Empirical Asset Pricing via Machine Learning: Evidence from China

Yue Pei[1, a,*]

[1]*School of Information Engineering, Tianjin University of Commerce, Tianjin, China*
a. 1099061503@qq.com
**corresponding author*

Keywords: random forest, firm characteristics, Chinese stock market

Abstract: This paper analyses the relative importance of 75 individual firm characteristics on expected earnings in China's A-share market, using random forests model in the field of machine learning. Empirically, we found that the stock turnover, industry momentum and RMB trading volume were the three most important characteristics. In addition, according to the importance of variables, the model was reconstructed with the top 23 firm characteristics after ranking, summing and dimension reduction. In comparison with the predicted results of 75 firm characteristics, we found the performance was basically unchanged, but the calculation time was saved by 68.21%. Our empirical results indicate that the random forest algorithm is not only more accurate but also more efficient in predicting the expected returns of stock portfolios based on firm characteristics. This paper contributes to the growing asset pricing study on Chinese stock market.

1. Introduction

Share price is influenced by many factors and is essentially determined by multi-dimensional index system. Because different portfolios will have different returns, the primary problem of quantitative investment is to choose a portfolio. The most fundamental one is to study the characteristics of firms that forecast the cross- section of stock returns, so as to select those firms with the greatest ability to predict future returns. For example, Goyal (2012) recorded dozens of firm characteristics that forecasted the cross-section of stock returns in the United States; Green, Hand, and Zhang (2017) examined the predictive power of 94 firm characteristics.

Traditional forecasting methods often fail due to the large number of stock features, many of which are highly similar. Machine learning, with its emphasis on variable selection and dimensionality reduction, is very suitable for studying such forecasting problems as portfolio building, based on the characteristics of many firms. For example, Rapach (2013) used LASSO regression in the field of machine learning to forecast global stock returns. Hutchinson (1994) used neural network to predict the price of derivatives. Kelly (2015) used the factor of dimensionality reduction method in machine learning and constructed a combination of multiple prediction models to predict the stock return rate. Harvey and Liu (2016) tested the effectiveness of 316 asset pricing factors with the self-help method of machine learning and achieved good prediction results.

In recent years, the random forest algorithm in the field of machine learning has become a new force. It is simple, efficient and widely used. The random forest algorithm can capture the long memory and other nonlinear characteristics of financial market and solve the problem of insufficient explanatory power of the linear pricing model. Gu et al. (2019) investigated 94 characteristics and 74 industry dummy variables of nearly 30,000 stocks in the United States. He used generalized linear model in machine learning, dimensionality reduction, regression tree enhancement, random forest and neural network methods to carry out a comparative analysis in predicting stock returns. The results show that machine learning is the most valuable method for predicting larger and more liquid stock portfolios. And among them, random forest algorithm is an integrated method that integrates many kinds of different trees and has better stability.

After the reform and opening up, China's economy has developed rapidly, and China's stock market has also rapidly developed into the world's second largest stock market. It is becoming more and more important to study the asset pricing law of Chinese market. However, there are few works studying Chinese stock market at present, among which the one by Jiang, Tang, and Zhou (2018) is representative. They created a large set of 75 individual firm characteristics of China's A-shares and compared various "big data research methods". It was found that the principal component analysis (PCA) method and the recently developed partial least squares (PLS) method are the most effective ways to aggregate all the feature information and generate the maximum long-short portfolio returns, compared with the Fama-MacBeth regression method in the linear research method. Studies on Chinese stock market using stochastic forest algorithm are rarer.

Following Jiang, Tang, and Zhou (2018), we created a comprehensive and large set of 75 firm characteristics in the Chinese stock market. On this basis, we studied the relative importance of single covariate to model performance. The results showed that the stock turnover, industry momentum and RMB trading volume were the three most important characteristics. It is suggested that only some firm characteristics can significantly predict the future stock returns of China's stock market. So, we ranked the importance of the covariate characteristics of the model, took the sum of the top priorities, and got the characteristics that make up 95% of the total, which were the top 23 firm characteristics.

Next, we used the 23 characteristics to construct the random forest model and compared it with the model of 75 characteristics. We found that the mean square errors of the two models were 0.228 and 0.0227 respectively, showing no change in performance. Then we calculated the calculation time to construct the model with all the characteristics and the calculation time to construct the model with 23 important characteristics. The comparison showed that the time of the model constructed with 23 important characteristics was saved by 68.21%.

In this paper, random forest algorithm is applied to explore the establishment and dimensionality reduction of firm characteristic model system in Chinese stock market. This paper contributes to the growing asset pricing study on Chinese stock market, which has not been reported in China.

2. Data and Methods

2.1. Data Sources

We obtained the data from Chinese Stock Market & Accounting Research (CSMAR) from July 2000 to December 2016. To ensure data quality, "ST" stocks, delisted stocks, and stocks in financial distress, lack of market liquidity and financial companies were excluded. Seventy-five firm characteristics were also used as variables, which largely followed the thesis of Jiang, Tang, and Zhou (2018). Specific definitions are shown in Table 1.

Table 1: List of Characteristics.

Acronym	Characteristics	Acronym	Characteristics
AM	Assets-to-market	BM	Book-to-market equity
CFP	Cash flow-to-price	DER	Debt-to-equity ratio
DLME	Long term debt-to-market equity	DP	Dividend-to-price ratio
EP	Earnings-to-price	LG	Liability growth
OCFP	Operating cash flow-to-price	PY	Payout yield
Rev1	Reversal	*SG*	Sustainable growth
SMI	Sales growth minus inventory growth	SP	Sales-to-price
TG	Tax growth	ACC	Accruals
PACC	Percent accruals	CAPXG	Capital expenditure growth
dBe	Change in shareholders' equity	dPIA	Changes in PPE and inventory-to-assets
IA	Investment-to-assets	IVC	Inventory change
IVG	Inventory growth	NOA	Net operating assets
ATO	Asset turnover	CFOA	Cash flow over assets
CP	Cash productivity	CTA	Cash-to-assets
CTO	Capital turnover	EBIT	Earnings before interests and taxes
EY	Earnings yield	GM	Gross margins
GP	Gross profitability ratio	NPOP	Net payout over profits
RNA	Return on net operating assets	ROA	Return on assets
ROE	Return on equity	ROIC	Return on invested capital
TBI	Taxable income-to-book income	Z	Z-score
CHMOM	Change in 6-month momentum	INDMOM	Industry momentum
MOM1M	1-month momentum	MOM6M	6-month momentum
MOM12M	12-month momentum	MOM36M	36-month momentum
VOLM	Volume Momentum	VOLT	Volume trend
B_DIM	The Dimson beta	B_DN	Downside beta
BETA	Market beta	BETASQ	Beta squared
B_FF	Fama and French (1992) beta	B_FP	Frazzini and Pedersen (2014) beta
B_HS	Hong and Sraer (2015) beta	IVOL	Idiosyncratic return volatility
ILLIQ	Illiquidity	MAXRET	Maximum daily returns
PRC	Price	PRCDEL	Price delay
RVOL	RMB trading volume	SIZE	Frim size
STD_RVOL	Volatility of RMB trading volume	STD_TURN	Volatility of turnover
RETVOL	Return volatility	TURN	Share turnover
ZEROTRADE	Zero trading days	AGE	Firm age
CFD	Cash flow-to-debt	CR	Current ratio
CRG	Current ratio growth	QR	Quick ratio
QRG	Quick ratio growth	SC	Sales-to-cash
SI	Sales-to-inventory		

2.2. Method Introduction

Because the machine learning algorithm of single classifier is easy to produce overfitting and so on, scholars put forward the idea of integrated learning of multiple classifiers. Boosting and Bagging are the two earliest ensemble learning algorithms. With the development of integrated learning, TinKamHo (1995) put forward the idea of random decision forest in 1995. And in 1998, he put

forward a new integration method of random subspace (TinKamHo, 1998). According to the idea of random subspace, Breiman (2001) proposed the random forest algorithm in 2001. Since then, random forest algorithm has become a representative integrated learning method in the field of machine learning.

Random forest algorithm can use multiple classification or regression trees to classify data. It can also give the importance score for each variable to assess the role of variables in the classification. Each decision tree in the random forest has its own independent sample training set. Each sample training set is extracted from the total samples by Bagging algorithm. They are all randomly selected. The input samples of each tree are not all samples, and will not duplicate samples, and they are not easy to overfit. The final result is the average value of each decision tree, so it has a good fault tolerance, which is suitable for the situation where the stock market has a lot of abnormal disturbance.

Random forest is a kind of grouping classifier. Its nature is a tree type classifier $\{h(x,\beta_k),\ k = 1,2,\ \cdots\}$aggregation, and its middle base classifier $h(x,\beta_k)$ is a kind of non-pruning decision tree constructed with CART algorithm. X is the input vector; β_k is an independent and incidental quantity that determines the growth process of a single tree. Output results are determined by simple multiple voting method. The basic principles are as follows:

A. Random selection of training set: Using bootstrap sampling technique, N training subsets are extracted from the original training set, and the size of each training subset is about two-thirds of the original training set. Each sampling is random sampling and put back sampling.
B. Random structure of forest: The algorithm establishes a decision tree for each training sub-set, and generates N decision trees to form a "forest". Each decision sub-tree is allowed to grow without pruning and management. In the growing process of each subtree, instead of splitting all the M generic parameters with nodes, the specified F (F≤M) genera are randomly selected to split nodes in the best way of F genera, so as to achieve the randomness of node splitting.
C. Node splitting: Node splitting is the kernel step of the algorithm. The generation of branches of each tree is the growth of one of M genera according to the smallest principle (or other evaluation principle) of node impurity (Gini system number). The calculation process of Gini coefficient index is as follows:

a. Calculate the required coefficients of the sample:

$$\text{Gini}(S) = 1 - \sum_{i=1}^{m} P_i^2 \qquad (1)$$

P_i represents the probability of category C_j appearing in sample set S.

b. Calculate the Gini coefficient of each partition:
If S is divided into two subsets, S1 and S2, the Gini coefficient of this division is:

$$Gini_{split}(S) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2) \qquad (2)$$

Among them, |S| is the sample number of the sample set S. $|S_1\ |$, $|S_2\ |$ are the sample numbers of two subsets S1 and S2 respectively.

In the case of node splitting, all divisions of each attribute are sorted according to their Gini coefficients. In the case of node splitting, the attribute with the smallest Gini coefficient is selected as the splitting attribute, and data classification is realized according to its division.

D. Take a large number of votes and divide them into groups. The final outcome of the algorithm is achieved by large majority voting. According to the N decision subtrees constructed with the

mechanism, some test samples will be classified, the fruit of each subtree will be aggregated, and the fruit of the classification with the most votes will be the final output of the algorithm.

3. Evidence and Analysis

We started our empirical study with investigating whether the individual firm characteristics could separately predict the cross-sectional stock returns. We extracted the firm characteristic data from the annual reports of Chinese stocks from July 2000 to December 2016. First, we sorted all stocks with respect to each characteristic depending on its data frequency. We formed 10 decile portfolios at the end of June of year t according to the ranked values of each firm characteristic for the fiscal year ending in year t-1. And then the average return rate of these portfolios in the next month was calculated with the random forest model to evaluate its effectiveness, and the factor estimation of the expected return of the model was obtained.

At the beginning of next month, we sorted firms into 10 decile portfolios according to the random forest factor estimation results of expected returns, held for one month and calculated the monthly portfolio returns. The latest available data would be updated at each end of the month, and the portfolios would be rebalanced on a monthly basis. The return predictability of each characteristic is the difference between the realized return on top and bottom decile portfolios, which is referred to as the long-short portfolio returns. The higher the yield is, the more important it is to predict the return ability of the firm characteristic factors implied in each group.

In this paper, Random forest is implemented with the class "Sklearn. Ensemble. Random Forest Regressor" in scikit-learn. The empirical results show that our strategy gains an average annual yield of 22.32% and the Sharpe ratio is 2.64. The investment benefit is relatively large and stable.

Although short selling is restricted in China, we believe that this does not hinder our next research on predicting the cross-sectional returns of stocks, because random forest also has an important function, which can give the importance score of each variable to evaluate the role of each variable in classification. On this basis, we studied the relative importance of single covariate to model performance. The importance of variables in a given model was normalized to sum to 1, allowing them to explain the relative importance of a particular model. After the calculation of the random forest model, Table 2 shows the degree of influence of each variable on the final result.

Table 2: Importance of Variable Importance=IMP.

Variable	IMP	Variable	IMP	Variable	IMP	Variable	IMP
TURN	0.23	INDMOM	0.15	RVOL	0.11	MOM1M	0.07
Rev1	0.06	SIZE	0.05	ROIC	0.05	MOM6M	0.03
EY	0.03	dBe	0.02	ACC	0.02	PRCDEL	0.02
MOM36M	0.01	RNA	0.01	PRC	0.01	SG	0.01
IA	0.01	NOA	0.01	QR	0.01	EBIT	0.01
PY	0.01	CFOA	0.01	MAXRET	0.01	STD_RVOL	0.01
B_HS	0.01	ILLIQ	0.01	BETA	0.01	BETASQ	0.01
MOM12M	0.0	VOLM	0.0	LG	0.0	TG	0.0
dPIA	0.0	IVC	0.0	IVG	0.	CR	0.0
CRG	0.0	QRG	0.0	CAPXG	0.0	GM	0.0
B_DIM	0.0	B_DN	0.0	B_FF	0.0	B_FP	0.0
VOLT	0.0	AM	0.0	BM	0.0	CFP	0.0
DER	0.0	DLME	0.0	DP	0.0	EP	0.0
OCFP	0.0	SMI	0.0	SP	0.0	PACC	0.0
ATO	0.0	CP	0.0	CTA	0.0	CTO	0.0

NPOP	0.0	TBI	0.0	Z	0.0	AGE	0.0
CFD	0.0	SC	0.0	SI	0.0	ZEROTRADE	0.0
CHMOM	0.0	GP	0.0	ROA	0.0	ROE	0.0
STD_TURN	0.0	RETVOL	0.0	IVOL	0.0		

It can be seen from table 2 that among the 75 characteristics of Chinese stock market, a large part cannot independently predict the future stock returns. Only 28 equal-weighted investment portfolios of variables generate statistically significant hedge returns. Among them, the effects of the stock turnover, industry momentum and RMB trading volume on the model accounted for 0.23, 0.15, and 0.11, respectively. Other than that, no other feature has a greater impact than 0.1. This suggests that only some firm characteristics can significantly predict the future stock returns of China's stock market.

It can be seen from the above that some characteristic variables do not make the model more effective, so next we ranked the importance of the covariate features of the random forest model, ranking the highest in the front and the lowest in the last. Then dimensionality reduction was carried out on the features, and the most important features were selected as the indicators of the following training to improve the performance and effect of the model. We reconstructed the random forest model with the top 23 characteristics, of which the cumulative importance exceeded 95%, and compared the results with the 75 characteristics mentioned above. It can be seen from Table 3 that the mean square errors of the two models are 0.228 and 0.0227, respectively. It shows that the performance does not change much. This is precisely because the selection of characteristic variables in the random forest algorithm is highly random.

At the same time, we calculated the calculation time of using all characteristics construction models and 23 important characteristics construction models. The comparison shows that the calculation time of the model constructed with 23 important characteristics is saved by 68.21%.

To sum up, the random forest algorithm in the field of machine learning has higher accuracy and lower time and labor costs in predicting the future returns of stock portfolios based on firm characteristics.

Table 3: Trade-off.

Features	MSE	Run-time(s)
All (75)	0.0227	368.39
Reduced (23)	0.0228	117.13

4. Conclusions

This paper adopts the random forest model in the field of machine learning to study the relative importance of a single covariable to the model performance and analyzes the relative importance of the impact of 75 firm characteristics on expected returns in China's A-share market. Empirically, we found that the stock turnover, industry momentum and RMB trading volume were the three most important characteristics. And according to the concept of variable importance ranking, based on the variable importance, summation, dimension reduction, we can get those top 23 firm characteristics that account for 95% of the total characteristics. The model is reconstructed with the top 23 firm characteristics after ranking, summing and dimension reduction. In comparison with the predicted results of 75 firm characteristics, we found the performance was basically unchanged, but the calculation time was saved by 68.21%.

References

[1] *Goyal, A. (2012). Empirical Cross-Sectional Asset Pricing: A Survey. Financial Markets and Portfolio Management, 26(1), 3-38.*

[2] *Green, J., Hand, J.R. and Zhang, X.F. (2017). The Characteristics that Provide Independent Information about Average US Monthly Stock Returns. The Review of Financial Studies, 30(12), 4389-4436.*

[3] *Rapach, E., Strauss, K. and Zhou, G. (2013). International Stock Return Predictability: What is the Role of the United States. Journal of Finance, 68(4), 1633-1662.*

[4] *Hutchinson, M.L., Andrew, P. and Tomaso, A. (1994). Nonparametric Approach to Pricing and Hedging Derivative Securities via Learning Networks. Journal of Finance, 49(3), 851-889.*

[5] *Kelly, B. and Pruitt, S. (2015). The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors. Journal of Econometrics, 186(2),294-316.*

[6] *Harvey, C.R., Liu, Y. and Zhu, H. (2016). ... and the Cross-Section of Expected Returns. The Review of Financial Studies, 29(1), 5-68.*

[7] *Gu, S., Kelly, B.T. and Xiu, D. (2019). Empirical Asset Pricing via Machine Learning. Social Science Electronic Publishing.*

[8] *Jiang, F., Tang, G. And Zhou, G.(2018). Firm Characteristics and Chinese Stocks. Social Science Electronic Publishing.*

[9] *TinKamHo. (1995). RandomDecision Forest. In Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montreal, Canada,8,278-282.*

[10] *TinKamHo. (1998). The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832-844.*

[11] *BreimanL. (2001). Random Forests. Machine Learning, 45(1), 5-32.*