Heliyon 5 (2019) e02310

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.heliyon.com

Stock selection with random forest: An exploitation of excess return in the Chinese stock market

Zheng Tan [a,*], Ziqin Yan [a], Guangwei Zhu [b]

[a] Xiyuan Hedge Fund, 388 Yizhou Road, Chengdu, Sichuan, PR China
[b] Institute of Chinese Financial Studies, Southwestern University of Finance and Economics, 555 Liutai Avenue, Wenjiang District, Chengdu, Sichuan, PR China

ARTICLE INFO

Keywords:
Computer science
Economics
Excess return
Random forests
Stock selection
Machine learning
Finance

ABSTRACT

In recent years, a variety of research fields, including finance, have begun to place great emphasis on machine learning techniques because they exhibit broad abilities to simulate more complicated problems. In contrast to the traditional linear regression scheme that is usually used to describe the relationship between the stock forward return and company characteristics, the field of finance has experienced the rapid development of tree-based algorithms and neural network paradigms when illustrating complex stock dynamics. These nonlinear methods have proved to be effective in predicting stock prices and selecting stocks that can outperform the general market. This article implements and evaluates the robustness of the random forest (RF) model in the context of the stock selection strategy. The model is trained for stocks in the Chinese stock market, and two types of feature spaces, fundamental/technical feature space and pure momentum feature space, are adopted to forecast the price trend in the long run and the short run, respectively. It is evidenced that both feature paradigms have led to remarkable excess returns during the past five out-of-sample period years, with the Sharpe ratios calculated to be 2.75 and 5 for the portfolio net value of the multi-factor space strategy and momentum space strategy, respectively. Although the excess return has weakened in recent years with respect to the multi-factor strategy, our findings point to a less efficient market that is far from equilibrium.

1. Introduction

The investigation of the primary drivers regarding stock returns has been of great interest for many decades. As seen in the classical financial theories, such as CAPM and various multivariate models (Fama and French, 1993, 2015), stock returns are attributed to the underlying fundamentals, including systematic risk, market cap, book to market ratio, etc., in a linear manner. The corresponding linear regression scheme, combined with the extensively developed factors covering various technical and fundamental aspects (Zhu et al., 2011), constitutes the primary workhorse for financial modelling in the academic and industry sectors. It is, however, uncertain whether the market is linear and whether the return is linearly regressible or purely the result of market anomalies (Zhu et al., 2012). The recent boom in machine learning techniques offers an alternative paradigm for illustrating the relationships between the stock price forward process and its relevant company features, thereby providing a higher degree of model diversification compared to traditional approaches. It is proved that by using powerful model classes, such as artificial neural networks (ANN) (Khashei and Bijari, 2010; Alberg and Lipton, 2017; Belciug and Sandita, 2017), decision trees (DT) (Sorensen et al., 2000; Andriyashin et al., 2008; Zhu et al., 2011, 2012), deep neural networks (DNN) (Chong et al., 2017), gradient-boosted trees (GBDT), random forests (RF) (Krauss et al., 2017), etc., the classification and prediction efficiency of stocks are significantly enhanced.

Generally, stock selection plays an important role in portfolio management. In contrast to the academic study where there is greater emphasis on the predictions of stock returns (Enke and Thawornwong, 2005; Khashei and Bijari, 2010; Hassan et al., 2007; Guresen et al., 2011; Jia et al., 2012; Ticknor, 2013; Babu and Reddy, 2014; Rather et al., 2015), investment managers focus primarily on the screening and categorization of stocks that can outperform the market cross-sectional median. The stock selection criteria, nevertheless, are complex and by no means comply with scientific standards. However, newly developed advanced algorithms, such as deep neural nets and tree-based models, have shed light on stock selection efficiency compared to the conventional factor weighted ranking system, thus leading to remarkable trading performance. A brief literature review is given below.

* Corresponding author.
E-mail address: tanz@xiyuanhedgefund.com (Z. Tan).

https://doi.org/10.1016/j.heliyon.2019.e02310
Received 23 December 2018; Received in revised form 2 June 2019; Accepted 12 August 2019

Andriyashin et al. (2008) applied decision trees to a ternary classification of stocks in DAX constituents by training the model on both the fundamental and technical variables. They achieved an annual return of 25.55% and a Sharpe of 1.59 in their weekly based stock selection strategy, thus notably outperforming the general market. Similar tree-based models were also employed (Sorensen et al., 2000; Zhu et al., 2012) to investigate the stock classification efficiency based on different feature spaces, where the decision tree framework was claimed to be superior to the linear weighting approach in providing risk-diversified portfolios and in illustrating higher-order relationships between stock returns and underlying variables.

Krauss et al. (2017) employed three state-of-the-art machine learning techniques, including DNN, GBDT and RF, as well as a combination of the three, referred to as ensembles, to the S&P 500 constituent stocks. A binary classification of stocks based on a one-day-forward excess return was predicted when the corresponding model was trained on the momentum feature space. When the trading strategy was executed in the context of statistical arbitrage, a 0.23% and 0.25% out-of-sample daily return for the RF and the equal-weighted ensemble model, respectively, was produced. Their empirical findings were promising, indicating that a sustainable profit opportunity in the short run is exploitable by means of machine learning, even in the case of a mature market.

Although there exists a vast volume of literature on the forecasting of the stock dynamics via machine learning, few articles illustrate the structural relationship between stock selection and its underlying factors with the aid of machine learning, especially in the case of emerging markets. While the Chinese stock market has been developed for decades, how the relevant fundamental and technical features relate to stock returns is rarely understood. Thus, it is of great value to establish such a framework for systematic investigation. To make predictions on stocks that belong to the first class, we employ random forest, which is understood as an uncorrelated decision tree ensemble that gives rise to a probability matrix for the classification of a sample. The model is trained on the fundamental/technical feature space and a pure momentum feature space, with the stocks accordingly being classified by long-term and short-term excess return, respectively. The trading strategy is implemented in a rolled training and trading scheme, which is detailed in the following sections. The influence of the number of trees in RF, the number of sample classes, the length of the training period and the length of the rolling period on out-of-sample performances is carefully tested, and the resultant trading performances derived from fundamental/technical feature space and momentum feature space are compared. Our aim is to properly connect the short-term and long-term excess returns with the appropriate feature space and exploit the profit opportunity in different time scales within an emerging market. The employed machine learning algorithm helps financial practitioners to improve their stock selection efficiency and, accordingly, bridge the gap between the complex stock pricing mechanism and portfolio management.

The paper is organized as follows. The second section is dedicated to the data sources and software used in the simulations. This is then followed by the methodology in which the rolled training and the trading scheme as well as the two types of feature spaces are explained in detail. Section 4 presents the results and discusses the main findings of our study, and a conclusion then brings the paper to an end.

2. Calculation

2.1. Data

Our study covers all listed firms on the Chinese stock market, while the CSI 500 index is selected to be the benchmark, as it reflects the overall performance of small-mid cap A-shares and tends to exhibit less industry bias. The stock data are acquired on a daily basis from the Wind financial database, which contains price, volume, and fundamentals for listed companies. The stocks that are traded in less than one year are eliminated to avoid the stock premium period shortly after the IPO.

2.2. Software

Preprocessing and data handling are conducted using MATLAB (MathWorks, Natick, MA, UNITED STATES), which provides a multi-paradigm numerical computing environment. The RF model is implemented in the context of scikit-learn (Pedregosa et al., 2011), which is a Python module that integrates a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems. The communication between MATLAB and scikit-learn is achieved via the python API within the MATLAB framework.

3. Methodology

The data sets are divided into a number of training parts and trading parts on a rolling basis, and the out-of-sample trading period covers February 8, 2013 to August 8, 2017. Features are generated for the respective training periods to capture the characteristics of stocks, whereas the RF model is deployed to train the training data sets in accordance with a certain classification criterion and to make predictions in the subsequent trading period.

3.1. Rolled training and trading scheme

The nearly five-year out-of-sample trading period is split into 20 trading sub-periods, with each consisting of 60 trading days (approximately three months). In this way, every trading sub-period is non-overlapping. Each corresponding training data set prior to the trading sub-period contains 252 trading days (approximately one year), where the RF model is applied to train the sample based on the classification as denoted to each stock. Herein, we prefer a classification instead of a regression problem, as the literature suggests that the former performs better than the latter in predicting financial market data (Leung et al., 2000; Enke and Thawornwong, 2005). In the training data set, stocks are divided into N classes based on the forward excess returns of each stock. The trained RF model is then used in the subsequent trading period to predict the probability for each stock that belongs to the first class, i.e., the category with the largest excess return, where the first 20 stocks with the highest probability are selected at each stock ranking date. The selected stocks constituting the portfolio are held for a certain period, and the portfolio constituents are then renewed based on the new probability ranking. All back-tests performed and discussed bear a transaction cost preset of 0.16%.

We design two types of feature spaces to examine the profitability of stock selection in the Chinese market using a factor modal framework, where those fundamental and technical factors likely to result in long-term excess returns and a pure momentum feature space with different time lags are both considered. A 20 trading-day holding period is assigned to the fundamental/technical feature space trading strategy, and a holding period of two trading days is assigned for the momentum feature space strategy.

3.2. Feature space generation

For each training period, we generate the feature space (input) and the response variable (output) as follows.

Input: For the model with fundamental/technical feature space, the input is a u × v matrix, where u is the sample number and is calculated as the total number of tradable stocks multiplied by the number of trading days in the training period, and v is the number of fundamental/technical features. For the model with momentum feature space, the input is also a u × v matrix similar to the matrix described above, where u is the number of samples and v denotes the number of momentum features. More details of fundamental/technical features and momentum features are described in sections 3.2.1 and 3.2.2, respectively.

Output: Let $P^S = (P_t^s)_{t\in T}$ denote the close price process of stock s, with

s$\varepsilon\{1, 2, \ldots, n\}$. At time t, the forward return of each stock $R^s_{t,m}$ and the CSI index $R^I_{t,m}$ over the subsequent m holding days can then be calculated. The excess return $ER^S = (ER^s_t)_{t\in T}$ is the difference between the stock return and the index return, as described in the equation. Finally, we equally split all stocks ranked with excess returns in descending order into N classes, which are the outputs of the training model.

$$R^s_{t,m} = \frac{P^s_{t+m}}{P^s_t} - 1 \tag{1}$$

$$R^I_{t,m} = \frac{P^I_{t+m}}{P^I_t} - 1 \tag{2}$$

$$ER^s_{t,m} = R^s_{t,m} - R^I_{t,m} \tag{3}$$

3.2.1. Fundamental and technical feature space

The two most common methods of stock valuation are known as fundamental and technical analyses. Fundamentalists consider that stock price would converge in the long run based on its underlying company characteristics, whereas technical analysts are prone to make predictions on stock trends via price and volume indicators, as they maintain that all information is already reflected in the price of a security. Both methods are widely used in the finance industry and separately capture different pricing mechanisms of the stock market. In this article, we design our proprietary fundamental and technical factors as related to the stock future return, and factor importance is examined in the framework of random forest.

It is well known that stock returns are related to company fundamentals in a variety of ways. Numerous literatures have interpreted stock cross-sectional returns in terms of a univariate or multivariate linear model, where the forward return is believed to be driven by book to price ratio (Brennan et a, 1997), market cap (Fama and French, 1993), earnings to price ratio (Basu, 1983), profitability, investment (Fama and French, 2015), etc. Meanwhile, many alternative nonlinear models, such as decision trees, artificial neural networks and random forest models, are deployed to illustrate the structural relationship.

The fundamental features used in the model training are listed in Table 1. These features are claimed to have considerable effects on stock price predictions in several markets, as evidenced in the references presented in Table 1. However, their effectiveness in emerging markets, such as the Chinese market, has not been systematically tested. Thus, we employ these factors to construct the fundamental feature space, which is further combined with the technical feature space to explore the classification efficiency of stocks in the Chinese market.

Table 1
Fundamental features. All factors are acquired from the Wind database. Some factors, such as EP, BP, ROE, etc., are strongly correlated with the average stock returns in mature stock markets, as indicated in the listed references therein, while others are believed to have significant explanatory power in practical stock investment.

Factors	Name	Description
EP (Basu, 1983)	The ratio of earnings to price	Determine whether shares are correctly valued in relation to one another
BP (Brennan et al., 1997)	The ratio of book to price	Used to compare a company's current market value to its book value
SP	The ratio of sales to price	Used to determine the value of a stock relative to its past business performance
Net profits yoy	The growth rate of net profits year on year	Used to estimate the company's business prospect
Business income yoy	The growth rate of business income year on year	Used to estimate the company's growth and development capabilities
ROA	The return on assets	Reflects by percentage how profitable a company's assets are in generating revenue
ROE (Chen et al., 2011)	The return on equity	Used to measure how well a company uses investments to generate earnings growth
Market cap (Fama and French, 1993)	Market capitalization calculated as price times shares outstanding	Reflects how much money is raised and the size of listed companies

With respect to the technical feature space, we design the proprietary technical factors that may influence the price dynamics in the preset holding period. All features are built upon the data of price, volume and turnover (calculated as the ratio of daily volume to tradable shares), and appropriate time lags are considered to be consistent with the length of the stock holding period. More details can be found in Table 2.

Note that the movavg and movstd in Table 2 are the average and standard deviation of a time series, respectively, over the past m trading days. Furthermore, all designed factors are systematically examined for their correlations with the stock forward returns, and they are able, to some extent, to establish a sensible stock classification. Further details of examinations of individual factors are not included in this article. Rather, the structural relationship between the stock selection and all the designed factors is explored in the framework of RF.

3.2.2. Momentum feature space

In previous investigations, the effect of price momentum with respect to stock returns has been extensively studied (Jegadeesh and Titman, 1993; Carhart, 1997; Hou et al., 2011; Moskowitz et al., 2012). Meanwhile, it is widely considered as one of the major factors leading to sensible stock classification and trading strategies. Recently, more advanced algorithms (Takeuchi and Lee, 2013; Krauss et al., 2017), such as deep neural networks, gradient boosted decision trees and random forests, are applied to enhance the strategy performance by intricately reconstructing the momentum feature space, thus significantly improving the corresponding annualized return, as seen in the literature (Krauss et al., 2017). Herein, we follow the design in Krauss et al. (2017) by building the momentum space as in Eq. (4).

$$\mathrm{Mom}^s_{t,m} = \frac{P^s_t}{P^s_{t-m}} \tag{4}$$

where m $\in \{\{1, \ldots, 20\} \cup \{40, 60, \ldots, 240\}\}$. Different time lags are considered with distinct resolutions, thus accounting for the price process in the former one trading year. We first focus on the stock return of the last 20 days, then switch to the preceding 11 months. In all, 31 features are generated in the momentum space, and we specifically test the classification efficiency of the Chinese stock market in the short forward period, which is assigned to be two trading days in this article, by means of the RF model.

3.3. Random forest model

We construct our random forest model by following the conventional approach given in Breiman (2001). No modification is made to the algorithm, as it is believed that the original RF can have enough capacity to handle large number of variables in datasets and give rise to unbiased estimate for real world classification problems (Ahmad et al., 2018), including finance.

In principle, the random forest consists of many deep but uncorrelated decision trees built upon different samples of the data (Breiman, 2001). The process of constructing a random forest is simple. For each decision tree, we first randomly generate a subset as a sample from the original dataset. Then, we grow a decision tree with this sample to its maximum depth of J_{RF}. Meanwhile, m_{RF} features used on each split are selected at random from p features. After repeating the procedure numerous times with the original dataset, n_{RF} decision trees are generated. The final output is an ensemble of all decision trees, and the classification is conducted via a majority vote. The computational complexity

Table 2
Technical features.

Factors	Description	Formula
turnover_20, turnover_40, turnover_60, turnover_120, turnover_240	Refers to the moving average of the turnover over a certain period	$movavg(turnover, m) m \in \{20, 40, 60, 120, 240\}$
close_0/close_9, close_0/close_19, close_0/close_39, close_0/close_59, close_0/close_119	Refers to the momentum with different time lags and can be used to help identify the trend of the price process	$\frac{P_t}{P_{t-m}} m \in \{9, 19, 39, 59, 119\}$
close_19/close_0, close_39/close_0, close_59/close_0, close_119/close_0	Refers to the reversal of momentum	$\frac{P_{t-m}}{P_t} m \in \{19, 39, 59, 119\}$
adjusted_close_0/close_59, adjusted_close_0/close_119	Refers to the momentum with different time lags, excluding the most recent month	$\frac{P_{t-19}}{P_{t-m}} m \in \{59, 119\}$
vol10/vol20, vol10/vol40, vol10/vol60, vol20/vol60, vol40/vol60	Refers to a rate of acceleration of a stock's volume and can be used to help identify trend lines of volume	$\frac{movavg(volume, m_1)}{movavg(volume, m_2)} m_1 \in \{10, 10, 10, 20, 20, 40\}\ m_2 \in \{20, 40, 60, 40, 60, 60\}$
volatility_10, volatility_20, volatility_40, volatility_60, volatility_120	Refers to the volatility over the past m trading days as calculated by the standard deviation of daily returns	$movstd(daily\_R, m) m \in \{10, 20, 40, 60, 120\}$
std(volume_10), std(volume_20), std(volume_40), std(volume_60), std(volume_120)	Refers to the standard deviation of trading volume time series over the past m trading days	$movstd(volume, m) m \in \{10, 20, 40, 60, 120\}$

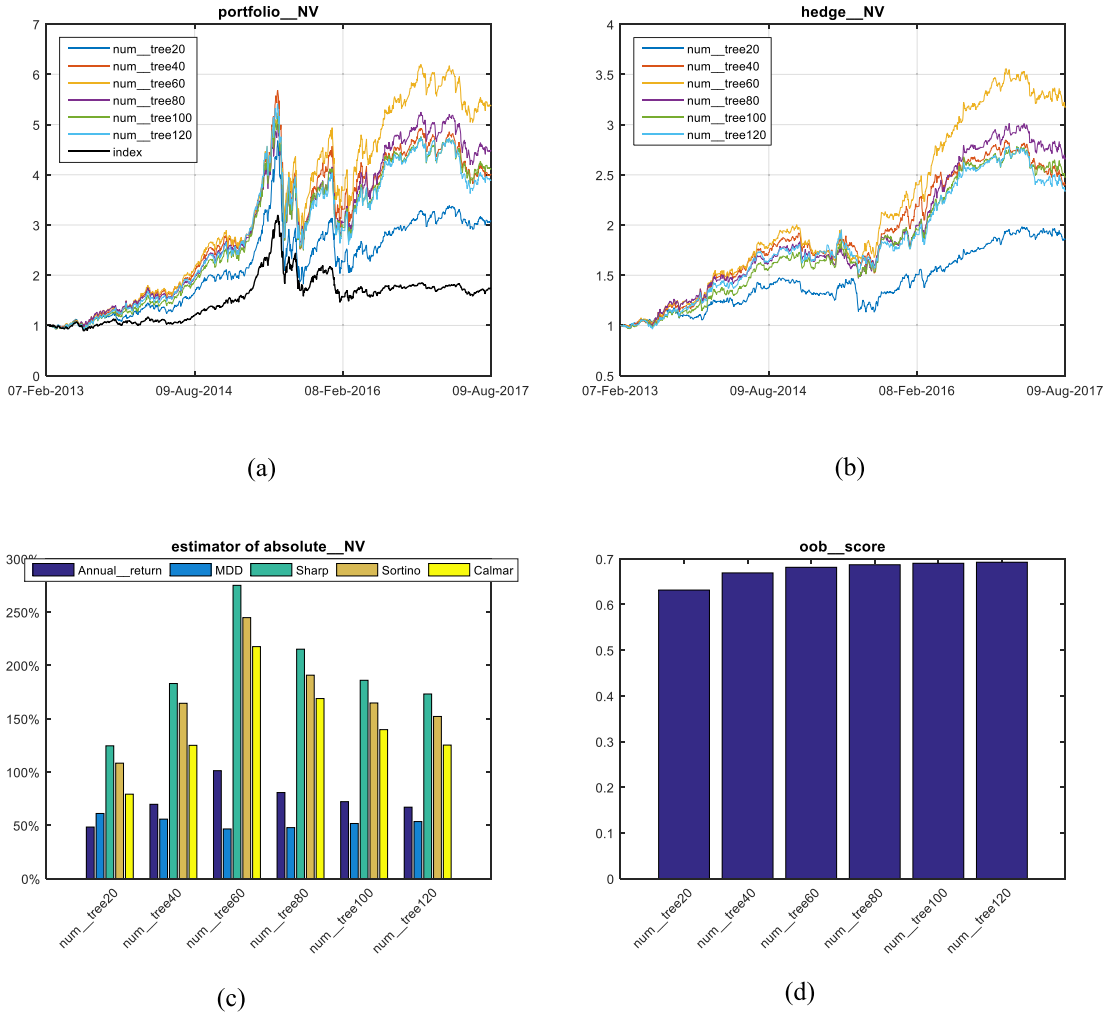


Fig. 1. The dependence of the strategy performance on the number of trees, which is set to be in {20, 40, 60, 80, 100, 120}. (a) Strategy performance, as represented by the net asset value portfolio, the dependence on the number of trees; (b) Hedged net asset value dependence on the number of trees; (c) Estimators of the net asset value portfolio and their dependence on the number of trees; (d) Oob score dependence on the number of trees.

can be simply estimated as $O(n_{RF}(p * n_{ins} * log n_{ins}))$, where n_{ins} represents the number of instances in the training datasets. Three parameters must be tuned to check the robustness of the RF on classification, i.e., the number of trees n_{RF}, the maximum depth J_{RF} and the number of features m_{RF} of each split. We set the maximum depth J_{RF} to be unlimited so that the nodes are expanded until all leaves are pure or until all leaves contain less than two samples. Regarding the feature subsampling, we typically choose $m_{RF} = \sqrt{p}$ (James et al., 2013). The influence of the number of trees on the classification accuracy and the out-of-sample performance is then systematically investigated.

Several extensions can be made to the current RF algorithm, such as the combined random forest-neural network machine (Wang et al., 2018), and the RF promoted multi-label learning machine (Yangming and Guoping, 2018; Chen et al., 2018). These methods are attractive in giving rise to broader model diversification, and could be used to further enhance our stock selection efficiency in the future.

4. Results and discussions

4.1. Dependence of strategy performance on model parameters

To inspect the classification effectiveness of the RF model on the Chinese stock market and the validity of the strategy in exploiting excess return, we carefully test the degree to which the strategy performance relies on modal parameters. The portfolio net asset value (NV) and hedged NV (note that the hedged NV is calculated using the accumulated excess return) are extracted depending on different tree numbers, sample class numbers, training period and rolling period. The annual return, maximal drawdown (MDD), Sharpe ratio, Sortino ratio and Calmar ratio are calculated for the NV portfolio, and the mean out-of-bag (oob) score (Breiman, 1996, 2001), as computed by the mean value of the oob score for each in-sample training, is used to evaluate the training accuracy. It is important to note that the parameter dependency analysis is performed by varying one parameter, while others remain constant in the following four sub-sections.

4.1.1. The number of trees

From Fig. 1(a) and (b), it is observed that the impact of tree number is significant, as both the NV portfolio and the hedged NV exhibit very different profiles due to different tree numbers. The portfolio reveals outstanding profitability in the market oscillating period, such as the time between 2013 and 2014 and the year after 2016. During the bull and bear market periods approximately 2015, the strategy exhibits a systematic rise and drawdown by following the index trend. The corresponding hedged NV profile illustrates a deteriorated excess return during the bullish and bearish periods and a steady excess profit in the market oscillating period, thus demonstrating a consistent outperformance compared to the general market.

Fig. 1(c) indicates that the portfolios give rise to remarkable annual returns that range from 50% to 100%, with similar maximal drawdowns that are irrespective of the tree number. The Sharpe ratio can reach 2.75 when the tree number is set to be 60, and similarly, the Sortino and

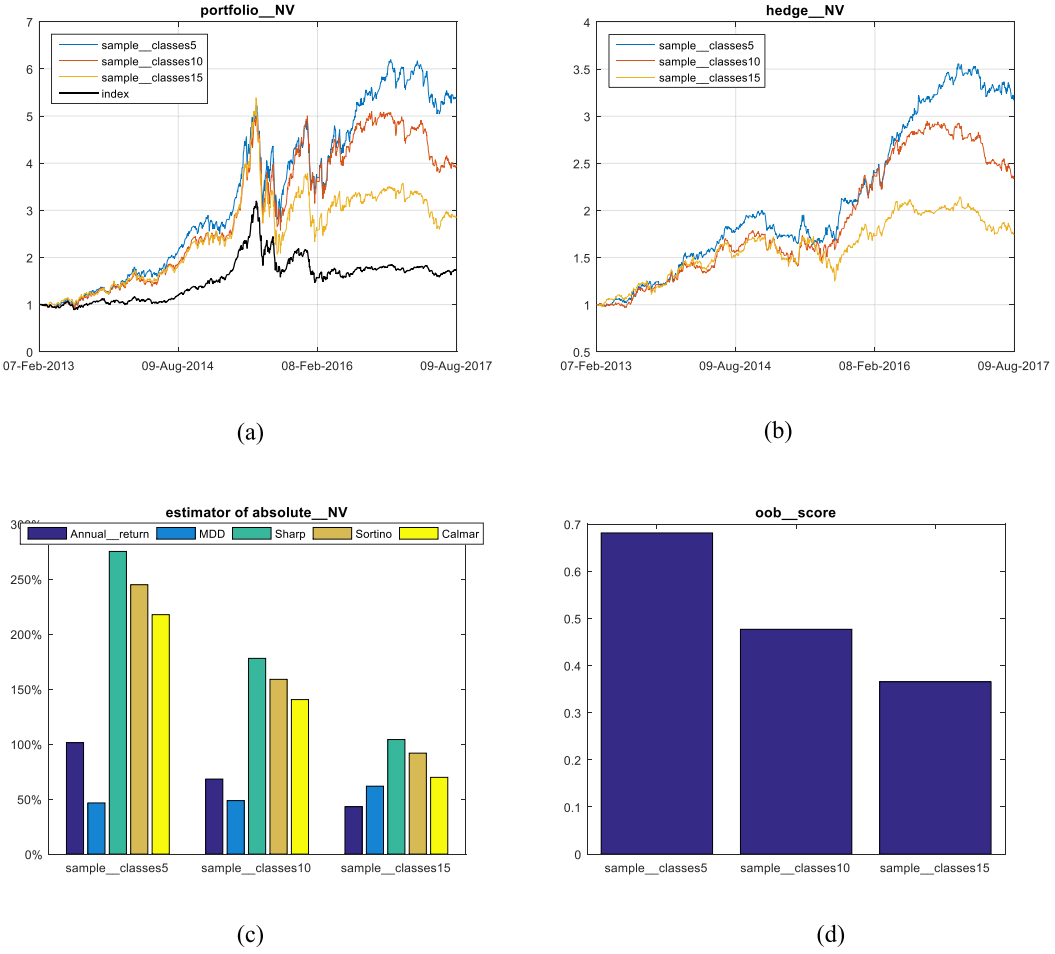


Fig. 2. The dependence of the strategy performance on the number of sample classes, which is set to be in {5, 10, 15}. (a) Portfolio NV dependence on the number of sample classes; (b) Hedged NV dependence on the number of sample classes; (c) Estimators of the NV portfolio and their dependence on the number of sample classes; (d) Oob score dependence on the number of sample classes.

Calmar ratio can also find their maxima. As seen in Fig. 1(d), the in-sample oob score exhibits a steady increase with the tree number, while the trend levels off as the tree number approaches 120. It is further noted that larger tree numbers lead to higher accuracy for the in-sample training, but by no means do they imply a better out-of-sample strategy performance.

4.1.2. The number of sample classes

As evidenced in Fig. 2(a) and (b), the out-of-sample performance portfolio and the hedged NV deteriorate with the increased sample class number, especially for the years 2016 and 2017, where larger numbers of sample classes lead to diminished excess returns. The estimators, namely, annual return, Sharpe, Sortino and Calmar, of the NV portfolio, which are presented in Fig. 2(c), demonstrate a similar declining trend with the sample class number, as revealed in the NV profiles. The in-sample oob score decreases with the sample class number as well, suggesting a less accurate classification of stocks when the class number increases.

4.1.3. The training period

The training period is important in the strategy design since it determines how many samples should be considered in training the model to be used for prediction in the subsequent trading period. Fig. 3(a), (b) and (c) indicate the strategy with different training periods can have notable annual returns and Sharpe, with consistent excess returns throughout the out-of-sample period. Training periods that are too short or too long produce a decreased outperformance, suggesting that a suitable training period length is required for the model to create an efficient stock classification within a certain period of time. A training period of 125 days reaches a Sharpe of 2.25 and an annual return of 84.96% in the strategy performance. The oob score, however, is almost indistinguishable among the different training periods, demonstrating a stabilized classification accuracy with respect to the number of training samples.

4.1.4. The rolling period

The rolling period represents the frequency by which we renew the model and indicates the length of time the model can be effective for predictions in the following period. As presented in Fig. 4, by setting the rolling period from 20 to 80 trading days, the portfolio NV and hedged NV are nearly unaffected regarding this parameter, This also holds for the NV estimators and the oob score.

4.2. Analysis of feature importance

The fundamental/technical feature space could lead to satisfactory classification efficiency and produce a steady strategy outperformance for a 20-day holding period. It is, therefore, interesting to analyse the feature relevance and explanatory power of stock classification. Fig. 5 presents the feature weight distribution for the 40 proprietarily designed factors, where the variable importance is determined by computing the relative influence of each variable, i.e., by assessing whether a particular variable is used during splitting when growing trees and by determining

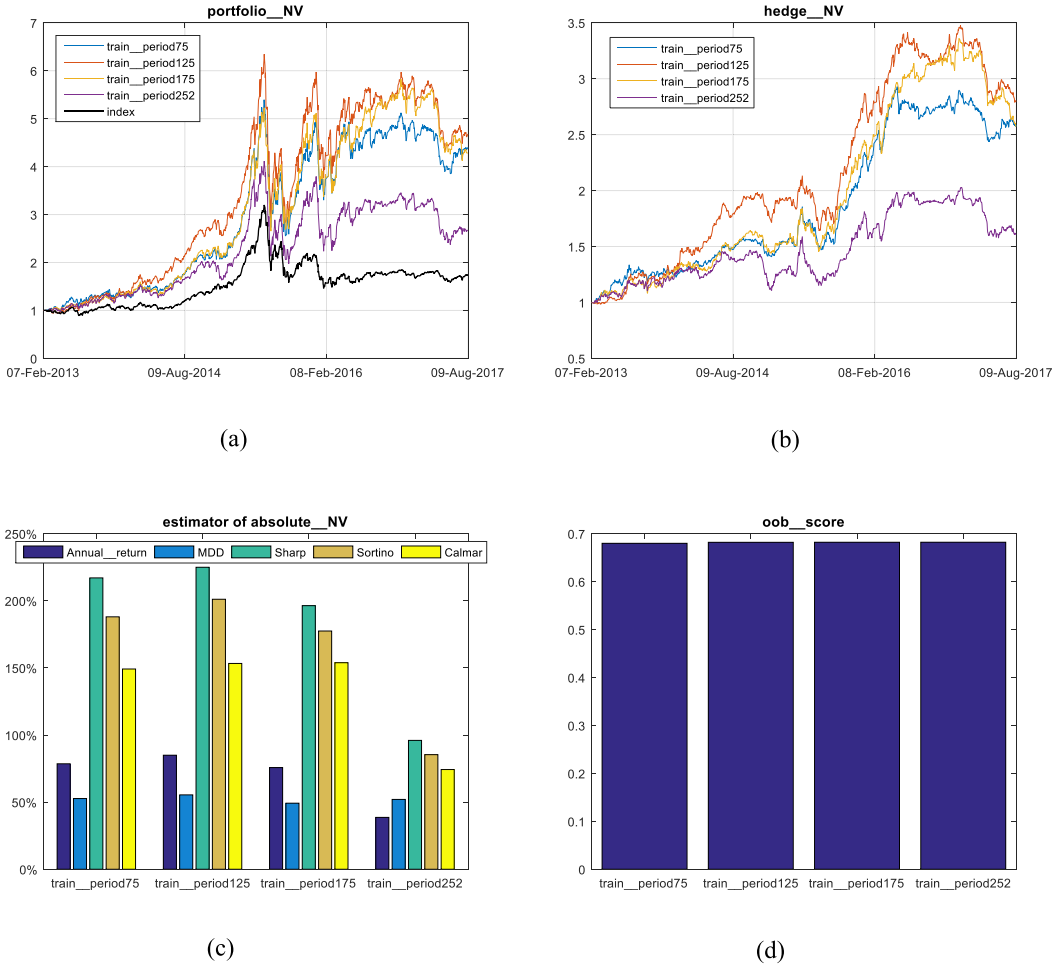


Fig. 3. Dependence of strategy performance on the training period, which is set to be in {75, 125, 175, 252}. (a) Portfolio NV dependence on the training period; (b) Hedged NV dependence on the training period; (c) Estimators of portfolio NV and their dependence on the training period; (d) Oob score dependence on the training period.

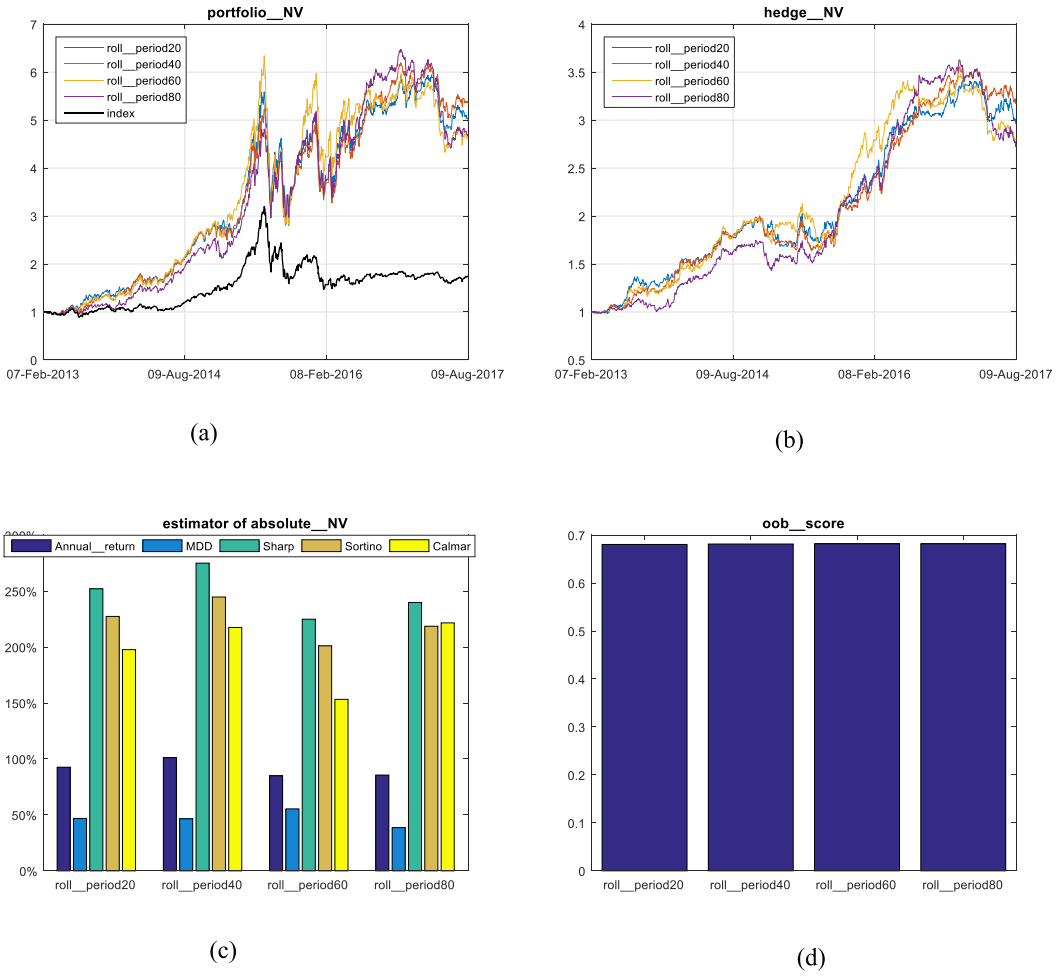


Fig. 4. Dependence of strategy performance on the rolling period, which is set to be in {20, 40, 60, 80}. (a) Portfolio NV dependence on the rolling period; (b) Hedged NV dependence on the rolling period; (c) Estimators of portfolio NV and their dependence on the rolling period; (d) Oob score dependence on the rolling period.

the degree to which the loss function improves as a result, on average, across all trees. We extract feature importance from every training data set and depict the average feature importance distribution within Fig. 5, which reveals that the modal parameters in the relevant training and trading scheme are properly set to guarantee a reliable outperformance.

It is evident that the market capitalization is the most prominent factor as it exhibits a weight approaching 0.04. The three fundamental factors, i.e., EP, BP and SP, demonstrate relatively higher importance in determining the stock forward direction. Regarding the technical factors, namely, turnover_240, volatility_120 and std (volume_120), they give rise to elevated weights compared to short-term correspondents, probably suggesting that a long-term price volume process characteristic would mainly account for the long-term excess return in the future. Our findings imply there is a significant relationship between the long-run profit and fundamental/technical factors that could be valuable in creating sensible strategies.

4.3. Performance comparison between multi-factor space and momentum space

From the analyses in the preceding sections, it is evident that by using the 40 fundamental/technical factors, long-term excess return is extractable in the Chinese stock market. As presented in Fig. 6, a steady outperformance for the multi-factor space strategy over the past five years is observed. By employing the parameters provided in Table 3, the annual return of the portfolio NV can reach 101.21%, while the MDD is calculated to be 46.51%, which is attributed to the systematic drawdown in 2015. Furthermore, the Sharpe ratio of the strategy is remarkable in that it approaches 2.75, whereas the mean oob score is approximately 0.7, which demonstrates satisfactory accuracy for the in-sample classification.

As described in Krauss et al. (2017), by designing a completely different feature space, such as a momentum series, profit and efficient classifications of stocks are expectable even in a mature market. Hence, it is appealing to employ a similar feature space to examine the strategy profitability in the emerging market. We follow the feature space paradigm in Krauss et al. (2017) by generating a price momentum series with different time lags in the preceding trading year. The strategy is implemented based on a two-day holding period, using the same modal parameters as in the multi-factor space strategy. The resultant portfolio NV and hedged NV performances are exhibited in Fig. 6. Compared to the multi-factor space strategy, the momentum space strategy exhibits an even more prominent outperformance, especially during the market oscillating period. The portfolio annual return reaches 185.56%, while the MDD is comparable with that of the multi-factor strategy. A Sharpe ratio of 5 is achieved, indicating that significant excess returns with diminished volatility are exploitable in the short forward period. The mean oob score, however, is relatively small in the momentum space strategy, pointing to reduced accuracy in in-sample trainings. The fact that the in-sample classification in terms of momentum series features is

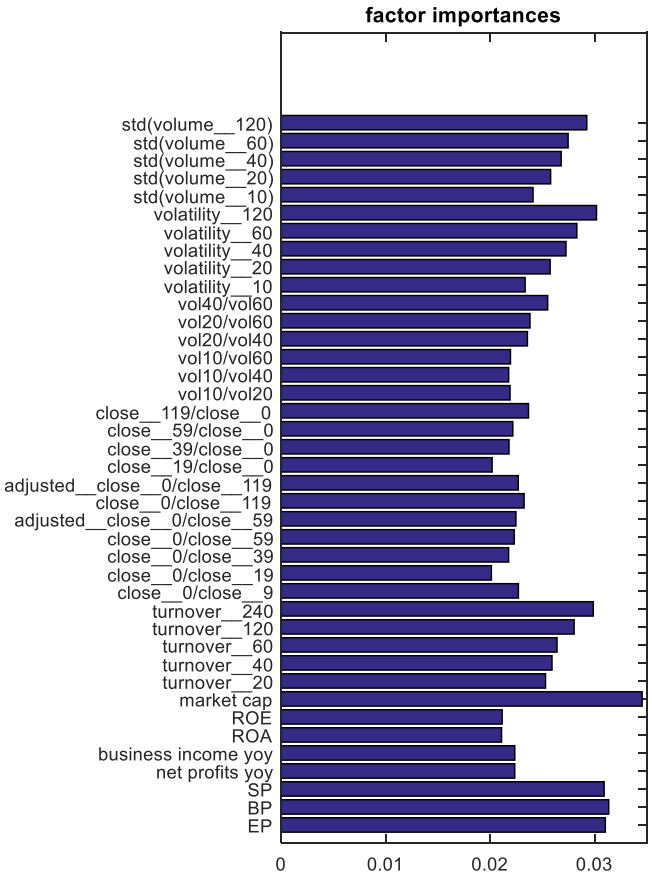


Fig. 5. Weight distribution of 40 factors used in RF model.

less efficient contradicts the enhanced out-of-sample strategy performance, implying a possible over-fit in the case of the multi-factor space strategy.

4.4. Performance comparison between different machine learning methods

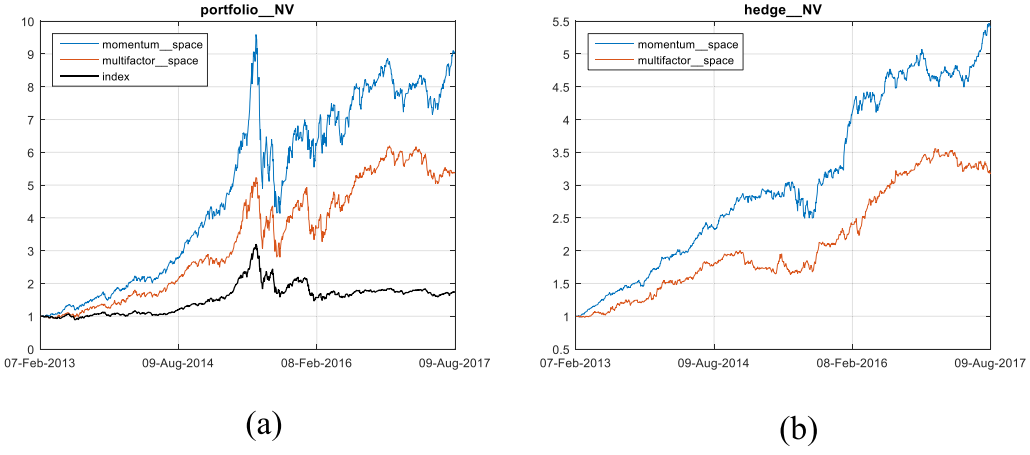
In order to test the feasibility for profit exploitation with other learning machines, we conduct data sets training and probability forecasting by using Logistic Regression (LR), Deep Neural Network (DNN) and RF, respectively. The DNN machine is set up by following the design in Krauss et al. (2017), where a feedforward architecture is used with all neuron layers being fully connected. A 4-layers topology, I–H1–H2–O, is constructed, with the corresponding number of neurons to be 40-40-20-5 in each layer. The 40 neurons within the input layer matches the 40 fundamental/technical features, while the number of neurons in the first hidden layer is set to be the same as that in the input. The subsequent layer follows a dimensionality reduction principle, where the number of neurons is reduced by half. The output layer corresponds to the five-fold classification space. The training set loss function minimization is performed via stochastic gradient descent in a backpropagation manner, with the learning rate and the maximal training epochs set to be 0.002 and 2000 respectively. All DNN relevant computations are carried out in the context of TORCH. The design of a LR machine is relatively simple compared to DNN, and can therefore be seen as a performance benchmark. The five-fold probit function is linearly regressed on the input variables in a multinomial manner (Engel, 1988), by setting the L2 regularization term to be 0.5. All LR relevant computations are done in the framework of scikit-learn.

All the AI models mentioned above are employed to predict the out-of-sample performance between February 8, 2013 to August 8, 2017 based on a rolled training and trading scheme, with the respective training and trading periods being specifically designed. The portfolio constitutes 20 stocks which are selected from the entire market according to the highest predicted probability as described in section 3.1. The 40 fundamental/technical features are prescribed as the input space, and the holding time is set to be 20-day. Relevant statistics of the resultant out-of-sample performances from different machines is exhibited in Table 4.

From Table 4, all learning machines can generate considerable out-performance compared to the general market, where DNN, LR and RF possess 0.2%, 0.12% and 0.18% mean returns per day respectively. The t-statistics extracted from daily return time series gives rise to significant values, demonstrating a statistically meaningful deviation from a zero mean return. In contrast to the general market, the machine learning

Table 3
Modal parameters and performance estimators for the momentum space strategy and multi-factor space strategy.

	Momentum space strategy	Multi-factor space strategy
training period (D)	125	125
rolling period (D)	20	20
sample class no.	5	5
tree no.	60	60
features	Momentum series with different lags	Proprietary multi-factors
annual return	1.8556	1.0121
maximal drawdown	0.569	0.4651
Sharpe	5.0068	2.7509
mean oob	0.2752	0.6814



(a) (b)

Fig. 6. Comparison of the momentum space and multi-factor space strategy performances. (a) Portfolio NV comparison; (b) Hedged NV comparison.

Table 4
Daily return characteristics of portfolio, after transaction costs for DNN, LR, RF compared to general market.

	DNN	LR	RF	General Market
daily mean return	0.002	0.0012	0.0018	0.0006
t-statistics	2.8365	1.8559	2.6435	1.219
standard deviation	0.0228	0.0216	0.0225	0.0186
Skewness	-0.7898	-0.9861	-0.7518	-1.0383
Kurtosis	6.2592	7.5038	5.9751	7.0337
5-percent VaR	-0.0414	-0.0333	-0.0368	-0.033
Maximum drawdown	0.5016	0.4534	0.4651	0.5435
Calmar	2.4618	0.9656	2.176	0.3171

assisted strategy performances exhibit slightly larger but still negative skewness, which is dissimilar from those reported in Krauss et al. (2017). The reason can be tentatively attributed to the effect of different markets and holding time scales. The 5-percent VaR and maximal drawdown of the three machines are comparable to those of the general market, showing that the portfolio might be exposed to systematic risk in 2015. The Calmar ratios for the three machines, however, are substantially higher than that of general market, suggesting exploitable excess returns via these machine assisted strategies.

In regard to the performance comparison between different machines, DNN and RF give rise to comparable daily return statistics, showing strong evidence for profit attainability. The strategy performance for LR is relatively weakened, probably indicating an inefficiency for linear classification in such a market context. Note that the DNN performance can possibly be improved subject to further optimizations of the neural net structure.

5. Conclusions

In conclusion, we have developed a stock selection strategy based on the random forest model and implement it in the Chinese stock market from February 8, 2013 to August 8, 2017. The model is trained by means of the decorrelated decision tree ensemble, and the stock classification is performed via the predicted probability matrix. The strategy conducts stock selection according to the probability ranking for stocks that belong to the first class, where the first 20 stocks with the highest probabilities are selected and held for a certain period of time until the next stock ranking date. The out-of-sample performance is extracted based on a rolled training and trading scheme, where the overall trading period is split into many trading sub-periods, and in each sub-period the RF model trained from the former training dataset is used for prediction. The resultant portfolio, NV, that originated from our fundamental/technical feature space exhibits extraordinary outperformance after transaction costs, compared to the chosen index benchmark, whereas the hedged profile, calculated as the accumulated excess return, reveals a steady growth of profit over the past five years. Although the tuning of modal parameters, such as the tree number, sample class number, length of training period and length of rolling period, would modify the out-of-sample performance to some extent, the strategy effectiveness in this market is guaranteed and the profit is exploitable throughout the testing period. Hence, we summarize that the methodology we employ is reliable and that the machine learning algorithm used here has good capacity of generalization when creating economically sensible stock classifications. Furthermore, the performance of the multi-factor space strategy deteriorates in the most recent year, regardless of the modal parameters, probably due to the popularization of the machine learning techniques or the inadaptability of a long holding period to a more mature market. This detailed reason, however, is subject to further research.

Our factor weight distribution analysis illustrates a structural relationship between the long run excess return and the relevant fundamental/technical feature space. The designed factors are more or less related to the classification, and some fundamentals and long-term technical factors are found to significantly drive the stock price dynamics. Our findings are consistent with the experience in practical stock investment that long-term profit is more relevant to the fundamentals and long-term technical features, thereby proving that the machine learning model can be economically explained.

By comparing the performances between the multi-factor and momentum space strategies, profit opportunities in different time scales are explored. It is interesting to note that the momentum feature space gives rise to even more significant outperformance than does the multi-factor space, demonstrating a larger profitability in the short time scale. In the case of multi-factor space, we see other learning machines, such as DNN, can also generate similar strategy performances as RF does, indicating that such a strategy framework might be suitable for a broad range of learning machine family. These observations suggest that machine learning assists in pattern discovery on the basis of daily data and is probably helpful for quantitative traders in building profitable strategies. Further works and forthcoming papers can be dedicated to the more valuable feature spaces and to the development of novel machine learning algorithms for stock selection strategies on a daily frequency.

Declarations

Author contribution statement

Zheng Tan: Conceived and designed the experiments; Performed the experiments; Wrote the paper.

Ziqin Yan: Analyzed and interpreted the data.

Guangwei Zhu: Contributed reagents, materials, analysis tools or data.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Ahmad, I., Basheri, M., Iqbal, M.J., Raheem, A., 2018. Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE Access 6, 33789–33795.

Alberg, J., Lipton, Z.C., 2017. Improving Factor-Based Quantitative Investing by Forecasting Company Fundamentals, (Nips).

Andriyashin, A., HHrdle, W.K., Timofeev, R.V., 2008. Recursive portfolio selection with decision trees. SSRN Electron. J.

Babu, C.N., Reddy, B.E., 2014. A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data. Appl. Soft Comput. J. 23 (November), 27–38.

Basu, S., 1983. The relationship between earnigns' yield, market value and return for NYSE common stocks: further evidence. J. Financ. Econ. 12 (1), 129–156.

Belciug, S., Sandita, A., 2017. Business Intelligence: statistics in predicting stock market. Ann. Univ. Craiova - Math. Comput. Sci. Ser. 44 (2), 292–298. Retrieved from. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85038626695&partnerID=40&md5=35b8b79a01a6b41338d3857b3e09c36c.

Breiman, L., 1996. Out-of-bag Estimation. Ftp.stat.berkeley.edu/pub/users/breiman/oobestimation.ps.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Brennan, Michael J., Chordia, Tarun, Subrahmanyam, Avanidhar, 1997. Alternative factor specifications, security characteristics, and the cross-section of expected stock returns. J. Financ. Econ. 49, 345–373.

Carhart, Mark M., 1997. On persistence in mutual fund performance. J. Financ. 52, 57–82.

Chen, Long, Novy-Marx, Robert, Zhang, Lu, 2011. An Alternative Three-Factor Model. Working Paper.
Chen, M., Wang, X., Feng, B., Liu, W., 2018. Structured random forest for label distribution learning. Neurocomputing 320 (3), 171–182.
Chong, E., Han, C., Park, F.C., 2017. Deep learning networks for stock market analysis and prediction: methodology, data representations, and case studies. Expert Syst. Appl. 83 (April), 187–205.
Engel, J., 1988. Polytomous logistic regression. Stat. Neerl. 42 (4), 20.
Enke, D., Thawornwong, S., 2005. The use of data mining and neural networks for forecasting stock market returns. Expert Syst. Appl. 29 (4), 927–940.
Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. J. Financ. Econ. 116 (1), 1–22.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33, 3–56.
Guresen, E., Kayakutlu, G., Daim, T.U., 2011. Using artificial neural network models in stock market index prediction. Expert Syst. Appl. 38 (8), 10389–10397.
Hassan, M.R., Nath, B., Kirley, M., 2007. A fusion model of HMM, ANN and GA for stock market forecasting. Expert Syst. Appl. 33 (1), 171–180.
Hou, Kewei, Andrew Karolyi, G., Kho, Bong-Chan, 2011. What factors drive global stock returns? Rev. Financ. Stud. 24, 2528–2574.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer, New York. Doctoral dissertation.
Jegadeesh, Narasimhan, Titman, Sheridan, 1993. Returns to buying winners and selling losers: implications for stock market efficiency. J. Financ. 48, 65–91.
Jia, C., Xu, W., Wang, F., Wang, H., 2012. Track irregularity time series analysis and trend forecasting. Discrete Dynam Nat. Soc. 2012.
Khashei, M., Bijari, M., 2010. An artificial neural network (p, d, q) model for timeseries forecasting. Expert Syst. Appl. 37 (1), 479–489.
Krauss, C., Do, X.A., Huck, N., 2017. Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500. Eur. J. Oper. Res. 259 (2), 689–702.
Leung, M.T., Daouk, H., Chen, A.-S., 2000. Forecasting stock indices: a comparison of classification and level estimation models. Int. J. Forecast. 16 (2), 173–190.
Moskowitz, T.J., Yao, H.O., Pedersen, L.H., 2012. Time series momentum. J. Financ. Econ. 104 (2), 228–250.
Pedregosa, F., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 2011 (12), 2825–2830.
Rather, A.M., Agarwal, A., Sastry, V.N., 2015. Recurrent neural network and a hybrid model for prediction of stock returns. Expert Syst. Appl. 42 (6), 3234–3241.
Sorensen, E.H., Miller, K.L., Ooi, C.K., 2000. The decision tree approach to stock selection. J. Portfolio Manag. 27 (1), 42–52.
Takeuchi, L., Lee, Y., 2013. Applying Deep Learning to Enhance Momentum Trading Strategies in Stocks. Cs229.Stanford.Edu, (December 1989), 1–5. http://cs229.stanford.edu/proj2013/TakeuchiLee-ApplyingDeepLearningToEnhanceMomentumTradingStrategiesInStocks.pdf.
Ticknor, J.L., 2013. A Bayesian regularized artificial neural network for stock market forecasting. Expert Syst. Appl. 40 (14), 5501–5506.
Wang, S., Aggarwal, C., Liu, H., 2018. Random-forest-inspired neural networks. ACM Trans. Intell. Syst. Technol. 9 (6), 1–25.
Yangming, Z., Guoping, Q., 2018. Random forest for label ranking. Expert Syst. Appl. 112 (1), 99–109.
Zhu, M., Philpotts, D., Stevenson, M.J., 2012. The benefits of tree-based models for stock selection. J. Asset Manag. 13 (6), 437–448.
Zhu, M., Philpotts, D., Sparks, R., Stevenson, M.J., 2011. A hybrid approach to combining CART and logistic regression for stock ranking. J. Portfolio Manag. 38 (1), 100–109.