Prediction of stocks with high transfer based on ensemble learning

Jie Ni¹, Linghong Zhang¹, Jiaming Tao¹ and Xiaorong Yang¹

¹College of Statistics & Mathematics, Zhejiang Gongshang University, Hangzhou, Zhejiang, 310018, China

*Corresponding author’s e-mail: xryang@zjgsu.edu.cn

Abstract. The “high transfer” dividend policy implemented by listed companies is a special phenomenon in the Chinese stock market. The stocks with high transfer can generate significant positive excess returns before and after the announcement date, so it is important for investors to predict the high transfer stock as precisely as possible. Previous studies mainly focused on regression models to make predictions, such as logistic and probit models, the prediction results of which are average. This paper uses the data of the Chinese A-share market for the last eight consecutive years and screens out 13 important factors through stock dividend theory. We construct three ensemble models for prediction, namely XGBoost, LightGBM, and CatBoost. The results show that the three ensemble models are far superior to the traditional Logistic model. And the prediction effect of the three ensemble models have been improved after Bayesian parameter tuning. Furthermore, three models are fused by simple voting and rank-score, it turns out that almost all metrics, including accuracy, precision, F1 value and AUC value, perform better than the single model. In addition, the top 10 and top 20 hit rate reach 100%, suggesting that the fusion model can predict the top 20 stocks with high transfer accurately, which has certain practical significance for strengthening the security of investment.

1. Introduction

In recent years, China's securities market has developed rapidly, and the Chinese stock market with a total value of more than 50 trillion has become the world's second largest capital market. As one of the core issues of corporate finance, dividend policy is always the focus of investors. Unlike foreign listed companies that pay more attention to cash dividends, Chinese listed companies show preference for stock dividends and stock splits. And the phenomenon of ‘high transfer’ refers to that the total proportion of stock dividends and stock splits is comparatively high, generally, more than 5 shares for every 10 shares.

The essence of high transfer is the internal adjustment of the owner's equity, which will not change the company's cash flow and profitability. However, since the stock price will be ex-righted after the implementation of the high transfer, it is feasible to profit from the stock appreciation in the secondary market by filling advantageous position. Historical data demonstrate that high transfer stocks can generate significant positive excess returns before and after the announcement date, that’s why high transfer turns into a hot spot for commercial speculation. In order to earn as much profit as possible in the high transfer, and avoid the retraction risk caused by share-buying after the ex-right, how to precisely predict the company with high transfer is crucial. If we can accurately predict the listed companies that may implement high transfer in the next year and purchase shares in advance, this has
certain practical significance for strengthening the security of investment and maintaining the stability of the capital market.

High transfer is actually a dividend policy. There are three mainstream distribution motivation theories of dividend policy in this area, namely signaling theory, optimal price theory and catering theory. Signaling theory believes that there is information asymmetry between the managers of listed companies and investors, and managers hope to release a positive signal to the market for the company's future development by implementing stock dividends[1-2]. Optimal price theory argues that the company implements the policy of dividend distribution, mainly to reduce the price of high-priced stocks and maintain the stock price in an optimal range, which attracts more investors[3]. Catering theory states that company management will cater to investors' preferences when formulating dividend policies, thereby maximizing company interests[4-5].

Based on the above theories, scholars incorporate corresponding factors into the models to predict the transfer situation. Hwang et al.[6] applied the Probit model to predict transfer situation within a year by variables, such as stock price, book value, and profit growth rate. Although this model only succeeded in predicting 30% of the companies with stock dividends and splits, it supported the idea of predicting the transfer situation by data containing companies' operation condition. Krieger et al.[7] further improved the accuracy of prediction based on the same Probit model. Xiong et al.[8] selected the sample data of all listed companies in Shenzhen and Shanghai stock markets from 2006 to 2010, and took advantage of Logit model for regression analysis, the accuracy rate of the high transfer in 2010 reached 90%. With the development of machine learning, all kinds of algorithm models are applied to the stock high transfer prediction. Wang et al.[9] used two different ensemble learning to build a forecasting model of high transfer stocks. One of them is a combined model based on the K-NN algorithm, Logistic regression and Decision Tree, while the other is built by AdaBoost and Random Forest algorithm. It turns out that the former combined model performs better with the evaluation index of G-mean value and Accuracy of prediction. Dong et al.[10] used the BP artificial neural network algorithm to establish a prediction model of high transfer, with a prediction accuracy of 92%, and explained factors contained in the established model.

In this paper, we construct a traditional logistic model as benchmark model based on the chosen factors. Then three Ensemble Algorithms, namely XGBoost, LightGBM, and CatBoost, are used to promote the model. We fuse three ensemble algorithms by simple voting and ranking methods, and this fusion model comprehensively improves the accuracy of high transfer prediction with excellent fitting effect, which has so far not been carried out.

2. Methodology

2.1. Data select and feature extraction

The data comes from the public data set of question A in the 2020 “Teddy Cup” data mining competition(www.tipdm.org). The data contains stock data of the Chinese A-share market for the last eight consecutive years. The dependent variable is whether the listed company implements high transfer, and the independent variables include multiple financial characteristics of the listed company and the market characteristics of the relevant stocks. Due to the large number of factors in the original data, we initially screen 23 factors according to the distribution motivation theories, then filter out 15 factors by single factor variance and mutual information and remove factors with a high correlation. Finally, we get 13 factors which have significant impact on high transfer for listed companies, they are shown in Table 1 below.
Table 1. Factors of high transfer.

| Factor category        | Factor name                             |
|------------------------|-----------------------------------------|
| Basic factor           | Stock price                             |
|                        | Total equity                            |
|                        | Time to market                          |
|                        | Amplitude                               |
|                        | Industry                                |
| Value factor           | Undistributed profit per share           |
|                        | Capital reserve per share                |
|                        | Cash flow per share                     |
|                        | Year-on-year growth of return on equity  |
|                        | Inventory turnover                       |
| Concept sector factor  | Whether small market capitalization      |
|                        | Whether private placement                |
| Timing factor          | Stock dividends and splits per share last year |

2.2. Ensemble learning

Ensemble learning builds and combines multiple weak learners to obtain a better and more comprehensive strong learner, thus improving the accuracy of prediction. The idea of ensemble learning was first proposed by Dasarathy and Sheela[11] in 1979. The basic principle is to generate a set of weak learners, and then combine them with a certain strategy to obtain superior generalization performance than a single learner. According to different combination strategies, it can be divided into two categories: Bagging[12] and Boosting[13], and XGBoost[14], LightGBM[15], CatBoost[16] involved in this paper all belong to Gradient Boosting Decision Tree framework (GBDT).

Compared with GBDT, XGBoost mainly uses the second derivative to construct a new loss function and make it the criterion of node splitting. In the process of constructing XGBoost, we perform one-hot encoding on category features so that it can be converted into the model with numerical features. Then, XGBoost searches the optimal split node by the presorting method, and generates leaf nodes by level-wise strategy. Finally, the sample's leaf node scores on each tree are synthesized as the raw score of the sample, which is input into the sigmoid function to obtain the predicted probability of the sample.

In order to reduce the calculation amount and improve the operation efficiency of the algorithm, LightGBM proposed the GOSS(Gradient-based One-Side Sampling) algorithm and the EFB(Exclusive Feature Budding) algorithm under the premise of maintaining a certain sample accuracy. The former is mainly used to reduce sample usage, while the latter aims to reduce the number of features. LightGBM can avoid performing one-hot encoding on category variables by specifying the index of category variables, which can solve the high-dimensional sparse problem caused by categorical variables. Then it uses a histogram algorithm for continuous variables, thereby reducing the amount of calculation when searching for the optimal node. The generation method of leaf nodes and the subsequent steps are exactly the same as XGBoost.

CatBoost can also process categorical features automatically, like LightGBM does. The model will convert categorical features into numerical features based on the target statistics of the categorical features. Catboost applies a completely symmetric tree strategy to generate leaf nodes. Similarly, the predicted probability of each sample is finally obtained.

2.3. Model training strategy

In our work, the first seven years' data of listed company is used as the training set, and data of the eighth year is used as the test set for model fitting. It is worth noting that the datasets are extremely imbalanced. Taking the training set as an example, the number of instances with high transfer is 3106
(positive samples), while the number of instances without high transfer is 15970 (negative samples), which is about 5 times the former. Therefore, the ratio of positive and negative samples in the training set is adjusted to 1:2 through the SMOTE oversampling technology. Then it is possible that the model can fully learn the characteristics of the high transfer samples, so as to distinguish between positive and negative samples better and improve the ability to identify companies with high transfer.

3. Results and Discussion

Table 2 shows the performance metrics of XGBoost, LightGBM, CatBoost, and logistic models on the test set under default parameters. The corresponding evaluation indicators are Accuracy, Precision, Recall, F1 value, AUC value and Hit rate. Among them, accuracy is to measure the overall accuracy rate of the model, which presents how many samples are predicted correctly in all samples; precision reveals the quality of the prediction result, which represents the probability of correct prediction in the positive judgment; recall means the probability of correct prediction in the affirmative category; F1 value is the harmonic mean of precision and recall; AUC is the area under the ROC curve (Receiver Operating Characteristic Curve); as for the hit rate, it is specifically designed for actual demand of investors. We forecast the top 10, 20, 50 and 100 stocks with the greatest probability of high transfer, the hit rate indicates the proportion of stocks with actual high transfer among the forecasted stocks.

| Index            | Logistic | XGBoost | LightGBM | CatBoost |
|------------------|----------|---------|----------|----------|
| Accuracy         | 78.62%   | 84.42%  | 84.85%   | 85.22%   |
| Precision        | 38.94%   | 49.66%  | 50.65%   | 51.49%   |
| Recall           | 67.85%   | 68.79%  | 72.71%   | 74.39%   |
| F1               | 49.48%   | 57.68%  | 59.70%   | 60.84%   |
| AUC              | 81.11%   | 89.66%  | 89.94%   | 90.60%   |
| Top 10 Hit rate | 60.00%   | 100.00% | 60.00%   | 80.00%   |
| Top 20 Hit rate | 55.00%   | 100.00% | 75.00%   | 90.00%   |
| Top 50 Hit rate | 68.00%   | 92.00%  | 86.00%   | 90.00%   |
| Top 100 Hit rate| 48.00%   | 87.00%  | 79.00%   | 84.00%   |

Comparing XGBoost with the Logistic model, it can be found that accuracy, precision, recall, F1 and AUC values predicted by the XGBoost model are superior to Logistic model. Accuracy is increased to 84.42%, a growth of nearly 6 percentage points. The recall rate has increased by nearly 1 percentage point. Futhermore, precision has been greatly improved by nearly 10 percentage points, indicating that the recognition effect of high transfer samples is remarkable. And hit rate of the top 10 and top 20 is 100%, indicating that the top 20 stocks predicted by XGBoost are exactly the same as the actual situation. Besides, the top 50 and 100 hit rates are still considerable, 92% and 87% respectively.

Similarly, the five basic evaluation indicators of LightGBM are also better than the benchmark model, even better than XGBoost. The recall rate has even been increased by nearly 4% relative to XGBoost. But unfortunately, LightGBM performs poorer on hit rate, which is far behind XGBoost.

Out of all the algorithms, CatBoost has the highest performance metrics with five basic indicators. It may be related to CatBoost's better use of categorical variables. The results displayed on hit rate are also well-behaved, compared with the logistic model and LightGBM. But there exists a certain gap between CatBoost and XGBoost in hit rate.

Overall, three ensemble learning models are far superior to the traditional statistical model. We may find that three models have enhanced the prediction effect from different aspects. It is still uncertain to determine the optimal model, so we need to make a deeper exploration based on three ensemble models.

To obtain better prediction results, the three ensemble models are developed through Bayesian optimization. Table 3 lists the prediction results after parameter optimization.
Table 3. Performance metrics after parameter optimization.

| Index         | XGBoost        | LightGBM       | CatBoost       | Untuned average |
|---------------|----------------|----------------|----------------|-----------------|
| Accuracy      | 92.53%         | 92.47%         | 91.66%         | 84.83%          |
| Precision     | 75.09%         | 76.35%         | 69.96%         | 50.60%          |
| Recall        | 77.20%         | 74.21%         | 80.56%         | 71.96%          |
| F1            | 76.13%         | 75.26%         | 74.89%         | 59.41%          |
| Top 10 Hit rate | 90.00%       | 90.00%         | 100.00%        | 80.00%          |
| Top 20 Hit rate | 95.00%       | 95.00%         | 100.00%        | 88.33%          |
| Top 50 Hit rate | 96.00%       | 96.00%         | 96.00%         | 89.33%          |
| Top 100 Hit rate | 94.00%      | 95.00%         | 96.00%         | 83.33%          |

Each performance metric of three ensemble models has been greatly improved after parameter adjustment. Specifically, compared with the simple average of three integrated models without adjustment, we get a better performance on accuracy, precision and recall. As for the hit rate indicators, they all have been increased to 90%. All of these results indicate that Bayesian tuning can significantly boost the prediction effect.

However, the prediction effects of three models with tuning are still relatively close. So we take the fusion of three ensemble models into consideration looking forward to better results. For a binary classification, there exists three primary methods, namely simple voting, weighted voting and rank-score. For the indicators of accuracy, precision, recall and F1 value, they are obtained by the fused model through simple voting, and the result with more votes is selected as the final prediction result. Moreover, the hit rate is sorted according to the predicted probability in advance, so the result of fusion is obtained by rank-score. Specifically, the predicted probability of high transfer by each model is sorted in descending order and scored. The highest probability gets 1 point, the second gets 2 point, and so on. We add the scores of three models and take the top 10, 20, 50, and 100 stocks with the lowest total score to calculate the hit rate. The detailed results are shown in Table 4.

Table 4. Performance metrics in Fusion model.

| Index    | Fusion model |
|----------|--------------|
| Accuracy | 92.60%       |
| Precision| 77.17%       |
| Recall   | 78.01%       |
| F1       | 77.58%       |
| Top 10 Hit rate | 100.00%   |
| Top 20 Hit rate | 100.00%   |
| Top 50 Hit rate | 98.00%    |
| Top 100 Hit rate | 96.00%    |

As can be seen from the table, all performance metrics of the fusion model have almost reached the highest level, except that Recall is slightly lower than CatBoost. It demonstrates that the fusion of models is definitely beneficial to prediction effect of high transfer and obtains the optimal prediction results.

4. Conclusions
Based on historical data of the Chinese stock market and factors selected through economic significance and statistical methods, the paper builds the traditional logistic model and ensemble models to predict stocks with high transfer, then fuses three ensemble models to get the optimal result. Finally, the following conclusions can be made.
(1) The three ensemble models of XGBoost, LightGBM and CatBoost perform better than the logistic model on each evaluation index, indicating that the ensemble model is better at capturing and learning the rules in the case of large-capacity and high-dimensional dataset.

(2) After the Bayesian parameter tuning, the prediction effects of the three ensemble models are improved and become comparable. Among them, Catboost performs best in recall and hit rate, while LightGBM in accuracy and XGBoost in F1 value. Three ensemble models have their own advantages in predicting high transfer stocks.

(3) XGBoost, LightGBM, and CatBoost are fused by means of simple voting and rank-score separately. In spite of a slight decrease in recall, the other performance metrics have been improved, suggesting that model fusion can effectively compensate for the defects of a single ensemble model.

(4) In order to meet the actual investment needs of investors, this paper constructs the top 10, 20, 50 and 100 hit rate indexs. The top 10 and top 20 hit rate have reached 100% in the fusion model, which means the top 20 high transfer stocks are completely consistent with the actual situation. The results can provide relevant references for investors and help them benefit from the stock market.

References
[1] Bhattacharya, S. (1979) Imperfect Information, Dividend Policy, and the Bird in the Hand Fallacy. Bell Journal of Economics, 10: 259-270.
[2] Grinblatt, M. S., Masulis, R. W., & Titman, S. (1984) The valuation effects of stock splits and stock dividends. Journal of Financial Economics, 13(4): 461-490.
[3] Baker, K., & Gary, K. (1993) Further evidence on managerial motives for stock splits. Quarterly Journal of Business and Economics, 32(3): 20-31.
[4] Baker, M., & Wurgler, J. (2004) A Catering Theory of Dividends. Journal of Finance, 59: 1125-1165.
[5] Li, W., & Lie, E. (2006) Dividend changes and catering incentives. Journal of Financial Economics, 80(2): 293-308.
[6] Hwang, S., Keswani, A. & Shackleton, M. (2005) Stock Splits: what does the Market Tell Us Ex Ante?. Working Paper, 40(2): 18-24.
[7] Krieger, K., & Peterson, D. R. (2009) Predicting stock splits with the help of firm-specific experiences. Journal of Economics and Finance, 410-421.
[8] Xiong, Y. M., Chen, X., Chen, P., & Xu, W. H. (2012) The motives of issuing stock dividends by Chinese listed firms—an empirical test based on a sample of high stock dividends. Research on Economics and Management, 000(005): 81-82.
[9] Wang, K., & Long, W. J. (2016) Research on High Transfer Stock Based on Ensemble Learning. Times Finance, (12).
[10] Dong, K. M., & Zhao, S. S. (2018) A Study on the Motivation of High transfer in Chinese Listed Companies—Analysis based on BP Neural Networks. Review of Investment Studies, 037(001):139-153.
[11] Dasarathy, B. V., & Sheela, B. V. (1979) A Composite Classifier System Design: Concepts and Methodology. Proceedings of the IEEE, 67(5): 708-713.
[12] Breiman, & Leo. (1996) Bagging predictors. Machine Learning, 24(2): 123-140.
[13] Schapire, R. E. (1989) The Strength of Weak Learnability. Proceedings of the Second Annual Workshop on Computational Learning Theory, 5(2): 197-227.
[14] Chen, T. Q., & Guestrin, C. (2016) Xgboost: a scalable tree boosting system. In: knowledge discovery and data mining. San Francisco. pp: 785-794.
[15] Ke, G. L., Meng, Q., Thomas, W. F., Wang, T. F., Chen, W., Ma, W. D., Ye, Q. W., & Liu, T. Y. (2017) LightGBM: a highly efficient gradient boosting decision tree. In: neural information processing systems. Long Beach. pp:3149-3157.
[16] Prokhorenkova, L. Q., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. CatBoost: unbiased boosting with categorical features. In: neural information processing systems. Montréal. pp: 6638-6648.