make appropriate risk reserve. We use the purchase and redemption data in 2014. The reasons why we don’t use later data are that we would like to avoid exogenous influence of state regulation in 2017. Intuitively, both macro and micro features might have impact on the inflow and outflow. One of the macro feature is SHIBOR, we can see from the graph that SHIBOR and Yu’e Bao’s rate of return are correlated. And the rate of return is likely to induce investors to buy the fund. The other macro feature is the stock market expected return, which is the opportunity cost of investors. Later we will see that this feature will be proxied by balance of margin trading and securities lending. Micro features are mainly user liquidity preference.

4. Feature Engineering

A time series chart of the total daily purchases and redemptions every day from July 13 in 2013 to September 14 in 2014 was drawn after the sum of 100,000 users. The results are shown in the figure below.



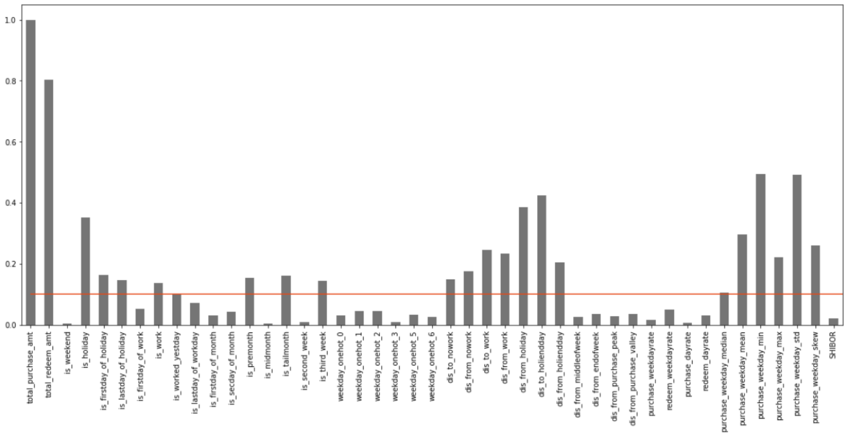
By observing the figure above, we can find that the overall data have significant characteristics: ①Obvious periodicity on a seven-day basis. ②The amount of working days is relatively high, while the amount of holidays is relatively low. ③The overall trend is relatively low from July 2013 to Nov. 2013, relatively large fluctuations from Nov. 2013 to Apr. 2014, and a stable period after Apr. 2014, which lasts until the end of Aug.2014. To model the data for September, we need regular user behaviors requiring data to be relatively smooth. In addition, we need to select samples which can describe the characteristics of September.



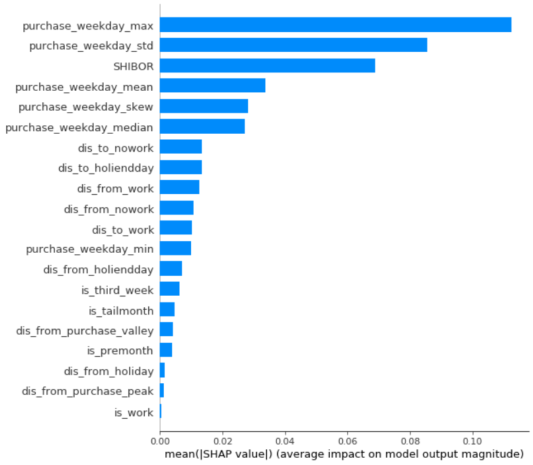
Combining the above cyclical and holiday characteristics, we can construct a static feature of the date. For time series, we will construct "is" feature based on date static feature and "distance" feature separately. As for macroeconomic factors, since monetary funds are main participator on the inter-bank market, so we will consider Shibor and Balance of Margin Trading and Securities Lending as features.

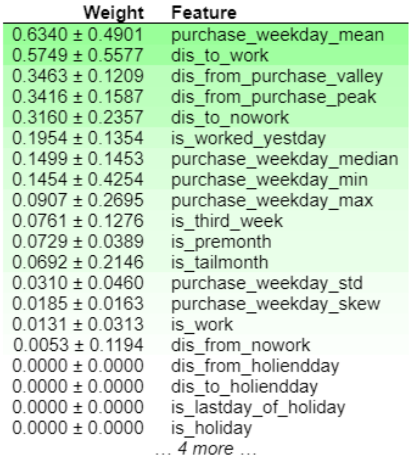
5. Feature Selection

We divided feature selection process into 5 steps: Low separation ability, Multi-collinearity, Low correlation, Shapley value and Permutation Importance. Feature selection is the process of selecting a subset of relevant, useful features to use in building an analytical model. Firstly, we delete features with low separation ability. Secondly, we delete features with multi-collinearity. Rules: Compute correlation coefficient between any two features. If two feature's correlation coefficient is larger than 0.8, then delete the one with small correlation with predicting variables. Use VIF statistic value to detect potential multi-collinearity, the truncated value is set to be 10. Thirdly, we delete features with low correlation. We set the threshold as 0.1.



Fourthly, we select the features with high Shapley value. Fifthly, we select the features with high permutation importance.





Weight	Feature
0.6340 ± 0.4901	purchase_weekday_mean
0.5749 ± 0.5577	dis_to_work
0.3463 ± 0.1209	dis_from_purchase_valley
0.3416 ± 0.1587	dis_from_purchase_peak
0.3160 ± 0.2357	dis_to_nowork
0.1954 ± 0.1354	is_worked_yestday
0.1499 ± 0.1453	purchase_weekday_median
0.1454 ± 0.4254	purchase_weekday_min
0.0907 ± 0.2695	purchase_weekday_max
0.0761 ± 0.1276	is_third_week
0.0729 ± 0.0389	is_premonth
0.0692 ± 0.2146	is_tailmonth
0.0310 ± 0.0460	purchase_weekday_std
0.0185 ± 0.0163	purchase_weekday_skew
0.0131 ± 0.0313	is_work
0.0053 ± 0.1194	dis_from_nowork
0.0000 ± 0.0000	dis_from_holiendday
0.0000 ± 0.0000	dis_to_holiendday
0.0000 ± 0.0000	is_lastday_of_holiday
0.0000 ± 0.0000	is_holiday
… *4 more* …	

Sixthly, we get the intersection of these two selection method.

6. Model Analysis

6.1 Outline

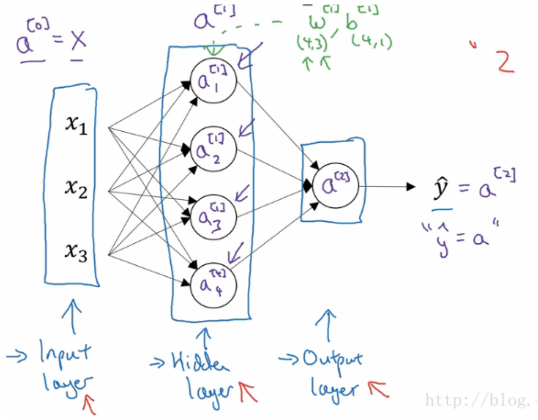
In our project we use six models, including tradition Linear regression, Decision tree regression, Random Forest regression, gradient boosting, MLP regression and XGBoost regression. For most of the selected models, we use the python package sklearn, which is a powerful tool for machine learning. These models could be divided into linear model, decision tree, ensemble method, neural network.

6.2 Multi-Layer Perception

For the neural network model, we use Multi-layer perception model here, which is called as MLP. MLP is a supervised learning algorithm that learns a function by training on dataset. Because it can learn a function for either classification or regression, we think that it is suitable for our purpose. For this model, we need to consider number of hidden neurons and number of delays. we can infer the number of hidden neurons by the empirical formula.

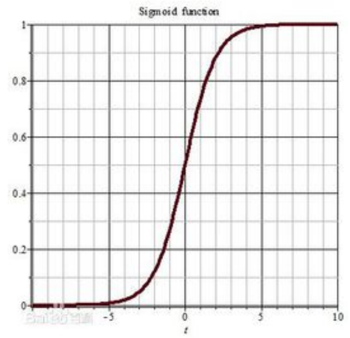
$$k < \sum_{i}^{n} C\binom{nl}{i}$$

Workflow of MLP method



Acitivation function

$$
\begin{aligned}
sigmoid'(x) &= (\frac{1}{1+e^{-x}})' \\
&= \frac{1}{1+e^{-x}}e^{-x}(-1) \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}}(1-\frac{1}{1+e^{-x}}) \\
&= sigmoid(x)(1-sigmoid(x))
\end{aligned}
$$



The result shows the number could be unit digitals or more than 10. For the number of delays, we can find that the week is one of the most significant features, so we can define the number of delays as 7.

6.3 Gradient Boosted Decision Tree

The second one is gradient boosted decision trees. GBDT is a generalization of boosting to arbitrary differentiable loss functions. It is also based on decision tree, the key difference here is that this method build several instances of a black-box decision tree and reduce the variance of the decision tree. As one kind of ensemble methods. This method combines the advantage of both linear model and decision tree, so we think that this could be one powerful model for our case. One of the most important parts related to the performance of GBDT is to add features efficiently.

6.4 Comparison between MLP and GBDT

From theoretical perspective, the key differences between these two methods are shown as below:

Model	MLP	GBDT
Advantage	• Capability to learn non-linear models • Capability to learn models in real-time	• Processing different kinds of dataset flexibly.
Disadvantage	• Non-convex loss function, different random weight initialization can lead to different validation. • Requiring a number of hyperparameters and is sensitive to feature scaling.	• Potential interdependence within weak learners, it is hard for training data in a paralleled way.

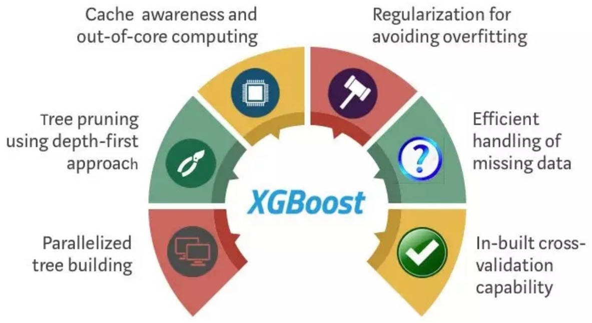
6.5 Random Forest

we also use random forest, it likes GBDT from many aspects, this method is more robust to outliers and the result doesn’t need to rely on parameters adjustment and we don’t need to do special treatments toward features by this model, but without adjusting weight of each learners this method may lose some accuracy.

6.6 XGBoost

XGBoost implements machine learning algorithms under the Gradient boosting framework. From bias-variance tradeoff perspective, this method could efficiently control the complexity of model by regularization. Beside by column subsampling, XGBoost highly save computing resource. Before doing actual test, we predict that its predicting power would be on the top.

Advantages of XGBoost regression method


Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

7. Training, evaluation and prediction

7.1 Method selection

There are three ways to solve this problem. The first is to model a single user and summarize after forecasting. The second is to classify the user groups and summarize the forecasts by groups. The third is to forecast according to the daily total amount.As the first and the second have the error accumulation problems, the third way is to solve the problem better.

7.2 Division of training set and test set

It is not good to choose August as the test set, because there is no holiday in August compared with September, and the score in August cannot reflect the score in September. Therefore, it is a better choice to choose April as the test set.

7.3 Self constructed scoring model

We built a score function by ourselves.

$$\sum Max(0, \frac{e^{1-x/0.3}-1}{e-1})$$

7.4 Prediction

It can be found that XGB has the best effect, XGB is not the best in the test set of April. The reason is that XGB has stronger generalization ability, and can avoid overfitting problems, as described in our analysis of the model above.

	Test Set	**Online Score**
Linear Regression	160.2887	108.7332
Decision Tree	179.1604	87.4493
Random Forest	192.7704	104.8427
Gradient Boosting	202.6532	115.5463
MLP	152.5565	109.3498
XGB	200.0075	119.9311

8. Result analysis and interpretation

8.1 Differences between results and expectations

We found three differences from our guess:

1. The situation of stock market has little influence on yu'ebao;
2. Users of yu'ebao are not sensitive to the change of interest rate;
3. "Whether it's a holiday", "whether it's a weekend" are more significant Features.

8.2 Result interpretation 1: motivation of early users of yu'ebao

The early users of the Yuebao mainly came from Alipay and Taobao users. Yuebao played a role of investment education for most users. Therefore, this part of the users mainly used Yuebao as a cash management tool. Therefore, Yu'ebao's early substitutes were not other financial instruments, but mainly bank deposits and cash. Compared with current deposit and fixed deposit, Yu'ebao has higher yield and more convenient access. Therefore, the purchase and redemption volume of Yu'ebao in 2014 was less affected by other financial products. After 2015, with the increasing number of users of Yu'ebao and more and more people taking yu'ebao as a choice of financial management, yu'ebao's substitution effect by other financial management products began to be obvious gradually.

8.3 Result interpretation 2: Destination of the redemption

Because of the relevance between Yuebao and Alipay and Taobao platform, users can choose to use the amount of Yuebao directly to make payment when they use Alipay payment. According to the 2014 big data report of Yu'ebao, 75% of the outflow of yu'e Bao's funds is used for consumption, and only 25% is used for withdrawal to bank cards. This explains why features such as "whether it's a holiday", "whether it's a weekend", "whether it's a working day" that have a great impact on consumption also have a great impact on redemption. At the same time, it also explains the reason why the stock market has little impact on yu'ebao in 2014.

8.4 Summary, prospect and improvement

Since both the training data and the test data are 2014 data, the conclusion drawn is quite different from the initial guess. On the one hand, we believe that the influencing factors of Yu'ebao purchase and redemption data in the future will change greatly with the change of user composition. For the research of the problem, it is worth doing more exploration in the future (if data available). On the other hand, after finding the data results and expectations are quite different, we actively look for the reasons from three aspects: data, model and fact, and finally find out the reasons successfully. We also feel that data analysis needs to be combined with the actual basic situation, and it is meaningless to do data analysis away from the fundamental information.

References

[1] Belsley, D. A. (1991). Conditioning diagnostics: Collinearity and weak data in regression (No. 519.536 B452). New York: Wiley.

[2] Chatterjee, S., & Price, B. (1977). Selection of variables in a regression equation. Regression Analysis by Example, 201-203.

[3] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpret- ing Model Predictions. In Advances in Neural Information Processing Sys- tems 30. Curran Associates, Inc., 4768–4777. http://papers.nips.cc/paper/ 7062- a- unified-approach- to- interpreting- model- predictions.pdf

[4] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang New- man, Jerry Kim, et al. 2017. Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. bioRxiv (2017), 206540.

[5] Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

[6] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

[7] Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), 1340-1347.

[8] Achard,S. et al. (2005) Criteria based on mutual information minimization for blind source separation in post nonlinear mixtures. Signal Processing, 85, 965–974.

[9] Strobl,C. et al. (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8, 25.

[10] Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. J. Mach. Learn. Res., 3, 1157–1182.

[11] Breiman,L. et al. (1984) Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

[12] Breiman,L. (2001) Random Forests. Mach. Learn., 45, 5–32.

[13] Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. J. Mach. Learn. Res., 3, 1157–1182.

[14] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.

[15] Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1-4.

[16] Quinlan, J. R. (1987). Simplifying decision trees. International journal of man-machine studies, 27(3), 221-234.

[17] Longstaff, I. D., & Cross, J. F. (1987). A pattern recognition approach to understanding the multi-layer perception. Pattern Recognition Letters, 5(5), 315-319.

[18] Prettenhofer, P., & Louppe, G. (2014). Gradient boosted regression trees in scikit-learn.

[19] Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E., & Suter, B. W. (1990). The multilayer perceptron as an approximation to a Bayes optimal discriminant function. IEEE Transactions on Neural Networks, 1(4), 296-298.

[20] Pal, S. K., & Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. IEEE Transactions on neural networks, 3(5), 683-697.

[21] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.