Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 755–759 ISSN 2300-5963 ACSIS, Vol. 39
DOI: 10.15439/2024F3560

Key Financial Indicators Analysis and Stock Trend Forecasting Based on a Wrapper Feature Selection Method

Chang Lin
State Key Laboratory of Information Photonics and Optical Communications
Beijing University of Posts and Telecommunications
Beijing, China
bupt.ipoc@yandex.com

***Abstract*—Predicting stock price trends is a challenging puzzle. The immediate price of a stock is affected by an uncountable number of factors. Thus there is essentially no way to accurately predict short-term stock price due to dynamic, incomplete, erratic, and chaotic data. However, by analyzing key financial indicators, it is possible to gain an accurate understanding of a company's operations, make a quantitative assessment of its value, and thus make a reasonable prediction of the long-term trend of its stock price. In this FedCSIS 2024 Data Science Challenge, participants are asked to predict the trends of the stocks which are chosen from the Standard & Poor's 500 index. In this paper, we apply a wrapper feature selection method that tightly combines the steps of feature selection and model building to result in better prediction models, and provide insight into the indicators. After selecting the best set of features, we train two kinds of gradient boost machine: multi-classification model and regression model for class and risk-return performance prediction respectively. Finally a high confidence voting strategy is used to determine the kind of trading action (buy, sell, or hold). Experimental and competition results demonstrate the effectiveness of the methodology in this paper.**

***Index Terms*—Financial Indicator, Stock Trend Prediction, Feature Selection, Gradient Boosting Decision Tree, Strategic Voting**

I. Introduction

PREDICTING stock price trends is a challenging puzzle. Researchers generally agree that there are few ways to accurately predict the direction of the stock market over the next few days or weeks, but it may be possible to make price predictions for next years with meticulous study. With the rapid development of technologies such as artificial intelligence and global digitization, the prediction of the stock market has entered a technologically advanced era. Many analysts and researchers have developed various Artificial Intelligence (including Machine Learning and Natural Language Processing) based tools and techniques to predict stock price movements and help investors in proper decision-making. For example, Leippold et al. investigate 11 machine learning method's (such as ordinary least squares regression, least absolute shrinkage and selection operator, elastic net, gradient boosted regression trees, random forest etc.) predictive power in the Chinese stock market [1]. They build and analyze a comprehensive set of return prediction factors of Chinese market and find that the most critical factors have entirely different characteristics than the US market. They also show that machine learning methods can be successfully applied to various markets with different characteristics. Wu et al. present BloombergGPT, a 50 billion parameter language model trained on a wide range of financial data, and demonstrate that their model outperforms existing models on financial tasks by significant margins [2]. However, it's worth mentioning that it took about 53 days to train the BloombergGPT at a cost of around $3M. Basically, all deep learning based algorithms (e.g., various time series analysis methods based on Long Short-Term Memory and its variants) require extensive training on large and versatile datasets, incurring high training costs.

In the FedCSIS 2024 Data Science Challenge: Predicting Stock Trends [3], [4], participants are asked to develop a predictive model to accurately forecast stock trend movements based on the provided financial fundamental data. The selected stocks are chosen from 11 industry sectors of the Standard & Poor's 500 index, spanning 10 years. The dataset contains 117 fields (58 key financial indicators and 58 absolute changes of these indicators, and 1 industry sector) for 300 companies. For this kind of non-time-series tabular data, conventional machine learning methods (e.g., GBM-like algorithms) are well suited for modeling and analysis. In this competition, we build two Gradient Boosting Machine(GBM) models to predict whether it is a good moment to buy, sell or hold the stock, and under what circumstances the performance of investments can be maximized, respectively. We wrapped the predictive models into our proposed feature selection framework [5], [6]. Therefore, we can eliminate the influence of ineffective indicators, and find out the factors that play a key role in the predictive models. The resulting models are concise, accurate, have strong generalization capabilities, and can be well interpreted. The similarities between our work and the work of Rakićević et al. [7] are: improve the predictor's performance and provide a deeper insight into each of the indicators used for prediction.

As usual, the competitions of KnowledgePit are always well organized. The organizers carefully reviewed the competitors'

solutions, objectively assessed the novelty of their approaches and the quality of the submitted reports. This effectively prevents improper behaviors commonly found in other competition platforms, and ensures each competitor's solution stands the test of time. Moreover, for almost all the competitions that have been held, the organizers analyzed and summarized the methods submitted by the competitors and present informative papers [8], [9]. This allows the participants to identify the shortcomings of their own approaches and learn from the strengths of others. Driven by this favorable atmosphere, we are very happy to share our findings. This paper is organized as following: In this section, we introduce the background of the research. In section II we provide the analysis of the data. In section III we present a feature selection method which embedding the gradient boosting decision tree (GBDT) algorithm into a sequential floating forward and backward framework. We select different feature subsets to train two GBDT models, and use the ensemble of these models to predict stock trends. Section IV shows the experimental results and analyzes the role of each financial indicator in the trend prediction. The last section draws the conclusions and makes some recommendations.

II. Data Analysis and Processing

The available training data in this challenge contain 8,000 instances with fundamental financial data in a tabular format. Each instance in the data represents a financial statement announcement for one of the chosen 300 companies. It contains information on the company's sector, values for 58 key financial indicators, 1-year absolute change of each indicator, target class information, and risk-return performance for a period after the announcement. The target class is a single number from the set {1, 0, -1} that indicates the predicted trading action (buy, hold, sell correspondingly) for the event. The test data which containing 2,000 instances have the same format and naming scheme as the training data but it does not contain columns 'target class' and 'risk-return performance'. The available data contain two distinct types of missing values that have different semantics. One corresponds to non-available/missing information which is marked by "NA" string and another one can be interpreted as non-applicable (there is no value) which is just an empty string.

We use target encoding method to encode the industry sector feature. For null and NA values, we simply set them to two specific numbers (e.g., -300 and -999) that are different from the other normal values, then our algorithm can handle automatically. Then we check the correlation coefficients between the financial indicators and the forecast targets (class and performance). Table I gives the 10 most correlated features' Spearman correlation coefficients. As can be seen from this table, there is no significant correlation between any financial indicators and the forecast targets. This demonstrates that stock change trends are the result of a combination of many complex factors. It is difficult to provide a comprehensive interpretation and make an accurate prediction of the stock trends based on a limited number of indicators.

TABLE I
10 most correlated features' Spearman coefficients

Class		Performance	
Coefficient	***Feature Id***	***Coefficient***	***Feature Id***
0.070830	dI6	0.102539	I57
0.064429	dI7	0.081737	dI6
0.064336	I57	0.080568	I9
0.058301	I6	0.080181	dI47
0.055730	dI57	0.080143	Group
0.054116	dI47	0.079727	dI52
0.053434	dI9	0.079392	I18
0.053369	I8	0.076056	dI7
0.051238	I18	0.073477	I4
0.050009	I9	0.070641	dI57

Fig. 1 gives the 'boxplot' of indicator I57 and dI6. As can be seen in Figure 1, the data of these two features are very concentrated in the center and have a long spread on both sides. This is very similar to a normal distribution. Many indicators have similar distribution properties. It can also be found from the figure that these features have the same distribution in the training set and the test set. By comparing the distribution of each feature, we are confident that the data in the training and test sets have the same distribution pattern.

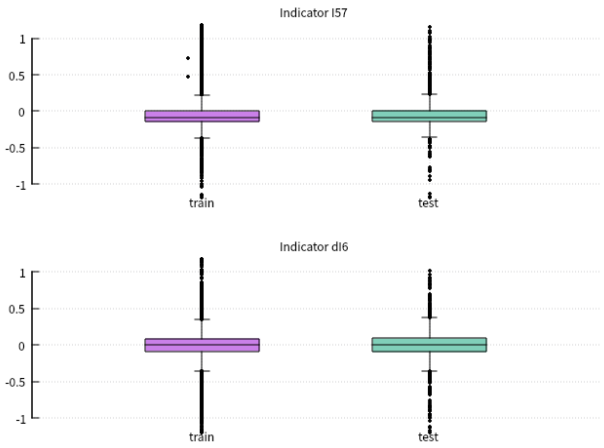

Fig. 1. 'boxplot' of indicator I57 and dI6.

Fig. 2 shows the risk-return performance of each industry sector. It is easy to find some interesting facts in this chart, for example, if one invests in energy(G3) stocks, there is a huge probability that one will lose money.

Fig. 3 shows the 'boxplot' of risk-return performance. From Fig.3 we can find that the data of risk-return performance are mainly concentrated in the range from -0.373 to 0.439, and it approximates a normal distribution. We also find that the correspondence between target 'class' and 'performance' can be described by the following equation:

$$class = \begin{cases} 1, & perform > 0.04; \\ -1, & perform < -0.015; \\ 0, & otherwise; \end{cases} \quad (1)$$

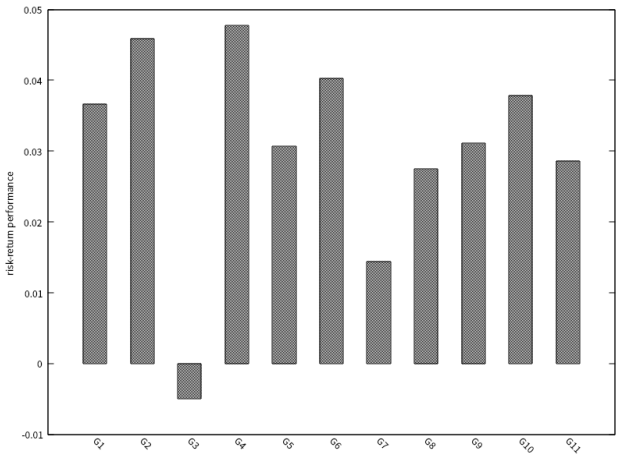

Fig. 2. risk-return performance of each industry sector.

Thus the stock trend prediction problem can be solved in two ways. First we can treat the task as a 3-classification problem for predicting a trading action (buy, sell, hold). We also can treat the task as a regression problem to fit the risk-return performance.

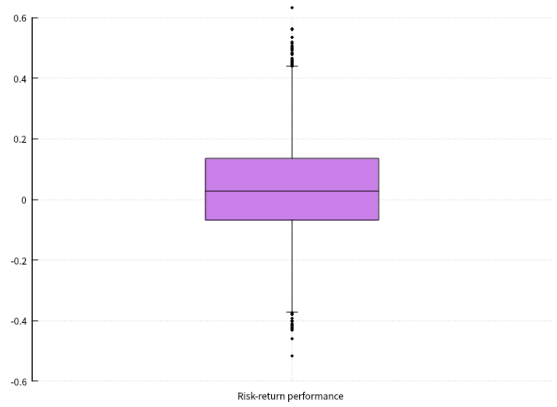

Fig. 3. boxplot of risk-return performance.

We also try to construct new features based on the provided financial fundamental data by adding, subtracting, multiplying and dividing. But they are found to be largely unhelpful in the prediction of stock trends.

III. Methodology

The methodology we use in this competition is very concise. We use the algorithm proposed in paper [5], [6] for feature selection. This algorithm essentially is a sequential floating forward and backward method. Its main improvement is that it embeds the GBM algorithm into the feature filtering framework. The procedure of feature selection is divided into forward and backward steps, as shown in Fig. 4.

In forward step, we sequentially select features one by one from the candidate set, add it to the selected set, use them to train the GBM, and evaluate the role played by each feature by comparing the results of each training, then move the L best features from candidate set to selected set. L is determined by the improvement of prediction accuracy.

In backward step, we sequentially drop a feature from the selected set and use the remains to train the GBM, evaluating the role played by each feature by comparing the results of each training, then move the R worst features from selected set to candidate set or directly drop the worst features according to the evaluation scores. R is determined by the loss of prediction accuracy.

Repeat this two steps until evaluation scores cannot be improved.

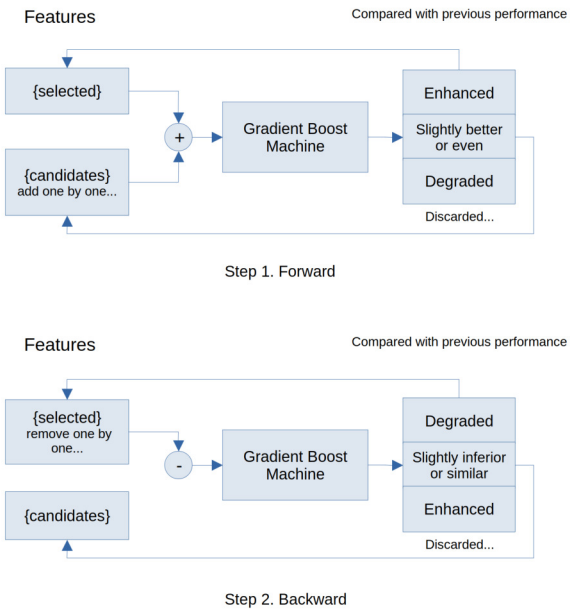

Fig. 4. Sequential floating forward and backward feature selection method.

After selecting the best set of features, we train 30 GBM multi-classification models and take their average for class prediction. We also have tried quite a few other methods, none of which are superior to the GBM. And we find that usual ensemble methods do not work here, because the prerequisite of ensembling a set of weak classifiers to a strong classifier is that the accuracy of each weak classifier must be slightly greater than 50%.

Using the same procedure, we train 30 GBM regression models and take their average for performance prediction.

Finally, we use (1) to transform the 'performance' value into a classification result, which is then combined with the 'class' value by voting. We calculate the weight of each vote according to its 'performance' value or the probability of its 'class'.

IV. Experimental Results and Analysis

We select 50 features from the indicators and sector information to train the three-classes classifiers. The learning curve (decrease of softmax loss) of our GBM is shown in

Fig. 5. Fig. 5 also gives the training results of xgboost. Using our classifiers to predict the trading action (buy, sell, hold) yields (by 4-fold cross-validation): softmax loss = 0.982, classification accuracy = 50.7%, and average error cost = 0.8405. The hyper-parameters of our GBM and xgboost are simply set to: learning rate = 0.01, gamma = 0.01, lambda = 4.0, min_child_weight = 20, num_round = 300. In this competition, the error cost matrix is defined as:

$$\begin{bmatrix} p/t & -1 & 0 & 1 \\ -1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{bmatrix}$$

p: prediction, t: truth

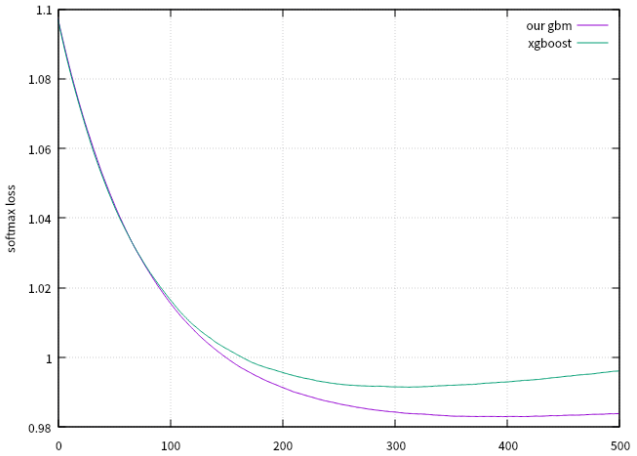

Fig. 5. The learning curve of the classifier.

We evaluate the importance of each selected feature by computing its contribution to the total gain. Fig. 6 shows the gain contribution of 10 most important features in the classification models. As can be seen from Fig. 6, indicators such as I57(Cash Flow from Operations Pct of Capital Expenditures), I5(Excess Cash Margin), dI52(1-year Absolute Change of Cash Ratio) et al. play important role in the classification, but their importance is not decisive. The contribution of each of these 50 features to the total gain only ranged from 1% to 3.3%. We think that the most crucial thing in this procedure is that we drop a large number of invalid features which are not closely related to the classification problem and tend to degrade the performance of the classifier, thus improving the accuracy and generalization ability of the classifier.

When training the regression models to fit the risk-return performance, we chose fewer features, just 39. The learning curve (decrease of mean-square error) of our GBM is shown in Fig. 7. Fig. 7 also compares the training results with xgboost. The mean-square error (MSE) of our regression model is 0.144. Using (1) to convert 'performance' value to 'class' probability, get average error cost = 0.8015.

Fig. 8 shows the gain contribution of 10 most important features in the classification models. As can be seen from

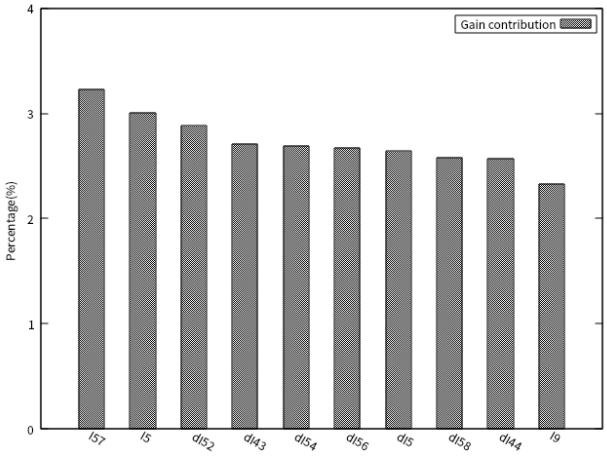

Fig. 6. Gain contribution of 10 most important features in the classification models.

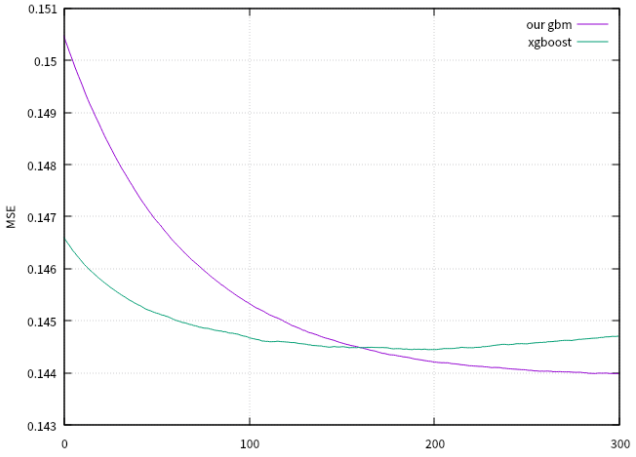

Fig. 7. The learning curve of the regression model.

Fig. 8, indicators such as I57(Cash Flow from Operations Pct of Capital Expenditures), dI58(1-year Absolute Change of Price to Cash Flow from Operations per Share), dI47(1-year Absolute Change of Cash & Cash Equivalents to Total Assets) et al. play important role in the regression, but their importance also is not decisive.

Combining the results of classification and regression, we get average error cost around 0.79x. Here we use a voting strategy that sets the prediction value to 0 by default; sets the prediction value to 1 when and only when 'performance' has a large positive value and 'class = 1' has a high probability; and sets the prediction value to -1 when and only when 'performance' has a large negative value and 'class = -1' has a high probability. This ensures that our predictions have a high degree of confidence.

We used 4-fold cross-validation in local test, and the obtained scores ranged from 0.790 to 0.799. Since the score of public leader-board was evaluated by only 200 instance, there

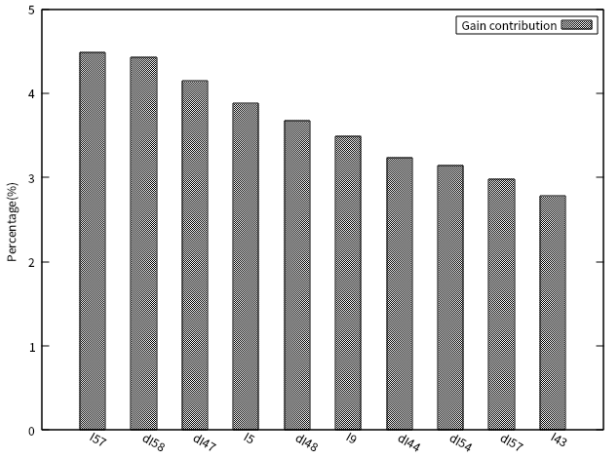

Fig. 8. Gain contribution of 10 most important features in the regression models.

is a significant difference between the public LB scores and the CV scores. Of our 20+ valid submissions (net of tests, obvious errors), most of our final scores largely better than 0.805, with 5 scores better than 0.790, and the best one is 0.7875.

V. Conclusion

The task of this competition is the prediction of stock trends. However it is more like estimating the return on investment of each company by analyzing its various financial indicators. With known data, we show that the methods proposed in this paper are concise, reliable and have excellent generalization ability. We achieved the desired results despite that we did not have enough time for fine-tuning the parameters and did not try hard to fit the test set. Our methods can provide a quantitative assessment of each financial indicator. It can be used as a good financial analysis tool.

Our research shows that the crux of stock trend forecasting is to select the indicators that are truly favorable for classification and regression in this task, and to buy or sell stocks when there has high degree of confidence, otherwise, 'hold' or just 'stay on the sidelines'. But our study also illustrates that there are no financial indicators that can directly influence stock trends, in other words there is no obvious causal relationship between them. Stock price movements are still governed by a large number of dynamic or unknown factors. Whether the methodology of this paper can be directly applied to stock trading needs to be verified by more tests. Interested researchers are welcome to share and discuss together.

We would like to thank the sponsors and organizers for providing such valuable research data and organizing the competition with great effort.

References

[1] M.Leippold, Q.Wang, W.Zhou. Machine-learning in the Chinese Stock Market. Journal of Financial Economics, 2022, 145(2): 64-82. DOI: https://doi.org/10.1016/j.jfineco.2021.08.017.

[2] S.Wu, O.Irsoy, S.Lu, V.Dabravolski, M.Dredze, S.Gehrmann, P.Kambadur, D.Rosenberg, G.Mann. BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564v3 [cs.LG]. https://doi.org/10.48550/arXiv.2303.17564.

[3] https://knowledgepit.ai/fedcsis-2024-challenge/.

[4] Aleksandar M. Rakicevic, Pavle D. Milosevic, Ivana T. Dragovic, Ana M. Poledica, Milica M. Zukanovic, Andrzej Janusz, Dominik Slezak: "Predicting Stock Trends Using Common Financial Indicators: A Summary of FedCSIS 2024 Data Science Challenge Held on KnowledgePit.ai Platform". In: Proceedings of FedCSIS 2024 (2024).

[5] C. Lin. Predicting Frags in Tactic Games using Machine Learning Techniques and Intuitive Knowledge. 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Brisbane, Australia, 2023, pp. 11-15, doi: https://doi.org/10.1109/ICMEW59549.2023.00008.

[6] C.Lin. Tackling Variable-length Sequences with High-cardinality Features in Cyber-attack Detection. Proceedings of the 18th Conference on Computer Science and Intelligence Systems, Vol.35, pages 1295-1299 (2023). DOI: https://dx.doi.org/10.15439/2023F2385.

[7] A.Rakićević, A.Poledica, B.Petrović. A Novel IBA-DE Hybrid Approach for Modeling Sovereign Credit Ratings. Mathematics 2022, 10, 2679. https://doi.org/10.3390/math10152679.

[8] A.Janusz, A.Jamiołkowski, M.Okulewicz. Predicting the Costs of Forwarding Contracts: Analysis of Data Mining Competition Results. Proceedings of the 17th Conference on Computer Science and Intelligence Systems, ACSIS, Vol.30, pages 399-402 (2022). DOI: https://dx.doi.org/10.15439/2022F303.

[9] M.Czerwinski, M.Michalak, P.Biczyk, B.Adamczyk, D.Iwanicki, I.Kostorz, M.Brzeczek, A. Janusz, M.Hermansa, L.Wawrowski, A.Kozlowski. Cybersecurity Threat Detection in the Behavior of IoT Devices: Analysis of Data Mining Competition Results. Proceedings of the 18th Conference on Computer Science and Intelligence Systems, ACSIS, Vol.35, pages 1289-1293 (2023). DOI: https://dx.doi.org/10.15439/2023F3089.