Journal of Banking and Financial Technology
https://doi.org/10.1007/s42786-021-00025-6

ORIGINAL ARTICLE

Machine learning applied to stock index performance enhancement

Tien-Yu Hsu[1]

Received: 23 January 2020 / Accepted: 2 January 2021

Abstract
The project constructs a stock selection model by machine learning methods to enhance the performance of the benchmark index for individual investors. Stock returns prediction is a highly researched topic. However, it is a difficult problem because the stock prices are complex, non-linear, and chaotic. Moreover, overfitting is always an important issue in machine learning field. In this article, it shows that how to solve these problems by dealing with time series data, feature engineering, and model construction. We apply the stock selection model on S&P 500 index and FTSE 100 index. The result shows that the portfolios with stock selection model outperform the benchmarks, and 2% of the number of constitution stocks is the best choice for the stock selection model. Besides, feature importance analysis shows that the stock selection model can measure import features appropriately, which means it has the ability to adapt to different economic environments. In addition, the portfolios with fewer stocks usually outperform the portfolios with more stocks shows the good prediction of the stock selection model. The results imply that machine learning techniques have a good application in stock markets.

Keywords AI · Machine learning · Random forest · Decision tree · Stock prediction · Index

Abbreviations

ETF	Exchange traded funds
ETN	Exchange traded note
ETP	Exchange traded product
NYSE	New York stock exchange
LSE	London stock exchange
MACD	Moving average convergence-divergence
RSI	Relative strength index
TRI	Total return index
TR	Total return
AR	Annualized return
STD	Standard deviation

1 Introduction

Passive investment is popular and popular in recent years, since it can gain the reasonable returns from the markets. Moreover, according to the research, seldom fund managers can beat the benchmark in the long term. Index plays an important role in passive investment. Furthermore, there are more and more indices in the financial markets. It is easy for investors to do passive investment by trading ETF, ETN, ETP and so on. However, not all indices are issued as financial products. Besides, it is difficult for individual investors to buy so much stocks with the specific weighting methods to track the performance of the index. Therefore, constructing a smart stock selection model to enhance the performance of the index is a good way for individual investors.

Factor investment is wildly used by investors because it is effective in the long term and easy to understand. The main factors include size, quality, value, momentum, low volatility, and high dividend yield. However, the effectiveness of these factors varies in different market situations. Some factors perform well in bull markets, while others are good during bear markets. It is difficult for investors to choose the ”right factor” when they invest. Moreover, some investors believe technical analysis works in the stock market, and some investors believe chip analysis is the most important issue in the stock market. There are manifold methods to analyze the stock markets. Unfortunately, most of methods usually only work in a specific period. On the other hand, artificial intelligence shows how smart it is in a lot of fields. For example, AlphaGo, which is an artificial intelligence software, beat the most professional Go player in the world in 2017. Therefore, it is possible to construct a model which

✉ Tien-Yu Hsu
tinahsukcl@gmail.com

[1] Department of Informatics, King’s College London, London, UK

Published online: 25 January 2021

can adapt market situations automatically by machine learning techniques to solve this problem.

1.1 Aims and objectives

The aim of this project is to construct a stock selection model by machine learning to enhance the performance of the index. There are 3 main problems should be solved for this project. First, how to construct features from time series data. Second, how to choose an appropriate machine learning method to stock selection. Third, how to avoid overfitting to enhance the prediction ability. Moreover, determining the best number of the stocks for the stock selection model is also an important objective of this project. Finally, individual investors can enhance the performance of the benchmark index with this stock selection model easily.

In this section, we discuss the topic and the overview of previous studies in this area. In Sect. 2, we give a brief introduction to the basic concepts of Decision Tree and Random Forest. In Sects. 3 and 4, we introduce the stock selection model construction in detail. In Sect. 5, the empirical results of the stock selection model applied to S&P 500 and FTSE 100 are given. Finally, conclusions and directions for the future work are discussed in Sect. 6.

1.2 Background and literature survey

Stock price prediction is a highly researched topic. However, the prices are dynamic, chaotic, and non-linear which make them difficult to be predicted. More and more researchers try to predict financial markets by machine learning techniques during these years [4]. A number of artificial intelligence techniques have been used to predict the stock markets over the past decades. Some researchers predict the movement of stock price by deep learning [15], some researchers construct global stock market investment strategies by machine learning techniques [13]. Moreover, Indu Kumar, Kiran Dogra, Chetna Utreja, and Premtata Yadav 2018 "A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction" shows the stock market trend prediction by 5 supervised machine learning methods, including support vector machine (SVM), Random Forest, K-nearest neighbor (KNN), Naive Bayes, and Softmax. The empirical result shows that Random Forest performs the best for large datasets [7].

On the other hand, investors believe fundamental analysis, technical analysis, and chip analysis are effective. Furthermore, a number of researches support these views. G. William SCHWERT, 1983, showed the empirical results of Size factor [12]. Momentum factors contains both price momentum [2] and trading volume [8, 9], some researches also generate other method to measure the momentum [6]. Low volatility [1] is also an important factor in recent years, because of the population aging. Low beta [10] is also a kind of low volatility factor with market view. The contents of financial statements belong to quality factor [5]. Some investors also care about the dividends [2]. Moreover, technical index such as RSI and MACD are also relative to the stock returns [14]. In addition, chip analysis such as major shareholders are also important. In general, investors construct their trading strategies based on these factors directly. However, it is difficult to rely on the same factors to get excess return all the time.

To solve this problem, we use these effective factors as raw data (input) to the model. In addition, we construct more features from these factors by dealing with time series data. Then we construct the stock selection model by machine learning techniques to make the model have the ability to adapt different economic environment automatically. To sum up, the idea of this stock selection model is to combine the effective factors and machine learning techniques.

2 Background theories

In this project, random forest which is based on decision tree model is the main method of the stock selecting model. Therefore, it is important to understand decision tree and random forest.

2.1 Decision tree

Decision tree is a kind of predictive model based on tree-like model. Moreover, decision tree is also a kind of supervised learning. There are two types of decision trees. One is classification tree, and the other one is regression tree. The target variable with discrete set variables is called classification tree, and the target variable with continuous set variables is called regression tree. In this project, we use regression tree to predict the returns of stocks.

The aim of decision tree learning is to find a small tree consistent with the training examples, then apply it on the testing data. There are a lot of algorithms for decision tree, such as ID3, CART, and so on. In this project, we choose Classification and Regression Tree (CART), because it is good to handle numerical variables.

2.1.1 Basic theorem

The idea of decision tree learning is to choose the most significant attribute as a root of tree recursively. CART creates a binary tree, which means there are only two edges for each node. Node Impurity is a good way to measure the contribution of the feature. The formula is as the following:

$$G(x_i, v_{ij}) = \frac{n_{left}}{N_s} H(X_{left}) + \frac{n_{right}}{N_s} H(X_{right})$$

where x_i is splitting variable, v_{ij} is the splitting value of x_i, n_{left} is the number of left training samples after splitting, n_{right} is the number of right training samples after splitting, N_s is the number of training samples at the node, and $H(x)$ is the impurity function.

There are two criterion for regression trees to calculate impurity [11]. One is variance reduction using Mean Square Error(MSE), and the other one is mean absolute error (MAE).

A. Mean square error (MSE):

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} (y_i - \mu_m)^2$$

where y_i is target i, and $\mu_m = \frac{1}{N_m} \sum_{i \in N_m} y_i$

B. Mean absolute error (MAE):

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} |y_i - \mu_m|,$$

where y_i is target i, and $\mu_m = \frac{1}{N_m} \sum_{i \in N_m} y_i$

In this project, we choose MSE as the node criterion.

2.1.2 Feature importance

The node of feature is higher in a tree, which means the feature is more important to the result. There are several ways to measure feature importance. In this project, we measure the importance of features by node impurity.

$$n_k = w_k \times G_k - w_{left} \times G_{left} - w_{right} \times G_{right},$$

where w_k, w_{left}, w_{right} are the ratio of training samples at node k to total training samples, the ratio of left training samples after splitting to total training samples, and the ratio of right training samples after splitting to total training samples respectively. G_k, G_{left}, G_{right} are the impurity of node k, left, and right respectively.

Moreover, we can calculate the importance of features referring to n_k.

$$f_i = \frac{\sum_{j \in \text{nodes split on feature i}} n_j}{\sum_{k \in \text{all nodes}} n_k}$$

Finally, feature importance normalization is necessary to make the sum of all feature importance equal to 1.

$$f_{n_i} = \frac{f_i}{\sum_{j \in \text{all features}} f_i}$$

2.2 Random forest

According to Leo Breiman, 2001 [3], "Random forests are a combination of tree predictors such that each tree depends on the value of a random vector sampled independently and with the same distribution for all trees in the forest". That is to say, Random Forest constructs many individual decision trees at training, and then get the prediction results by voting. For classification, the result is the mode of prediction of all classification trees. For regression, the result is the mean of prediction of all regression trees. Furthermore, the feature importance of Random Forest is also based on the mean of feature importance in decision trees.

2.2.1 Bagging method

Bagging is an important part of Random Forest. In Random Forest algorithm, it selects a random sample with replacement of training set for each decision tree to fit repeatedly. Therefore, it can get a lot of results from different decision trees, then average them to get the final result.

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

Bagging is a good way to reduce variance without increasing bias. Moreover, it can reduce the relevance between each decision tree.

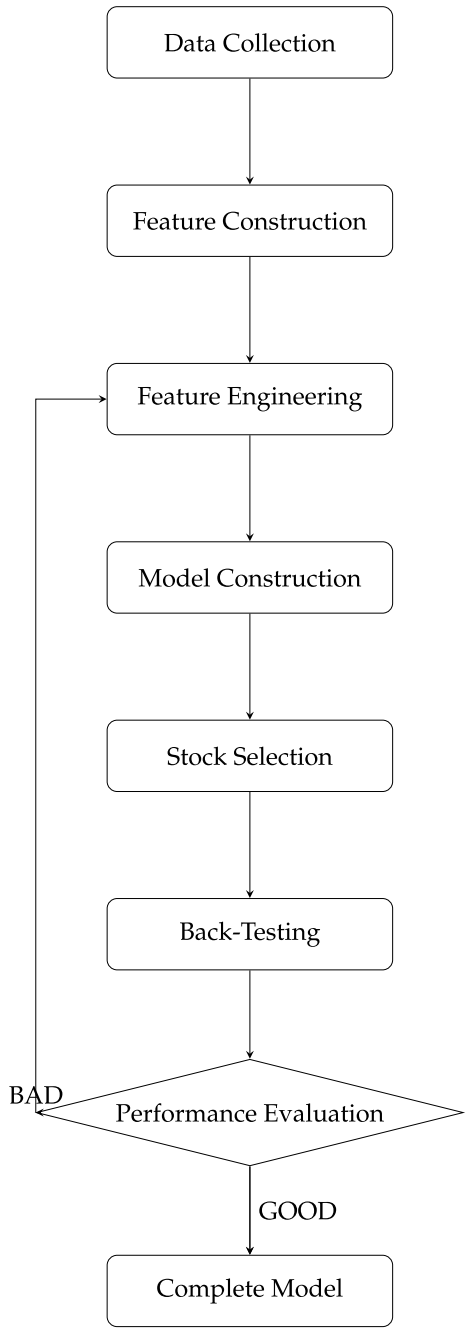
3 Objectives

The article shows how to construct a stock selection model with machine learning methods to enhance the performance of the index for individual investors. In this project, the input of the model are the factors based on fundamental analysis, technical analysis, and chip analysis. In order to make the stock selection model have the ability to adapt to any situations of the financial markets, it constructs the main model with machine learning techniques. Moreover, it can find the useful factors automatically by feature importance.

We construct five portfolios with the stock selection model from the benchmark index. Furthermore, we choose 1%, 2%, 3%, 4%, and 5% of the constituent stocks of the benchmark respectively to represent the number of stocks of the portfolios. Therefore, the appropriate percent of constituent stocks of the model can be found by evaluating the performance of these portfolios.

4 Methodology and implementation

The process of stock selection model construction is as follows. To begin with, we collect both market data and financial statements from the Bloomberg Terminal. next, constructing features base on fundamental analysis, technical analysis, ans chip analysis from the raw data. Moreover, feature engineering is an important process to avoid overfitting. In this article, we construct more features by dealing with time series data, and do feature selection based on f-regression. Furthermore, model construction is to choose an appropriate machine learning model and try to enhance the performance by adjusting some parameters logically to fit the data. In this paper, we choose random forest as the main algorithm model because it is more stable with bagging method and less overfitting with pruning method. Finally, we construct the price-weighted portfolios with the stocks which have been selected from the model, then do back-testing to evaluate the performance of these portfolios. Finally, the stock selection model has been completed.



4.1 Data description

All data is collected from the Bloomberg terminal in this project.

4.1.1 Data for benchmarks

S&P 500 total return index and FTSE 100 total return index are the benchmarks of this project.

A. Data period:
   31/12/2012-31/12/2018
B. Constituent stocks:
   Both indices are reviewed on a quarterly basis, effective after the close on the third Friday of March, June, September, and December. Therefore, we collect the constituent stocks of the indices on the next trading day of effective date to get the new constituent stocks of the indices.
C. Benchmark:
   The daily values of S&P 500 total return index and FTSE 100 total return index.

4.1.2 Data for features

Both market data and financial statement data of the constituent stocks are collected in this project.

A. Data period:
   31/12/2012-31/12/2018
B. Market data:
   Close price, adjusted close price, trading volume, market value, enterprise value, outstanding shares, and institutional held percentage.
C. Financial statement:
   Asset growth, EBITDA growth, leverage, profit margin, revenue growth, return on asset, and return on equity.
D. Ratios:
   Dividend yield, EPS, PB ratio, and PE ratio.

4.2 Feature construction

Features are constructed from the raw data, including fundamental analysis, technical analysis, and chip analysis. Moreover, the project splits the time series data into 12 groups for some features. Furthermore, it calculates the mean, maximum, minimum, amplitude, standard deviation, and so forth of the data for each group and whole periods. Therefore, it can get much information from raw data by constructing the features in this way.

4.2.1 Fundamental analysis

Fundamental analysis contains six main factors, including size, quality, momentum, value, low volatility, and dividend. The project constructs some features from the raw data to represent these factors.

A. Size:
   Market value (MV), Enterprise value (EV)
B. Quality:
   Earnings per share (EPS), Profit margin (PM), Leverage (average asset / average equity), Return on assets (ROA), Return on equity (ROE)
C. Momentum:
   Asset Growth(AG):

$$AG = \frac{Asset_t}{Asset_{t-1}} - 1$$

   EBITDA growth (EG):

$$EG = \frac{EBITDA_t}{EBITDA_{t-1}} - 1$$

   Revenue growth (RG):

$$RG = \frac{Revenue_t}{Revenue_{t-1}} - 1$$

D. Value:
   Price-to-earning ratio (P/E), Price-to-book ratio (P/B)
E. Low volatility:
   The standard deviation of daily returns (STD):

$$STD = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_t - \mu_r)^2}, \quad where \quad r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

F. Dividends:
   Dividend yield (DY)

4.2.2 Technical analysis

Technical analysis is based on the market information, including prices and volumes. In addition, investors usually use a variety of technical analysis indices to check trading signals. In this project, it constructs some features from the raw data on behalf of technical analysis factors.

A. Momentum:
   Price return (PR):

$$PR = \frac{Price_t}{Price_{t-1}} - 1$$

   Volume growth (VG):

$$VG = \frac{Trade\ \ Volume_t}{Trade\ \ Volume_{t-1}} - 1$$

   Market value growth (MVG):

$$MVG = \frac{MV_t}{MV_{t-1}} - 1$$

B. Liquidity:
   Turnover rate (TOR):

$$TOR = \frac{Trading\ \ Volume}{Outstanding\ \ Shares}$$

C. Technical analysis index:
RSI:

$$RSI = (1 - \frac{1}{1+RS}) \times 100\%, where\ \ RS = \frac{EMA(U,n)}{EMA(D,n)}$$

$$U = \begin{cases} P_t - P_{t-1}, & \text{if } P_t \geq P_{t-1} \\ 0, & \text{if } P_t < P_{t-1} \end{cases}$$

$$D = \begin{cases} 0, & \text{if } P_t > P_{t-1} \\ P_{t-1} - P_t, & \text{if } P_t \leq P_{t-1} \end{cases}$$

MACD:

$$OCF = DIF - DEM, where\ \ DIF = EMA_{(close,12)} - EMA_{(close,26)}$$

$$DEM = EMA(DIF, 9)$$

4.2.3 Chip analysis

Chip analysis refers to stock holders. The major stock holders usually can affect the price of the stocks. It chooses institutional held percentage and the growth of institutional held of the stocks to represent the chip analysis factor in this project.

4.2.4 Dealing with time series data

It is a good way to describe the features from time series data by statics. The statics of time series data $\{X_t\}_{t=1}^{T}$ are as follows.

- Mean:

$$\mu = \frac{\sum_{t=1}^{T} X_t}{T}$$

- Standard deviation:

$$\sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_t - \mu_r)^2}, \ \ where\ \ r_t = \frac{X_t - X_{t-1}}{X_{t-1}}$$

- Maximum:

$$max = max\{X_t\}, t = 1, 2, ..., T$$

- Minimum:

$$min = min\{X_t\}, t = 1, 2, ..., T$$

- Amplitude:

$$amplitude = max - min$$

Moreover, the frequency of raw data is different. Therefore, it splits the data with daily frequency into 12 groups, which means a group usually contains 5 trading days. Thus, we can get much information from the time series data. Finally, the number of features expends to 667 from 22 by splits time series data into groups.

4.3 Model construction

The goal of the model is to select top N stocks from the universe of the index. First, splitting the time series data refer to the benchmark index review days. Second, constructing the features from the raw data as above.

4.3.1 Training data and testing data

The project splits the training data and testing data refer to the benchmark index review days.

4.3.2 Feature selection

Overfitting is always an critical issue in machine learning, especially when the number of features is more than the number of samples. To avoid this problem, we keep the number of features at 60% of the number of samples by feature selection.

Univariate feature selection is to select the best features based on univariate statistical tests. "SelectBest" is one of the methods to remove all but the k highest scoring features. The score is based on F-value between labels and features for regression tasks in this project.

The process is as follows:

First, compute the correlation between each regressor and target.

$$r_i = \frac{(X_i - \bar{X})(y_i - \bar{y})}{\sigma_X \sigma_y}$$

Second, compute the f score:

$$f = \frac{r_i^2}{1 - r_i^2} \times (n-2)$$

Finally, we can convert it to p-value by the F-Distribution table.

To sum up, the feature is more effective to the target if its f score is bigger.

4.3.3 Machine learning method

In this project, we choose Random Forest as the main model for two reasons. First, it is based on decision trees, which means it can show which features are more important clearly.

These important features are important for us to understand the financial markets. Second, the data of financial markets is complicated and non-linear. Random Forest is easy to avoid overfitting by bagging process. Moreover, it combines a number of decision trees to enhance the prediction ability and stability.

On the other hand, computational overhead is a shortage of Random Forest, because it contains many decision trees which means it needs to calculate much more times than a pure decision tree when it outputs a result. The number of decision tree is an important issue in Random Forest algorithm, and it is a trade off between prediction stability and calculation efficiency. In this project, we choose 100 as the number of decision trees because it is enough to keep the stability of the result. Moreover, consider the maximum number of samples is about 500, and the number of features is only 667 for each sample, because time series data has been dealing with by feature construction. Therefore, the size of data is not so big, which means it still can work well with Random Forest algorithm. Moreover, in order to enhance the calculation efficiency with Random Forest, we also use parallel computing.

In addition, if the noise of sample data is too large, it still causes overfitting in Random Forest algorithm. Unfortunately, the prices of stocks are difficult to predict because it is complicated, chaotic, and non-linear. In order to solve this problem, it prunes the leafs after constructing the decision trees. In this project, it prunes the leafs with samples lower than 10% of the training samples. The idea is to make the model split the returns of the samples into about 10 groups, because all of us know it is difficult to predict the stock returns precisely. However, it is a good way to predict the returns of stocks roughly, then average the results from the decision trees. Finally, we will get the prediction of stock returns from the model, and select top N stocks which perform better relatively as the constituent stocks of the portfolio.

4.4 Back-testing

Firstly, the process of constructing portfolios are as the following:

Step 1. Check the universe, which means the constituent stocks of the benchmark index on the portfolio-reviewing day.

Step 2. Select N stocks with the stock selection model as the portfolio. In this project, N is 1%, 2%, 3%, 4%, and 5% of the number of constituent stocks of the benchmark index respectively.

Step 3. In order to make the model easy to be used for individual investors, we choose price-weighted portfolio in this project. It means all stocks in the portfolio have the same unit.

Secondly, we do back-testing based on adjusted price, which adjusted the dividends and corporate events.

4.5 Performance evaluation

We calculate total return, annualized return, standard deviation, Sharpe ratio, and hit ratio to evaluate a portfolio performance during the back-testing period. Moreover, we calculate the quarterly return to check if a portfolio performed stably between each index review day. Furthermore, we compare the performance of all portfolios with the benchmark to evaluate if the portfolio outperformed steadily.

A. Total return (TR):

$$TR = \frac{P_T}{P_1} - 1,$$

where P_t is the value of the portfolio at time t.

B. Annualized return (AR):

$$AR = (1 + TR)^{\frac{1}{n}} - 1,$$

where n is the year of back-testing period.

C. Annual standard deviation (ASTD):

$$ASTD = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_t - \mu_r)^2} \times \sqrt{252}$$

$$, \ where \ r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

D. Sharpe ratio (SR):

$$SR = \frac{AR - r_f}{ASTD},$$

where r_f is risk-free rate.

We choose $r_f = 0$, because the interest rate is very low over these years. Furthermore, the goal is to compare the performance of these portfolios with the performance of the benchmark in this project. It is reasonable to make $r_f = 0$ because it will not affect the relative results.

E. Hit ratio (HR):

$$HR = \frac{\text{Numbers of the portfolio outperformed the benchmark}}{\text{Total numbers}}$$

The back-testing period is 5 years, and it covers 20 quarter periods. Therefore, the total number is 20 in this project.

5 Empirical results

The project applies the stock selection model to S&P 500 and FTSE 100 indices, which are the benchmark of the US stock market and the UK stock market respectively. The back-testing period is 5 years between 01/01/2014 and 31/12/2018, which contains business cycles.

In this chapter, we focus on the performance of the portfolios, and the appropriate number of stocks for the stock selection model. Moreover, feature importance analysis is also important. It makes us understand the key factors to stock markets.

5.1 S&P 500 index

S&P 500 index has about 500 constituent stocks. Therefore, we choose 5, 10, 15, 20, and 25 as the number of stocks of the portfolios, which are the 1%, 2%, 3%, 4%, and 5% of the universe respectively. Moreover, the benchmark is S&P 500 total return index, which implies the dividends are reinvested into the index.

5.1.1 Performances of portfolios

First, Fig. 1 shows the total returns of the portfolios, and the performance of S&P 500 total return index. Orange line, green line, red line, purple line, and brown line represent 5, 10, 15, 20, 25 as the stock number of the portfolios respectively. The blue line on behalf of the benchmark. It is clear that all portfolios with stock selection model outperformed the benchmark. Moreover, in general, the portfolio with fewer number of stocks performed better than the portfolio with more number of stocks. In addition, the trend of portfolios was similar to the trend of the benchmark, except to a few specific periods. Furthermore, whatever in bull markets or bear markets, it seems that the volatility of portfolios were much higher than the benchmark.

Second, move on to the Fig. 2. It shows the portfolio returns and the benchmark returns between every index review days. Most of the time, the returns of the portfolios beat the benchmark. Moreover, when the portfolios beat the benchmark, they usually outperformed the benchmark a lot, which means the returns of the portfolios were many times of the return of the benchmark. In addition, look at the third quarter of 2015, most of portfolios had positive returns while the market faced recession. However, the portfolios not always performed so well, especially in 2017. Look at

Fig. 1 Performance of ML portfolios and S&P 500 TRI

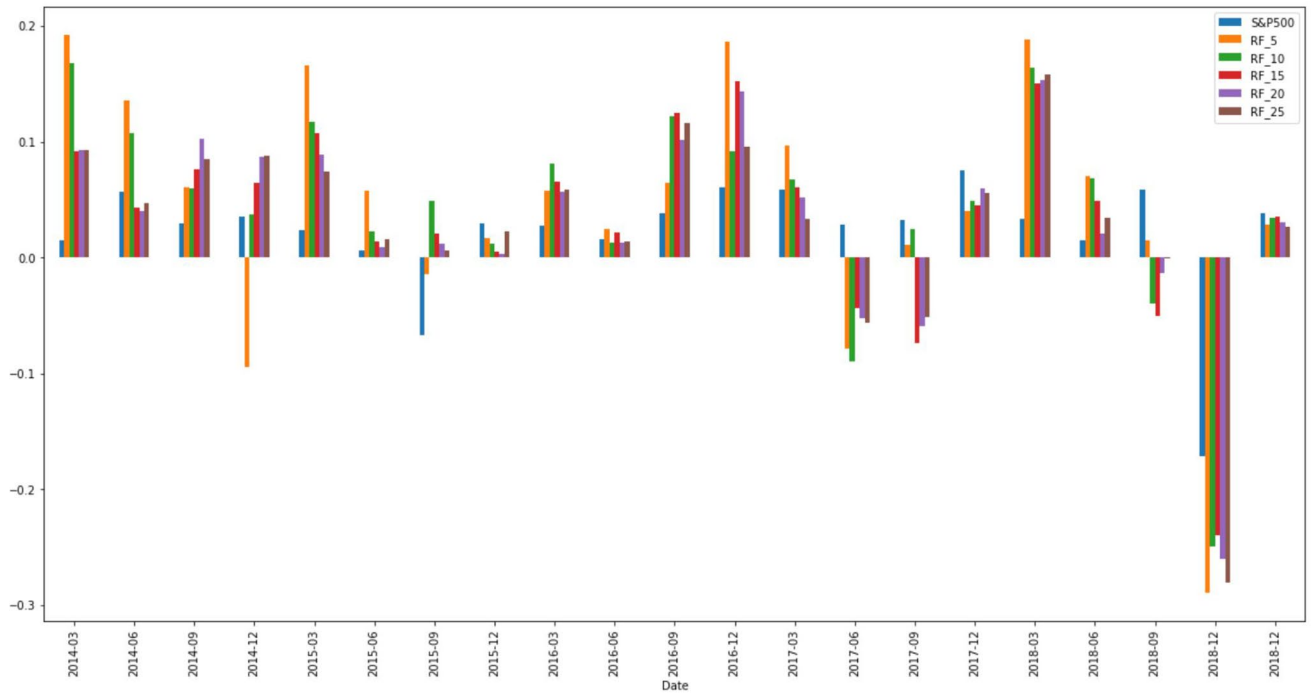


Fig. 2 Quarterly performance of ML portfolios and S&P 500 TRI

the second and third quarter of 2017, most of portfolios had negative returns while the benchmark still had the positive returns. Although the portfolios had the positive returns at the end of 2017, All of them still could not catch up to the return of the benchmark. Additionally, the volatility of portfolios were grater than the volatility of the benchmark. In general, the portfolios with fewer stocks had higher volatility than the portfolios with more stocks.

Finally, move on to the summary table of portfolios in Table 1. It shows the performance of the portfolios with different numbers of stocks and the benchmark. It is clear that the annualized return of all the portfolios were better than the benchmark. However, the Sharpe Ratio of the portfolio with 25 stocks was lower than the benchmark, because the standard deviation of it was much higher than the benchmark. The Sharpe ratio represents the ability of getting returns with the same risk. Most of portfolios with stock selection model performed better than the benchmark excepted for the portfolio with 25 stocks. Besides, the hit ratios of all portfolios were greater than 50%, which implied they usually beat the benchmark.

Clearly, the portfolios with 5 stocks and 10 stocks performed better than other portfolios, because both of them had higher returns and hit ratios. Moreover, the 10-stocks portfolio with higher annualized return and lower standard deviation made its Sharpe ratio higher than the 5-stocks portfolio, which means it could get more profit with the same risk. In addition, the hit ratio was 65%, which means it beat the benchmark 13 times during the past 20 quarters. Therefore, 10-stocks portfolio was the best choice for the stock selection model.

5.1.2 Feature importance

Finding important features is also an important issue, since it can make us understand the financial market deeply by the useful features. The back-testing period is 5 years, and it had 22 sets of training data

Table 1 The performance of portfolios and S&P 500 TRI

Number of stocks	5	10	15	20	25	Benchmark
Total return (%)	121.1	124.2	88.5	80.4	72.0	50.3
Annualized return (%)	17.2	17.5	13.5	12.5	11.5	8.5
Standard deviation (%)	23.2	19.6	18.1	18.3	17.9	13.2
Sharpe ratio (%)	74.3	91.1	74.6	68.3	64.1	64.2
Hit ratio (%)	65	65	65	55	55	–

(4 *quarters* × 5 + 1 for the first + 1 for the last), which means we had 22 training times.

Originally we constructed 667 features for each sample every time. To avoid overfitting, it only kept 60% of the number of samples by feature selection. That is to say, it only kept 300 features each time. 664 features had been selected at least once, and the mean of the frequency of the features was 9.9, which means a feature showed up about 10 times during the past 22 training times. The maximum of the feature appearance frequency was 19 times, which included "RSI_20_4average", "RSI_20_6average", and "TOR_9min".

However, considering it also pruned the leafs which less than 10% of samples in Random Forest algorithm, which means the useful features for each tree should much less than 300. Therefore, it is dangerous to judge a feature only based on the appearance frequency. Thus, we found top 10 features for each training by calculating feature importance. The results showed that 157 features appeared at least once in the top 10 important features over the 22 training times. However, the mean of these feature appearances was only 1.4, which means important features changed all the time. Furthermore, the maximum of the important feature appearance frequency only 3 times, including "MACD_2min", "MV_5std", "PB_0std", "PB_11std", "PB_1std", "PB_8std", "PE_11amplitude", "PE_1std", "PMlast", "RGlast", "RSI_10_11std", "RSI_10_4average", and "TOR_11amplitude". These features showed that dealing with time series data is very important.

To sum up, feature importance analysis shows the effectiveness of the stock selection model. The stock selection model can weight the features appropriately by itself, which means it has the ability to face the different financial environments.

5.2 FTSE 100 index

FTSE 100 index has about 100 constituent stocks. In this case, we choose 1, 2, 3, 4, and 5 as the number of stocks of the portfolios, which are also 1%, 2%,3%, 4%, and 5% of the universe respectively. Besides, the benchmark is FTSE 100 total return index.

5.2.1 Performances of portfolios

Firstly, look at Fig. 3, orange line, green line, red line, purple line, brown line, and blue line represents the portfolio of 1, 2, 3, 4, 5 stocks, and the benchmark. Clearly, all portfolios performed better than the benchmark. Moreover, the volatility with the portfolios were also greater than the benchmark, especially for the portfolios with 1 stock and 2 stocks.

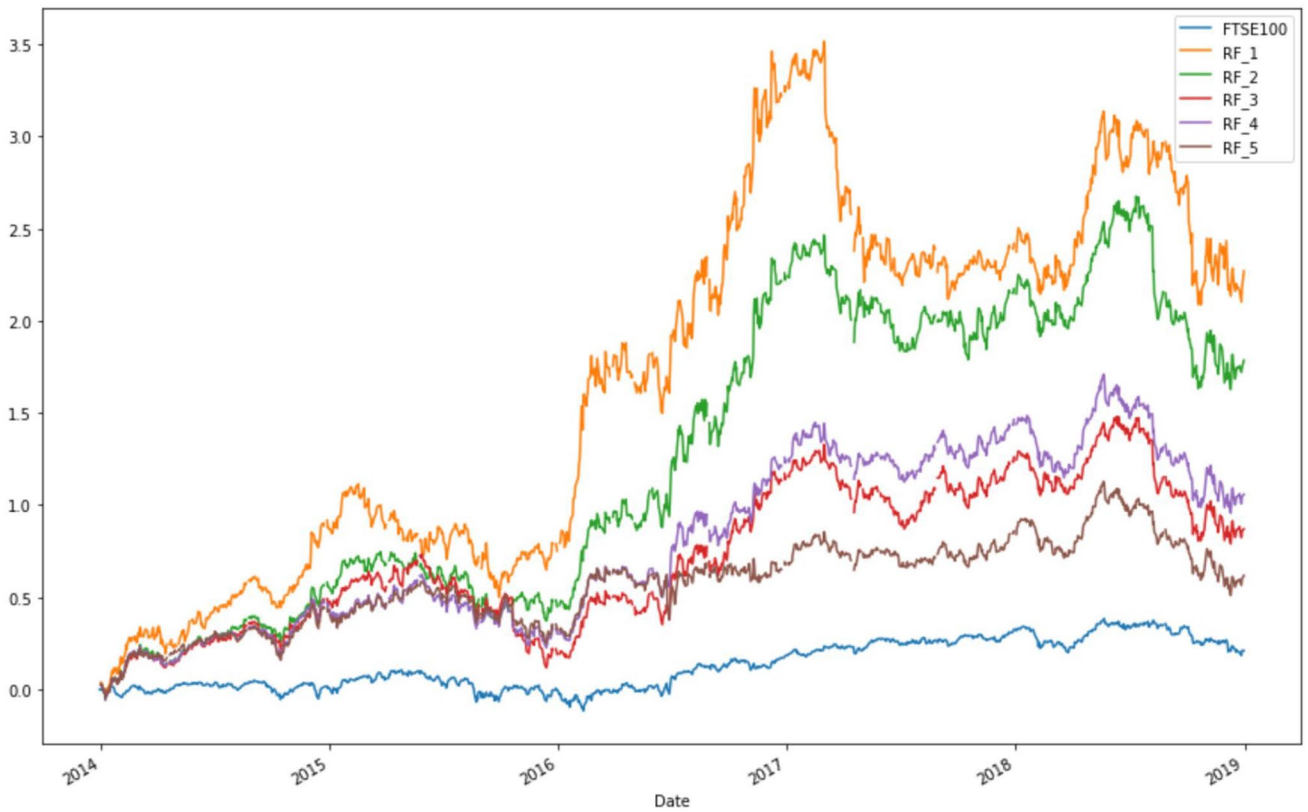


Fig. 3 Performance of ML portfolios and FTSE 100 TRI

In addition, portfolios with fewer stocks outperformed the portfolio with more stocks besides portfolio with 3 stocks. On the other hand, the trend of portfolios with 3, 4, and 5 stocks are much similar to the benchmark than the portfolios with 1 and 2 stocks.

Secondly, move on to Fig. 4. It shows the performance of each portfolio and the benchmark every quarters. The volatility of portfolios were greater than the benchmark. In general, the performance of portfolios with fewer stocks fluctuated much more than the portfolios with more stocks. In addition, the portfolios usually outperformed the benchmark a lot. look at the first and forth quarter of 2014, all portfolios got higher positive returns while the benchmark got negative returns. Moreover, look at the first and forth quarter of 2016, returns of the portfolios were many times of the returns of the benchmark. However, sometimes portfolios not performed as well as the benchmark. For example, most portfolios got more negative returns than the benchmark in the second and third quarter in 2015 and the third and forth quarter of 2018. Besides, all portfolios got negative returns while the benchmark got positive return in the second quarter of 2017.

Finally, Table 2 shows the summary of the performance of the portfolios and the benchmark. Clearly, all portfolios outperformed so much of the benchmark. Although the standard deviation of all portfolios were greater than the benchmark, the Sharpe ratio of portfolios were still better than the benchmark. In general, the portfolios with fewer stocks performed better than the portfolios with more stocks. However, there was something wrong about the portfolio with three stocks, because the 4-stocks portfolio outperformed the 3-stocks portfolio. It implied the stock returns prediction was a little bit imprecise.

On the other hand, the portfolio with one stock had the highest return. However, the volatility of the portfolio was also the highest, which led to the lower Sharpe ratio. Although the annualized return of the portfolio with two stocks was not as good as the portfolio with one stock, but the volatility of it was

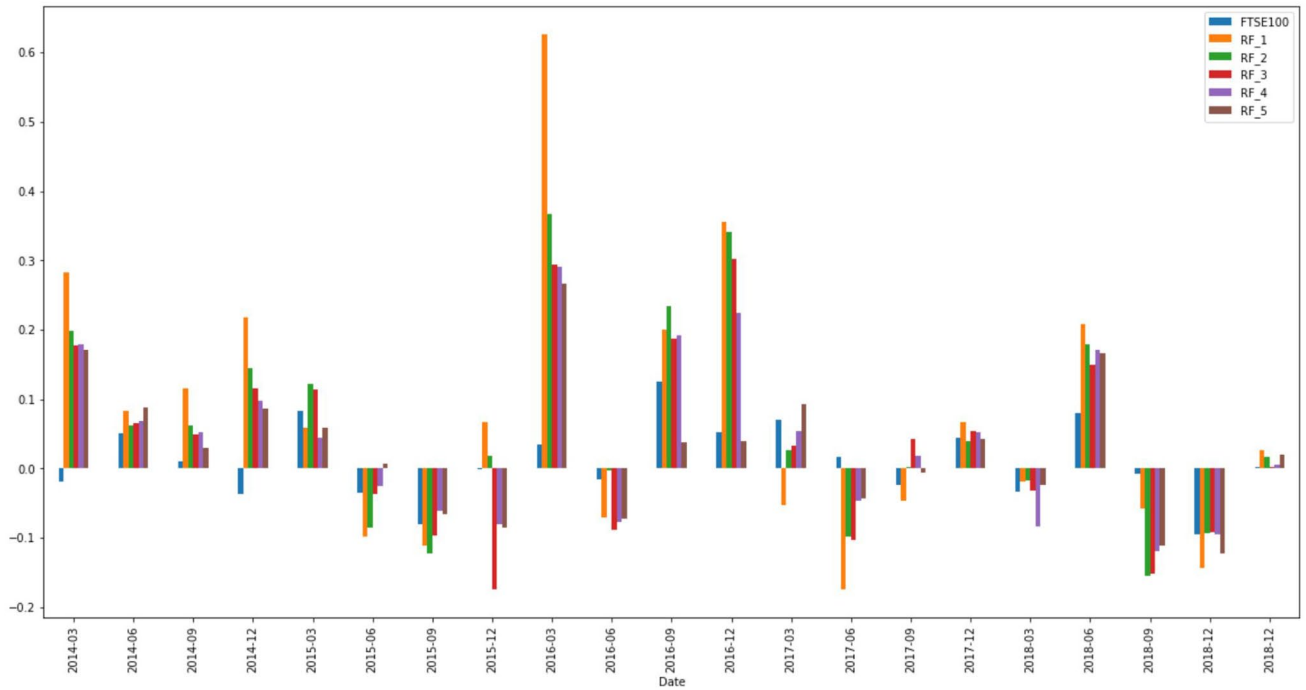


Fig. 4 Quarterly performance of ML portfolios and FTSE 100 TRI

Table 2 The performance of portfolios and FTSE 100 TRI

Number of stocks	1	2	3	4	5	Benchmark
Total return (%)	226.9	178.5	87.1	105.9	61.8	20.9
Annualized return (%)	26.7	22.7	13.3	15.5	10.1	3.9
Standard deviation (%)	27.2	22.8	21.4	20.7	20.6	13.6
Sharpe ratio (%)	98.1	99.8	62.4	75.2	49.1	28.6
Hit ratio (%)	55	70	65	65	55	–

much lower than the portfolio with one stock. Therefore, the portfolio with two stocks had the highest Sharpe ratio, which means it could earn more profit than other portfolios with the same risk. In addition, the hit ratio of all portfolios were higher than 50%, which means the performance of the portfolios with stock selection model usually beat the benchmark. Furthermore, the hit ratio of the portfolio with two stocks was the highest (70%). To sum up, the portfolio with two stocks had the highest Sharpe ratio and hit ratio. Therefore, 2-stocks portfolio was the best choice for the stock selection model.

5.2.2 Feature importance

Originally, we also constructed 667 features for each sample every time. However, it merely kept 60% of the best features by feature selection to avoid overfitting. Thus, it only kept 60 features for each training. The results show that 545 features had been selected at least once, and the mean of feature appearance was 2.4 times. However, the maximum of the feature appearance times only 6, which means the same features only showed up up to 6 times over the past 22 training times. Moreover, these features were "AdjPriced", "DY_11amplitude", "DY_5max", "DY_6max", "PB_10std", "PB_4std", "PE_10std", "Price_10std", "Price_9std", and "RSI_20_9max". The result also showed how important of dealing with time series data again.

On the other hand, pruning the leafs which less than 10% of samples made a lot of features become useless. In order to measure the importance of features precisely, we calculated the feature importance for each feature in every training. Then we picked up the top 10 important features at each training. The result shows that 171 features were top 10 important features at least once over the past 22 training times. However, the mean of it was only 1.3, which means the same features were difficult to become an top 10 important features again. Furthermore, there were only 8 features showed up on the top 10 feature list more than 2 times, including "DY_1amplitude", "MV_0amplitude", "MV_1amplitude", "MV_9amplitude", "PB_10std", "PB_4std", "PB_9amplitude", "RSI_10_9max". The result shows that the important features changed all the time.

According to the feature analysis, it is clear that the stock selection model has the ability to measure the important features by itself, which means it can face the different market situations automatically. Besides, it shows that dealing with time series data works in the model.

6 Conclusion

According to the empirical result, it shows effectiveness of the stock selection model by applying it on both S&P 500 index and FTSE 100 index. Most portfolios with stock selection model preformed much better than the benchmarks. In addition, portfolios with fewer stocks usually performed better than portfolios with more stocks, which implied the accuracy of the prediction of stock selection model. Moreover, the stock selection model showed how smart it was by feature importance analysis. The result shows that it could measure the importance of features by itself, which means it could adapt appropriately to different financial market environments.

Furthermore, the stock selection model with 2% of the number of constitution stocks is the most effective to enhance the performance of the index in the long term. The portfolios with 2% of the number of constitution stocks had the highest Sharpe Ratio and hit ratio for both indices, which implied that it could diverse the risk effectively. Therefore, 2% of the number of constitution stocks is a good choice for the stock selection model.

On the other hand, there are some important issues which we should focus on. First, although both total returns and annualized returns of portfolios with stock selection model performed much better than the benchmarks, they were not the "real returns" because we did not consider transaction costs in the project. Second, the stock selection model didn't consider the investable capacity, which means users could not make sure if it is easy to construct the portfolios. In this case, the size of constitution stocks of both S&P 500 index and FTSE 100 index were big enough, it is difficult to face this problem. However, we still need to keep this in mind when we apply the stock selection model to other indices.

Finally, there are still some works for the future research, such as improving the method of dealing with missing value, importing other effective features, and applying the model to other type of the index. In this project, we roughly fill all missing value to be zero because the data only had a few missing values. However, if we can construct a good method to fill these missing value, the prediction of the stock selection model might be much precisely. Moreover, it will provide us an opportunity to import more features into the model. There are still manifold features we can construct from fundamental analysis. However, most of features have missing values in specific industries, it is difficult to choose these features without a well done missing value filling method. Furthermore, both S&P 500 index and FTSE 100 index belong to size factor index. It is a good idea to apply the stock selection model to other type indices, such as value, quality, low volatility, and so forth.

In conclusion, the stock selection model is effective, and 2% of the number of constitution stocks is the best choice for the model. Most importantly, it implies that the machine learning techniques can have a good application in stock markets.

References

1. Yuhang X, Xiaoyan Z, Andrew A, Hodrick J (2006) The cross-section of volatility and expected returns. J Financ 61:259–299
2. Asem E (2009) Dividends and price momentum. J Bank Financ 33:486–494
3. Breiman L (2001) Random forest. Mach Learn 45:5–32
4. Kimura H, Henrique BM, Sobreiro VA (2019) Machine learning techniques applied to financial market prediction. Literature review. Expert Syst Appl 124:226–251
5. Khairurizka R, Mulyono DM (2009) The effect of financial ratios, firm size, and cash flow from operating activities in the interim report to the stock return. Chin Bus Rev 8(6):44–55
6. Hühn H, Scholz HL (2018) Alpha momentum and price momentum. Int J Financ Stud 6
7. Ultreja C, Yadav P, Kumar I, Dogra K (2018) A comparative study of supervised machine learning algorithms for stock market trend prediction. IEEE
8. Christopher GL, William DL (2012) Endogenous trading volume and momentum in stock-return volatility. J Bus Econ Stat, pp 253–260
9. Lee CMC (2002) Price momentum and trading volume. J Financ 55:2017–2069
10. Levy RA (2018) Beta coefficients as predictors of return. Financ Anal J 67:61–69
11. Stacey R (2018) The mathematics of decision trees, random forest and feature importance in scikit-learn and spark
12. Schwert GW (1983) Size and stock returns, and other empirical regularities. J Financ Econ 12:3–12
13. Kwon DS, Sohn SY, Lee TK, Cho JH (2019) Global stock market investment strategies based on financial network indicators using machine learning techniques. Expert Syst Appl 117:228–242
14. Ng W-K, Chong TT-L (2008) Technical analysis and the London stock exchange: testing the macd and rsi rules using the ft30. Appl Econ Lett 15:1111–1114
15. Cui L, Long W, Lu Z (2019) Deep learning-based feature engineering for stock price movement prediction. Knowl Based Syst 164:163–173