2018 IEEE International Conference on Big Data (Big Data)

Subscription and Redemption Prediction in Mutual Funds Using Machine Learning Techniques

Morteza Mashayekhi*, Iman Rezaeian†, Annie Z. Zhang ‡ and Jonathan Anders§
Royal Bank of Canada (RBC), Toronto, Canada
Email: {*morteza.mashayekhi, †iman.rezaeian, ‡annie.z.zhang, §jonathan.anders}@rbc.com

***Abstract*—There are various factors that derive mutual funds' flow. Among them, factors related to investment patterns and behaviors of the investors are highly informative. It is very beneficial for a fund manager to know who are the most probable investors that are going to subscribe to or redeem from a particular fund in near future. In addition, extracting the important underlying factors involved in this process helps fund managers to plan for optimizing their fund's performance. Our experiments on historic transaction data of about 400 mutual funds show that we can extract most informative patterns and use them to predict mutual funds' flow with relatively high accuracy. In addition, the proposed investors' ranking method gives a curated list for running more effective targeted campaign.**

I. Introduction

A fund is a supply of capital that is belonged to different investors and can be used to purchase various securities while each investor retains ownership and control of his/her own shares. A fund provides better investment opportunities and management while at the same time lowers the investment fees than investors might be able to obtain on their own. There are different types of funds including mutual funds, exchange-traded funds (ETFs), money market funds and hedge funds among others.

Mutual fund, in particular, is a registered open-end investment that invests the investor's money into stocks, bonds, short-term money-market instruments, or other types of securities or assets, or as a combination of the aforementioned investments. Over the past two decades, mutual funds have become the main type of investment for small investors across the globe. In fact, in past few years, the number of mutual funds in United States have exceeded the number of securities [15]. In comparison with other types of investments such as direct investments in individual stocks and bonds, mutual funds offer better liquidity and diversification at a lower cost.

There are different types of transaction for a given mutual fund such as transfer and switch. But the most important transactions in terms of their influence on fund's performance are subscription (contribution or buy) and redemption (sell). According to Chen et al. [3], there are a few main factors that are usually considered when the fund's investors want to contribute to or redeem from a fund: 1) whether the funds investors are mainly individuals and small stakeholders, or banks and large institutions, 2) the past performance of the fund; and 3) the fund's liquidity.

One of the most informative factors that a given fund's investment manager hope to know is that who are the most probable individuals that are going to subscribe to or redeem from a particular fund in near future. That way, an investment manager not only can plan ahead in order to optimize the fund's performance based on the predicted cash flow of the fund, but also can engage with the top potential subscribers and redeemers and persuade them in such a way that maximize the fund performance.

The common approach for engaging with the investors is to get in touch with them using various outlets (email, call, mail, etc.) or in-person to realize if an investor is going to subscribe or redeem based on their feedback. However, this approach is not very efficient most of the time. First of all, the cost of engagement is high as it tries to approach investors rather blindly. Second, frequent engagement with clients that are not considering any subscription/redemption in near future may cause negative impact on their relation with the fund manager and/or the fund itself.

Machine learning techniques can facilitate these engagement in a more target oriented fashion. Targeting only those investors that are high likely to subscribe to/redeem from the fund in near future, not only decreases the engagement cost, but also increases the probability of persuading the right investors to subscribe to a new fund or stay with the current fund by providing more personalized insights and offers. It also can provide the fund manager a broader and more comprehensive view of the fund behaviors such as seasonality, trend, and predicting the fund activity in near future. Also since not all investors act the same way across different market conditions [1], [8] ranking significant factors based on their amount of influence in decision making process of investors is a very important insight toward tailoring and customization of the process for each investor more efficiently.

It is known that there is no single factor that derives the mutual funds flow. *Investor Economics*, an independent research and consulting firm specializing in the financial services sector, found that mutual fund flows can be influenced by a complex interplay of as much as 40 different factors [7] from investors IQ [9] to their conflict of interest [6]. However, it has been shown that there are three main factors to advancing the understanding of the volumes and the directionality of mutual fund sales and redemptions [7]: Macro-economic and demographic factors, individual fund investment return characteristics, and finally preferred access to distribution.

Wang et al. [17] tried to identify the relationship between purchase and redemption behavior of flow-return and flow-

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: ASTON UNIVERSITY. Downloaded on August 20,2025 at 19:37:54 UTC from IEEE Xplore. Restrictions apply.

fund characteristics within different group investors by using quantile regression and comparing return performance of insured vs non-insured investors.

Chen et al. [5] categorized users into five different groups: small users, large inactive users, large more active users, large active users, large ultra-active users based on their operating frequency. Then an ARIMA model was used to model each of the five classes independently.

Qamar et al. [14] used a non-parametric method to analyze the efficiency and performance of mutual funds. Using Data Envelopment Analysis (DEA), they predicted the performance of fund in coming years. Different factors such as mutual fund returns, turnover rate, volatility, and expense ratio were used to find the relative efficiency of funds using DEA.

In this work, our aim is to capture valuable insights and patterns from historical redemption/subscription (red/sub) transactions performed by mutual funds' investors. More specifically, we want to investigate the possibility of accurately predicting the next month's red/sub for a given investor. In addition, we want to derive the most important factors involved in the investors behaviors. Finally, we aim to rank the investors based on the probability of red/sub and other factors in order to provide a more tailored and targeted campaign.

II. Method

In this section, we will describe the applied methods for data processing, prediction models construction, evaluation mechanism, investor ranking, and rule extraction from prediction models.

A. Data preparation and feature engineering

We used two years (2016-2017) worth of the transaction data that represents the investment patterns of more than 800,000 active investors within around 400 mutual funds across different fund classes. There are two broad types of investors: individuals and corporations. We only consider the transaction records that represent funds share's redemption and subscription. For each investor, monthly net amount of redemption/subscription (Net) is computed. The reason for computing Net amount is to capture switch transactions, where an investor redeems from one fund and subscribes into another fund. Here, the positive amounts show subscription and negative amounts show redemption. Afterward, we generate a feature set consisting of two parts. The first part includes high-level information about the investor where applicable such as age, province, investor type, and type of accounts. The second part is extracted from historical transaction data as a representation of investor's investment behavior pattern using a sliding window method. Given a window of length k, at each point in time we compute average($mean$), minimum(min), maximum(max), median, standard deviation(std) of Net amounts as well as the number of subscriptions ($subscr\_count$) and redemptions ($redmpt\_count$), in past k months. Month features($M1$ to $M12$), which shows the red/sub prediction month, is also added to the feature set to capture likely seasonality in the data. In addition, we include the average ($mean\_incpt$), minimum (min_incpt), maximum (max_incpt), median, and standard deviation (std_incpt) of Net amounts as well as number of subscription ($subscr\_count\_incpt$) and number of redemption ($redmpt\_count\_incpt$) from the beginning of the transactions which shows these value since inception of the fund. These features represent a simple embedding of the investors' investment pattern. As an example and considering 24 months of the available data, if $k = 6$ then we will generate 18 records per investors including all aforementioned features. Out of these features, month, province, and investor types were encoded to a one-hot encoding representation, results in generating 59 features per sample.

Finally, the label is determined by the next month Net amount right after the window with size k. If the Net is negative it shows redemption ($class1$) unless it is non-redemption ($class0$) when we want to predict redemption. Similarly, if the Net is positive it shows subscription ($class1$) unless it is non-subscription ($class0$) when we want to predict subscription.

B. Redemption/subscription prediction and evaluation

In order to predict redemption/subscription, the data was split into two sets: training set and validation set. To do so, depending on the month of prediction, all of the records corresponding to that specific month are considered as our validation set and the rest is used as training data. Since most of the mutual funds' investors invest rather passively, only 10% of investors either subscribe to or redeem from a fund on average. To resolve this issue, we employ an under-sampling technique and reduce the number of instances corresponding to $class0$ in such a way that results in a balanced distribution of samples for both classes. The combination of under-sampling along with ensemble learning can help to achieve better results for classification problems [18]. Moreover, we employ logistic regression [11] as our baseline classifier and compare it against two ensemble methods i.e. random forest [2] and XGBoost [4] models using *scikit-learn* package [12]. For logistic regression model, we standardize all the continuous features.

For random forest we use 500 trees. To tune the XGBoost, we try to minimize AUC for prediction of a specific month (June 2017) by modifying hyper parameters such as learning rate, number of estimators, maximum depth of estimators, regularization (reg), minimum sum of weights of all observations required in a child (min_child_weight), fraction of observations to be randomly samples for each tree ($subsample$), and fraction of columns to be randomly samples for each tree ($colsample\_bytree$). The hyper-parameters used in training our XGBoost model is listed in table I. We leave default values for the rest of hyper-parameters in both models.

We apply a cross-validation scheme to evaluate and compare the models' performance. For every month of the last six months of 2017, we build training and validation sets as explained in section II-A and finally compute the average and standard deviation of various evaluation metrics.

Authorized licensed use limited to: ASTON UNIVERSITY. Downloaded on August 20,2025 at 19:37:54 UTC from IEEE Xplore. Restrictions apply.

TABLE I
HYPER PARAMETERS USED IN TRAINING THE PROPOSED XGBOOST MODEL

Hyper-parameter	Value
learning rate	0.035
n_estimators	57
max_depth	10
min_child_weight	7
subsample	0.53
colsample_bytree	0.9
objective	binary:logistic
Reg	0.0025

C. Investor ranking

In order to rank investors based on their importance, We select a combination of features including prediction probability to rank the investors based upon them. For redemption, we rank the investors based on the *prediction probability*, *min* (which shows the maximum redemption amount), *avg*, and $min \times predictionprobability$ (shows likely redemption amount). Similarly for subscription, we use prediction probability, *max* (likely subscription amount), *avg*, and $max \times predictionprobability$ (likely subscription amount). We employ two different methods to find the final rank for all the investors: average ranking and optimal ranking. The first one is simple average ranking of the selected features and when there is a tie the average value is considered. The second one, $RankAggreg$, [13] is a ranking aggregation method, which tries to find an aggregated rank via the cross-entropy (CE) Monte Carlo algorithm. The algorithm searches for the super-list which is as close as possible to the ordered lists based on the data. For each list, we use min-max normalization values of the same features. The *Spearman footrule* distance is used to measure the closeness of any two ordered lists. Finally to evaluate applied ranking methods, we compute the percentage of the red/sub amount covered with top investors in comparison with the whole amount of red/sub.

D. Rule extraction

Because of the black-box nature of XGBoost and random forest, explaining their prediction outputs is not straightforward. Hence, we use a rule extraction method to find the main reasons of red/sub. We employ RF+MSGL [10] to extract rules from random forest model built on the data. In RF+MSGL model, the rule extraction problem is converted to a regression problem and solved using the sparse group lasso method [16], where the extracted rules from RF are considered as features and each row indicates whether the sample instance matches with each rule (feature).

III. RESULTS

For generating the training/validation set, we use a window size of $k = 12$. We did not observe any significant differences when we tried different window sizes.

A. Investors Redemption/Subscription Prediction

For each experiment, we report average and standard deviation of sensitivity, specificity, F-score, area under ROC curve (AUC), and overall accuracy for the last six month of 2017. Also for each model, we report the effect of using the balanced data created using under-sampling technique, as described in section II-B. Figures 1,2 show the performance of each model for subscription and redemption prediction respectively.

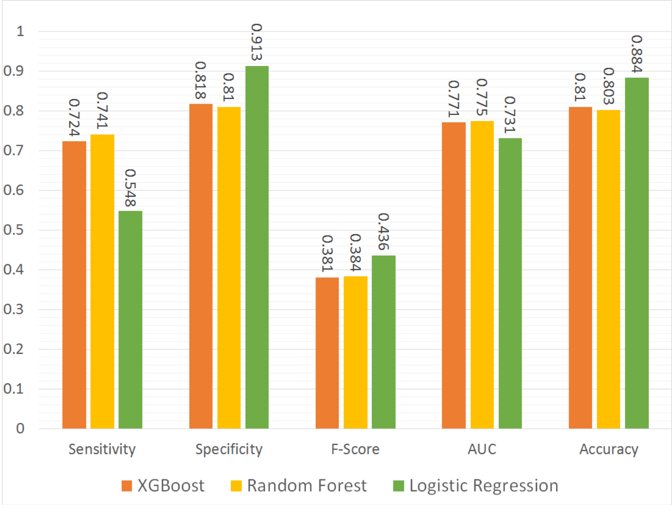

Fig. 1. Subscription prediction results

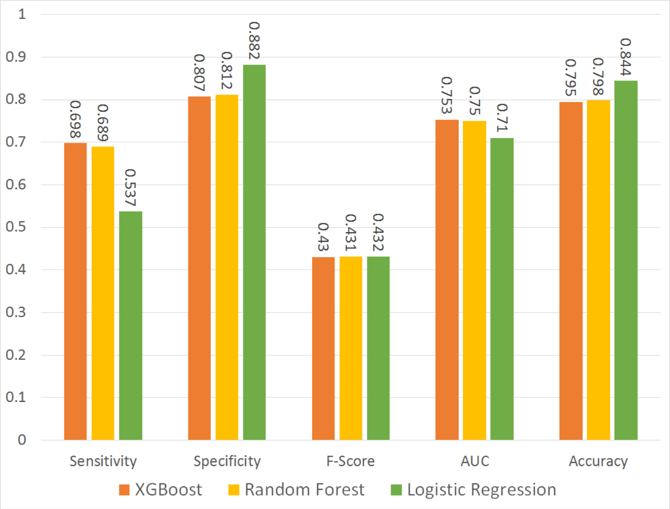

Fig. 2. Redemption prediction results

Based on the obtained results, prediction accuracy of redemption is slightly less than subscription. Sensitivity of logistic regression is far less than sensitivity in both random forest and XGBoost. This metric is really important for this problem, since the aim is to capture as much positive samples as possible, which represent redemption/subscription cases. At the same time we want to have higher precision. However, there is always a trade-off between sensitivity and precision (as shown by F-Score). In addition, AUC metric is higher in

Authorized licensed use limited to: ASTON UNIVERSITY. Downloaded on August 20,2025 at 19:37:54 UTC from IEEE Xplore. Restrictions apply.

both random forest and XGBoost in comparison with logistic regression. We found both XGBoost and random forest equally effective for red/sub prediction.

Table II shows an example of different investment strategies used by 14 different investors and their red/sub net amount from mid to end of 2017. The last column (probability) shows the probability of redemption which is predicted by the XGBoost model for the month of December 2017. Also, the highlighted rows show some of the cases that our model failed to predict the correct scenario based on historical red/sub data.

As shown in table II, there are various types of investment strategies. One strategy is a rather regular monthly red/sub pattern with relatively fixed amount (for example, investors one to four). All classifiers have high probability for such instances. This means such patterns are easy to be captured by the classifiers. Another strategy is regular monthly red/sub pattern with no fixed amounts (for example as investors five to 12). For most of these cases, all three classifiers still can predict correctly with relatively high degree of accuracy. As irregularity increases in the investment pattern, the probability is reduced which shows increasing uncertainty, such as for investors 13 and 14. For completely irregular scenarios, the prediction model is not able to capture them because of lacking related and contextual information. For example, investor two has a very regular redemption each month, but suddenly stopped redeeming in December 2017. Or for investors four and six who have a very regular monthly pattern, there is a big subscription on December 2017. These are the patterns that the prediction model has not experienced in the historic data and logically it is hard to predict them correctly. Nevertheless, there are some extreme cases that the prediction model did very good job and predict them correctly for instance investors 13 and 14.

Redemption prediction results for last six months of 2017 shows that in average 24.86% of the total investors are predicted as positive correctly, which covers 76.95% of the total redemption amount for all investors.

B. Top investors

To run a targeted campaign, we need to find top investors based on both subscription and redemption rates. We predict red/sub probability for the next month to validate our findings. One way to find the top investors is to sort them based on prediction probability. As we mentioned before, the highest prediction probability is related to the investors with regular investment pattern. Therefore, using prediction probability alone, does not capture investors who have irregular patterns and in most cases they have large red/sub amounts that are desirable to be captured. On the other hand, if we only focus on these type of investors and sort them for example based on *min* feature and keep only those with highest redemption amounts, then there would be a lot of false positive cases in the top list. Here we are looking for an intermediate solution that enables us to consider both high probability investors as well as investors with large amount of red/sub.

Therefore, we rank the investors based on various factors explained in section II-C and then we compare the average ranking and optimal ranking to obtain the final rank. Considering the prediction for June to December 2017, Table III shows the percentage of average redemption amounts captured by top investors along with the false positive percentage. Optimal rank is much better in terms of capturing redemption amounts (about 16 times more than average ranking) while they have relatively the same false positive percentages. Also Figure 3 shows the results of using optimal ranking for selecting the top 25 investors in terms of their predicted redemption. The first plot in the top left corner shows the objective function over time with the global minimum of 67.076. The histogram of the objective function scores at the last iteration is displayed in top right corner, while the third plot at the bottom shows the four individual features (*prediction probability*, *min*, *avg*, and $min \times predictionprobability$) and the obtained solution (CE) along with the average ranking (Mean) for each of the top 25 investors.

C. Discovering Prediction Results Reasons

Employing RF+MSGL method, we obtain 47 rules for redemption prediction with accuracy of 79% and AUC of 0.72. These results are very close to the results of the black box models such as random forest and XGBoost. This shows the RF+MSGL is able to mimic the random forest and XGBoost performance with much fewer number of rules. Table IV shows the most important extracted rules. column *Prediction* specifies the prediction label for each single rule and *accuracy* and *coverage* show the accuracy and percentage of the data which is covered by each rule, respectively. *Importance* is the weight associated to each rule which has been obtained during the optimization process in MSGL. Rules with higher importance absolute value have more effect to determine the final classification result.

Figure 4 shows the relationship between each top rule and its corresponding features. As shown here, $Redmp\_count$ has the highest influence in top extracted rules with participating in eight of them while Max and $M1$ were involved in only one of the top rules.

IV. Conclusion

In this work, we use historical transaction data of about 400 mutual funds and investigate the possibility of predicting subscription and redemption rates corresponding to each fund in the next month. Experimental results show that most of the patterns and behaviors can be predicted with high accuracy. In addition, the proposed investors ranking method covers considerable amount of total red/sub value. Finally, we are able to extract the most important features that derive the redemption transactions. For the future work, we want to add features related to the mutual funds, as we believe that fund features such as type of the fund, performance and holdings have important effect on red/sub events. Also, we are going to use deep neural networks to generate features from transaction

Authorized licensed use limited to: ASTON UNIVERSITY. Downloaded on August 20,2025 at 19:37:54 UTC from IEEE Xplore. Restrictions apply.

TABLE II
An example of different investment strategies used by 14 different investors based on their red/sub net amount from mid to end of 2017

Investor	2017-06	2017-07	2017-08	2017-09	2017-10	2017-11	2017-12	Probability
1	-79.98	-82.01	-82.76	-82.65	-83.41	-82.13	-82.21	0.932
2	-132	-132	-132	-132	-132	-132	0	0.932
3	-300	-1250	-1800	-300	-300	-300	-300	0.924
4	-1200	-1200	-1200	-1200	-1200	-1200	34892	0.924
5	-1905.71	-1761.95	-1761.97	-241762	-1761.95	-1796.14	-1761.95	0.913
6	-2714.47	-2714.47	-2714.47	-2714.47	-2714.47	-2714.47	79232.4	0.913
7	-4409.25	-1158.7	0	-1833.21	-4207.13	-1818.68	-5762.39	0.902
8	-2065	-2904.97	-3065	-3065	-2474.34	-2600	18400	0.902
9	192000	-117600	23000	-844000	-236400	60000	-34000	0.883
10	-10939.1	346.78	33733.71	-43320.7	-68493.3	-8935.15	-32457.8	0.882
11	20572.47	1180.92	-2817.43	35758.04	42687.44	-34591.9	-199.25	0.857
12	-3125	5702.66	-4673.36	24377.94	-13539.4	-9179.31	0	0.857
13	0	0	0	-80000	37486.75	0	-58429.9	0.857
14	0	-532	0	0	583.55	0	-16448.7	0.766

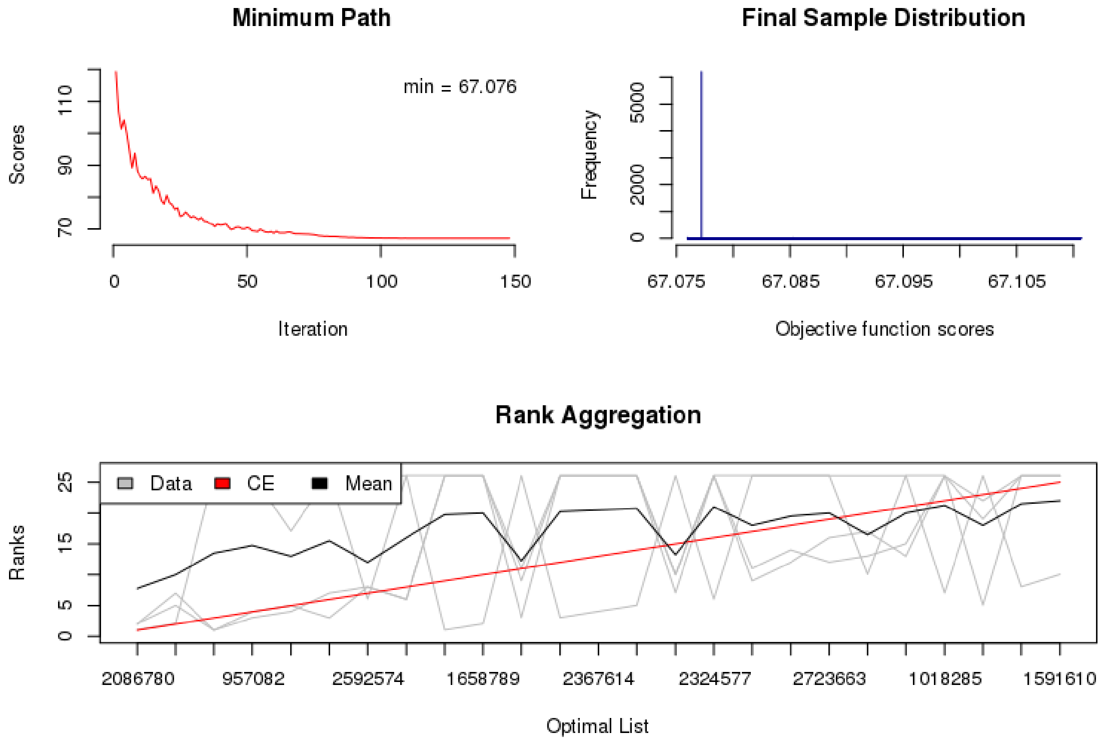

Fig. 3. Using optimal ranking for selecting the top 25 investors in terms of their predicted redemption. x-axis shows the investors' id in the dataset and the y-axis shows their corresponding rank

TABLE III
Comparison between optimal and average ranking

	Optimal ranking		Average ranking	
	Captured Redemption	False Positive	Captured Redemption	False Positive
Top 25 Investors	3.2%	33.7%	0.2%	32.6%
Top 50 Investors	4.83%	37.14%	0.31%	36%
Top 100 Investors	6.02%	30.14%	0.42%	38.6%

data, instead of hand-crafted features in the current work, that can potentially improve the performance of the model.

References

[1] Brad M Barber, Xing Huang, and Terrance Odean. Which factors matter to investors? evidence from mutual fund flows. *The Review of Financial Studies*, 29(10):2600–2642, 2016.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Qi Chen, Itay Goldstein, and Wei Jiang. Payoff complementarities and financial fragility: Evidence from mutual fund outflows. *Journal of Financial Economics*, 97(2):239–262, 2010.

Authorized licensed use limited to: ASTON UNIVERSITY. Downloaded on August 20,2025 at 19:37:54 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
THE MOST IMPORTANT EXTRACTED RULES USING RF+MSGL METHOD

Rule	Definition	prediction	Accuracy	Coverage	Importance
R1	redmp_count<=2.5 & median_incp>-0.265	0	0.68	0.59	0.1555
R2	redmp_count<=1.5	0	0.70	0.49	0.124
R3	median>-0.537 & mean<=923	0	0.63	0.65	0.0927
R4	mean_incp<=-2.84e-15 & redmp_count<=3.5	0	0.75	0.26	0.08
R5	std_incp>0 & min>-9.4	0	0.73	0.23	0.0734
R6	median>0 & mean<=1010	0	0.63	0.66	0.051
R7	mean<=500 & median>0	0	0.64	0.60	0.0423
R8	m_1<=0.5 & median>0	0	0.58	0.76	0.037
R9	mean<=834 & redmp_count<=3.5	0	0.70	0.52	0.0311
R10	max<=6440 & median>-0.537	0	0.67	0.53	0.0295
R11	mean_incp<=0 & redmp_count_incp<=3.5	0	0.77	0.22	0.0192
R12	redmp_count>8.5 & median_incp<=0	1	0.94	0.12	-0.0255
R13	mean_incp>59.7 & min<=-9.21	1	0.75	0.25	-0.0491
R14	median<=-0.537	1	0.92	0.15	-0.0506
R15	mean>49.6 & min_incp<=-9.4	1	0.74	0.24	-0.0547
R16	redmp_count>7.5	1	0.88	0.20	-0.0551
R17	min<=-0.45 & redmp_count_incp>4.5	1	0.79	0.34	-0.0703
R18	redmp_count>6.5 & redmp_count_incp >3.5	1	0.87	0.23	-0.081
R19	min<=-1.53 & redmp_count>3.5	1	0.80	0.34	-0.1379

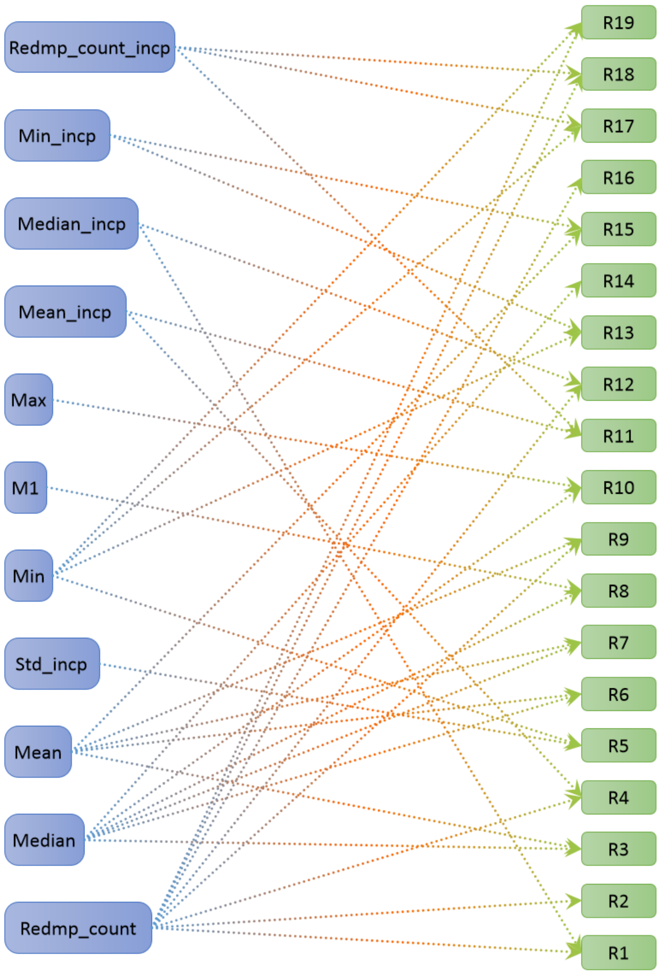

Fig. 4. Relationship between each top rule and its corresponding features

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[5] Zhao Chen. Research on forecasting method of balance treasure fund flow. In *Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2017 16th International Symposium on*, pages 173–176. IEEE, 2017.

[6] Diane Del Guercio, Egemen Genc, and Hai Tran. Playing favorites: Conflicts of interest in mutual fund management. *Journal of Financial Economics*, 128(3):535–557, 2018.

[7] Investor Economics. Analysis of factors influencing sales, retention and redemptions of mutual fund units, 2015 (accessed July 3, 2018).

[8] Vincent Glode, Burton Hollifield, Marcin Kacperczyk, and Shimon Kogan. Is investor rationality time varying? evidence from the mutual fund industry. In *Behavioral Finance: WHERE DO INVESTORS'BIASES COME FROM?*, pages 67–113. World Scientific, 2017.

[9] Mark Grinblatt, Seppo Ikäheimo, Matti Keloharju, and Samuli Knüpfer. Iq and mutual fund choice. *Management Science*, 62(4):924–944, 2015.

[10] Morteza Mashayekhi and Robin Gras. Rule extraction from decision trees ensembles: new algorithms based on heuristic search and sparse group lasso methods. *International Journal of Information Technology & Decision Making*, 16(06):1707–1727, 2017.

[11] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] Vasyl Pihur, Susmita Datta, and Somnath Datta. Rankaggreg, an r package for weighted rank aggregation. *BMC bioinformatics*, 10(1):62, 2009.

[14] Hassan Qamar and Sanjay Singh. Mutual fund performance prediction. In *Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), IEEE*, pages 185–189. IEEE, 2016.

[15] K Geert Rouwenhorst. The origins of mutual funds. 2004.

[16] Martin Vincent and Niels Richard Hansen. Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis*, 71:771–786, 2014.

[17] Nan Yu Wang, Sen Sung Chen, Chih Jen Huang, and Cheng Hsin Yen. Purchase and redemption decisions of mutual fund investors of variable life insurance-using quantile regression. *International Journal of Economics and Financial Issues*, 4(4):714–725, 2014.

[18] Mingzhu Zhu, Chao Xu, and Yi-Fang Brook Wu. Ifme: information filtering by multiple examples with under-sampling in a digital library environment. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 107–110. ACM, 2013.

Authorized licensed use limited to: ASTON UNIVERSITY. Downloaded on August 20,2025 at 19:37:54 UTC from IEEE Xplore. Restrictions apply.