Forecasting A-share stock return by using machine learning-based Financial Statement and Macroeconomic Variables analysis

**Research background**

Recently, an increasing number of scholars and traders have researched various factors related to changes of stock price and constructed possible stock return patterns to complete prediction, discovering the return anomalies that could not be explained by traditional asset pricing models. According to Ou and Wang (2009), it is feasible for investors to hedge against potential market risk and for speculators to capture the chance of making the profit, if they obtain the accurate prediction of the stock market. Among the numerous empirical factors, financial statements and economic variables are two prior categories that may indicate the movement of stock price, and previous studies documented the credible correlations between indicators in these categories and stock returns.

The volume of electronical historical data dramatically increases during last three decades and will continue to grow in the future (Kannan et al., 2010). Therefore, considerable amounts of dissertations have verified the feasibility of constructing machine learning models to predict the stock market by either fundamental analysis and technique analysis. Baba and Kozaki (1992) developed a neural network system for forecasting the stock price in Japanese market. The reinforcement learning was also proved as a prediction method for stock trading (Lee, 2001).

The most previous literature selected top important indicators respectively from financial statement and macroeconomic variables by using various machine learning algorithms. However, when financial statement variables and macroeconomic variables are simultaneously covered in the algorithm, the empirical result may show the distinct list of top indicators. Another gap in this field is that researchers devote more time to exploring developed stock markets instead of developing markets, and therefore, less information about indicators in developing countries could be obtained. This paper focuses on researching the list of top important indicators in the Chinese stock market and predicts the stock returns based on these indicators, providing appropriate trading strategy to investors.

**Research questions**

This paper is going to use the machine learning model to predict the China’s A-Share return for one month, three months and six months (considered the effect of drift) after the announcement of quarterly financial statements and quarterly macroeconomic reports (the project converts monthly variables into quarterly variables by following conversion rules or calculating average value, and if the quarterly form is unavailable, the project will use last year’s annual variables). The empirical results will answer the following questions:

- What are important indicators in the prediction model?

- What is the contribution of each important indicator to the prediction model?

- Is it possible to use indictors to construct a scoring system for selecting individual stocks?

- What is the return performance of the stocks selected by this scoring system if the long-only position strategy is adopted?

**Research objectives**

- To identify important indicators from financial statement variables and macroeconomic variables by using Random Regression Forest.

- To quantitatively measure the contribution of each selected indicator by Random Regression Forest’s the variables importance measures (VIM).

- To determine the accuracy and reliability of the stock scoring model by comparing the monthly return of the selected stock with the market return.

**Literature review**

The earliest concepts related to fundament analysis could be traced into work of Graham et al. (1934), who indicated that purchasing portfolio of common stocks should carefully investigate three crucial factors, dividends information (dividend rate and history of paying dividends), earnings gained from income statement, and asset value evaluated from balance sheet. Ou and Penman (1989) researched the relationship between 68 variables and stock price by using LOGIT in that coefficient estimates with , proving that stock price did not reflect some fundamentals captured by financial statement. Different from research of Ou and Penman (1989), the work of Holthausen and Larcker (1992) predicted three excess return metrics, making the contribution to the quantitative analysis of stock return forecasted by variables in financial statement. Recently, top fundamental signals consisting of accounting variables and financial ratios were selected from 18000 records by using Bootstrap model, and in the data mining approach these signals show predictive ability for stock return (Yan and Zheng, 2017).

In the period of 1980s, the fundamentals increased 70% to the predictive ability of excess return, and macroeconomic variables would strengthen the fundamental analysis, adding an important dimension to contextual analysis of stock return (Lev and Thiagarajan, 1993). Ratanapakorn and Sharma (2007) used Vector Error-Correction Model (VECM) and Granger Causality Testing to reveal causality between the US stock return and 6 variables, emphasizing the direct effect of variables on the stock return in the long-run causality channel. Vijh et al. (2020) compared 6 different machine learning algorithms on predicting stock returns by analyzing from financial statement variables.

Dhar and Chou (2001) compared 4 nonlinear machine learning methods (including artificial neural network (ANN), genetic algorithm (GA), classification and regression tree (CART), and Naïve Bayes (NB)) with linear regression on predicting the distinctive cumulative abnormal returns in the proximity of the earnings announcement day. Unfortunately, the work of Dhar and Chou did not delve into the correlation between specific variable and stock returns. Zhang et al. (2004) enhanced the predicting accuracy of EPS by including a group of fundamental accounting variables and emphasized that only nonlinear method could achieve incremental value of fundamental information.

Though a series of research have minutely selected top indicators from financial statement and macroeconomic variables as well as achieved the objective for forecasting stock returns, there is a gap that should be focused. Financial accounting variables and macroeconomic variables are independently considered in previous models, and less empirical results represent the list of top indicators when these two categories are integrated. Therefore, this paper makes an improvement to fundamental analysis, and it is covering financial statement variables and macroeconomic variables simultaneously in the algorithm and listing the top indicators in China’s A-share market with the short-selling constrains in China. Under the long-only position trading strategy, which adds values to return of stocks (Leippold et al., 2021), I assume a scoring system for selecting individual stocks, and test the reliability of the system in a statistical way.

**Research approach and Methodology**

Random Regression Forest (CART) (in [appendix 1](#appendix1_1)) is selected as the main machine learning model to train data, completing prediction. Considering the huge amount of data to be processed in a limited time, this paper does not compare the Random Regression Forest model with other models, which is also in line with our expected academic contribution.

Data

This project uses all quarterly variables from China Stock Market & Accounting Research (CSMAR) file which contains about 234580 balance sheets, 222403 income statements, and 146807 cash flow statements from 2000 to 2020, and the project can initially get 291 financial statement variables for around 1800 firms. Meanwhile, CSMAR also provides nearly 457000 trading records on A-share stocks.

In this paper, 42 macroeconomic variables are from Wind Financial Terminal (Wind), a professional platform providing financial information services, and shown in [appendix 2](#appendix2_1).

In the data cleaning, the project may delete missing data and neglect irrelevant variables, so the number of variables and records participating in the prediction will be much less than the raw data.

Research Design

1) Assumptions of Predictions and Features

Because of the short position constrains in China, this paper deploys buy and hold abnormal return (BHAR) to be representative of monthly stock return, the mathematical symbol of which is , where stands for the Year, and stands for the quarter of the same year.

where stands for company c, stands for company’s return on day , and stands for the return of CSI 300 Index on day . Returns are compounded from one day before the announcement day utile .

In this paper, denotes to company c’s financial and macroeconomic variables in quarter of Y year, therefore stands for the number of variables.

Therefore, stands for one sample.

2) Rolling Window Construction

Data set is divided into training data and test data. The research assumes that the samples in the first three quarters of each year are used as training data and the samples in the fourth quarter are used as test data, and therefore, the rolling window is shown below:

Testing sample: 1 quarter

Training sample: 3 quarters

The training sample is:

The testing sample is:

3) Data Processing and Data Analysis

The research will follow the standard approach of random regression forest to complete the selection of important indicators and the calculation of the contribution of important indicators in prediction. Then, the research will assume a stock scoring system based on the results obtained from the random region forest to select valuable stocks, and then use traditional statistical methods, such as p-value, to test feasibility and reliability of the system. Each link in data processing and data analysis will have appropriate evaluation criteria, but the specific evaluation criteria need to be constantly adjusted in the research process, so relevant literature may need to be read.

Rationales

Random Regression Forest is defined by Breiman (2001), and is an effective tool in prediction because of avoiding overfitting and injecting appropriate randomness. After this model was proposed, it has been applied to various fields, and the field of stock return is also a hot topic. Krauss et al. (2017) developed a statistic arbitrage strategy based on Random Forest and successfully used it on the S&P 500. Nti (2019) used Random Forest model to select the top features from 42 macroeconomic variables and answered which macroeconomic variables affect which sector stocks. Amel-Zadeh et al. (2020) regarded Random Forest as a better model to select the top financial accounting variables and then to predict stock returns, after testing 5 machine learning models. Therefore, it is confident to use Random Regression Forest algorithm to achieve the research objectives.

**Execution plan**



The Gantt chart above has described the tasks in the project and shows the start time and duration of each task. The following four tasks are difficult to perform:

Learning CART

The principle of random region forest should be learned, including the principle of selecting important indicators and calculating indicators’ contribution.

Data Preprocessing

In this task, the project will calculate the financial ratios of each listed company from 2000 to 2020. In addition, it is also necessary to convert the cleaned data into data sets that can be recognized by the machine learning algorithm.

Data Processing

The random region forest algorithm will be implemented with Python code, and then the relationship between variables and stock return is explored, and the experimental results are visualized.

Testing System

The project intends to use SAS or SPSS software to test the effectiveness of the stock scoring model we constructed in research by hypothesis testing (p-value). The specific operation process can be defined around Feb 10 next year.

**Potential risks and feasibility analysis**

The project does have potential risks, which are mainly reflected in the following:

1. Raw data must contain missing data. If missing data is simply preprocessed, there may be underfitting, which will affect the accuracy of our model.

2. The cycle of rolling window may be unreasonable. Each company has weak and peak business every year; if the data of the previous three quarters are used as training data and the data of the fourth quarter are used as test data, the results may be influenced.

3. Due to the huge amount of data, the requirement for computing power of the computer is particularly high.

Despite the potential risks, the project can still be implemented. Firstly, the missing data can be simply processed, such as deleting directly. Because random forest selects samples by bagging, the problem of underfitting can be effectively avoided. Secondly, previous literature used different rolling windows, and no conclusion has been reached in terms of. Before the obtaining results of experiment, it is difficult to judge the accuracy of rolling windows assumed. Finally, the previous research used computers to complete such complex tasks, so I believe that these data can be processed under the cost limit of existing computer operations.

**Reference list**

Amel-Zadeh, A., Calliess, J.-P., Kaiser, D. & Roberts, S. (2020). Machine Learning-Based Financial Statement Analysis. *Available at SSRN 3520684*.

Baba, N. & Kozaki, M. (1992). An intelligent forecasting system of stock price using neural networks. *Proceedings of IJCNN International Joint Conference on Neural Networks*. IEEE, 371-377.

Breiman, L. (2001). Random forests. *Machine learning,* 45**,** 5-32.

Dhar, V. & Chou, D. (2001). A comparison of nonlinear methods for predicting earnings surprises and returns. *IEEE Transactions on Neural Networks,* 12**,** 907-921.

Graham, B., Dodd, D. L. F. & Cottle, S. (1934). *Security analysis*, McGraw-Hill New York.

Holthausen, R. W. & Larcker, D. F. (1992). The prediction of stock returns using financial statement information. *Journal of accounting and economics,* 15**,** 373-411.

Kannan, K. S., Sekar, P. S., Sathik, M. M. & Arumugam, P. (2010). Financial stock market forecast using data mining techniques. *Proceedings of the International Multiconference of Engineers and computer scientists*. 4.

Krauss, C., Do, X. A. & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research,* 259**,** 689-702.

Lee, J. W. (2001). Stock price prediction using reinforcement learning. *ISIE 2001. 2001 IEEE International Symposium on Industrial Electronics Proceedings (Cat. No. 01TH8570)*. IEEE, 690-695.

Leippold, M., Wang, Q. & Zhou, W. (2021). Machine learning in the Chinese stock market. *Journal of Financial Economics*.

Lev, B. & Thiagarajan, S. R. (1993). Fundamental information analysis. *Journal of Accounting research,* 31**,** 190-215.

Nti, K. O. a. A., Adebayo and Weyori, Benjamin (2019). Random Forest Based Feature Selection of Macroeconomic Variables for Stock Market Prediction. *American Journal of Applied Sciences 2019*.

Ou, J. A. & Penman, S. H. (1989). Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics,* 11**,** 295-329.

Ou, P. & Wang, H. (2009). Prediction of stock market index movement by ten data mining techniques. *Modern Applied Science,* 3**,** 28-42.

Ratanapakorn, O. & Sharma, S. C. (2007). Dynamic analysis between the US stock returns and the macroeconomic variables. *Applied Financial Economics,* 17**,** 369-377.

Vijh, M., Chandola, D., Tikkiwal, V. A. & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia Computer Science,* 167**,** 599-606.

Yan, X. & Zheng, L. (2017). Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach. *The Review of Financial Studies,* 30**,** 1382-1423.

Zhang, W., Cao, Q. & Schniederjans, M. J. (2004). Neural Network Earnings per Share Forecasting Models: A Comparative Analysis of Alternative Methods. *Decision Sciences,* 35**,** 205-237.

**Appendix**

1. The principle of Random Regression Forest

文本, 信件

描述已自动生成

2. Macroeconomic variables used to predict stock return

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Macroeconomic Variables | Abbreviation | Source | Frequency |
| National Accounts |  | |  |  |  |
| 1 | Gross Domestic Product | | GDP | Wind | Quarterly |
| 2 | Producer Price Index | | PPI | Wind | Monthly |
| 3 | Gross National Savings | | GNS | Wind | Monthly |
| 4 | Gross Domestic Savings | | GDS | Wind | Monthly |
| 5 | Fixed Assets Investment | | FAI | Wind | Monthly |
| 6 | Total Retails Sales of Consumer Goods | | TRSCG | Wind | Monthly |
| 7 | Value of Exports | | VoE | Wind | Monthly |
| 8 | Value of Imports | | VoI | Wind | Monthly |
| 9 | Trade Balance | | TB | Wind | Monthly |
| 10 | M2 | | M2 | Wind | Monthly |
| 11 | Purchasing Managers' Index | | PMI | Wind | Monthly |
| 12 | Total State-owned Assets | | TSOA | Wind | Annual |
| Industry |  | |  |  |  |
| 13 | Business Revenue | | BR | Wind | Monthly |
| 14 | Main Business Income | | MBI | Wind | Monthly |
| 15 | Total profit | | TP | Wind | Monthly |
| 16 | Interest Expense | | IE | Wind | Monthly |
| 17 | Inventory | | I | Wind | Monthly |
| 18 | Accounts Receivable | | AR | Wind | Monthly |
| 19 | Total Assets | | TA | Wind | Monthly |
| 20 | Total Liabilities | | TL | Wind | Monthly |
| 21 | Total Owners’ Equity | | TOE | Wind | Monthly |
| Price Index |  | |  |  |  |
| 22 | Consumer price Index | | CPI | Wind | Monthly |
| 23 | Retail price index | | RPI | Wind | Monthly |
| 24 | Price Index | | PI | Wind | Monthly |
| Interest Rate & Exchange Rate |  | |  |  |  |
| 25 | Deposit Rate | | DR | Wind | Daily |
| 26 | Loan Prime Rate | | LPR | Wind | Daily |
| 27 | Bond Yield | | BY | Wind | Daily |
| 28 | Bill Rate | | BR | Wind | Daily |
| 29 | China Financial Condition Index | | CFCI | Wind | Daily |
| Security |  | |  |  |  |
| 30 | Total Market Value | | TMV | Wind | Daily |
| 31 | Total Capitalization | | TC | Wind | Daily |
| 32 | Stock Turnover | | ST | Wind | Daily |
| 33 | Stock Trading Volume | | STV | Wind | Daily |
| 34 | Turnover of A Shares (SSE) | | ToAS (SSE) | Wind | Daily |
| 35 | Turnover of A Shares (SZSE) | | ToAS (SZSE) | Wind | Daily |
| People’s Living Conditions |  | |  |  |  |
| 36 | Gini Coefficient | | Gini | Wind | Annual |
| 37 | Private Wealth Report (CMB) | | PWR (CMB) | Wind | Annual |
| 38 | Private Wealth Report (CCB) | | PWR (CCB) | Wind | Annual |
| Economic Climates Survey |  | |  |  |  |
| 39 | Macroeconomic Climate Index | | MCI | Wind | Monthly |
| 40 | Business Climate Index | | BCI | Wind | Quarterly |
| 41 | Consumer Confidence Index | | CCI | Wind | Quarterly |
| 42 | Economist Confidence Index | | ECI | Wind | Quarterly |