A Study of Stock Portfolio Strategy Based on Machine Learning

Zhuoyuan Ouyang

SUSTech Business School, Southern University of Science and Technology, Shen Zhen ,518055, China
Corresponding author. Email: 646491441@qq.com

ABSTRACT

At present, artificial intelligence is a hot topic in the field of finance. With the continuous development of domestic quantitative investment technology, it is increasingly difficult to obtain excess returns from traditional quantitative investment methods. Artificial intelligence, as a new data analysis and forecasting tool, has excellent processing capability for high-dimensional and serial data in the field of quantitative investment. As a result, quantitative investment has become one of the key areas where artificial intelligence is empowering the financial industry. In this paper, the data of listed companies in the New York Stock Exchange was used as the fundamental dataset. Twelve factors were selected as input variables for machine learning training. In terms of research methodology, portfolios were first selected based on different model algorithms, then the actual performance of each algorithm was back-tested, and investors were simulated to hold the portfolios for a long period. To ensure that the conclusions are better guided in practice, this paper attempts to apply the emerging machine learning algorithms and classical machine learning algorithms to the study of New York stock market returns, and to compare and discuss the predictive power of the algorithms on portfolio performance. The results of the study show that the portfolios selected by support vector regression and neural networks outperform the Dow Jones Index in the face of high noise and small sample space. In contrast, the emerging machine learning algorithms Adaboost regression and Bayesian Ridge regression performed slightly worse than the Dow Jones Index.

Keywords: *quantitative investment, machine learning, support vector machine, neural network, rate-of-return prediction*

1.INTRODUCTION

The financial industry, as the core of the modern economy, is an important core competence of the country. Currently, the financial industry of China is growing rapidly. Financial markets are evolving to become more professional, broader, and deeper. People are also becoming more financially aware and sophisticated in their investments. The financial markets are a very complex system, full of uncertainties. In such a complex financial market, how financial institutions or individual investors can effectively control risk while pursuing high returns is a major theoretical research question. As a result, the portfolio problem has become a central issue in financial theory and an important part of financial engineering.

The fundamental purpose of an investment portfolio is to diffuse risk. Institutional and individual investors both want to maximize returns while minimizing investment risk. A portfolio is a reasonable allocation of assets with the goal of maximizing returns and minimizing risk. Research on this issue can not only improve the efficiency of financial services, promote financial business innovation, lower the threshold of financial services, enhance the coverage of financial services to promote financial inclusion, but also reduce the cost of financial services, reduce the risk to ensure the stability of the financial system. The traditional portfolio is mainly based on the mean-variance theory of Markowitz (1952) [1] and the Capital Asset Pricing Model (CAPM) established by three economists, Sharpe (1964) [2], Lintner (1965) [3] and Mossin (1966) [4]. However, the traditional theoretical models which are based on a large number of strict assumptions does not take into account various dynamics, stochastic, and other uncertainties under realistic conditions.

In recent years, with the development of intelligent algorithms, the use of intelligent algorithms to solve portfolio optimization problems is gradually recognized by most scholars. Researchers have proposed many intelligent optimization algorithms by imitating some phenomena and processes in the natural. For example, by drawing on Darwinian evolution, Deb et al. (2002) proposed the NSGA-II algorithm based on NSGA [5]. Nowadays, these intelligent optimization algorithms have been widely used to solve various portfolio optimization problems. Some of these algorithms have had initial applications in solving portfolio optimization problems. For example, Anagnostopoulos & Mamanis solved a multi-objective portfolio model based on mean-variance approach using multi-objective genetic evolutionary algorithms such as NSGA-II, PESA, and SPEA2 [6].

This paper focuses on the application of machine learning in the field of portfolios with the real stock market as the research object. By introducing a machine learning approach, an intelligent set of smart methods for portfolio optimization is constructed, which leads to an efficient allocation of capital.

2.LITERATURE REVIEW

There have been some preliminary studies by some scholars on the application of machine learning methods to the stock market. Ye Wang et al. (2012) used support vector regression to forecast the future CPI index and explored the relationship between FCI (Financial Conditions Index) and future inflation [7]. Felipe Dias Paiva et al. (2019) evaluate the São Paulo Asset Stock Exchange Index (Ibovespa stocks index) using a machine learning-based classifier, a support vector machine approach, and a mean-variance approach for portfolio selection [8].

In a practical stock market return forecasting application, Rapach et al. (2013) applied Lasso regression to predict global stock market returns using global lagged returns [9][10]. Harvey et al. (2018) used a self-help procedure to study the multiple comparison problem [11]. Bin Li (2019) used ridge regression, Lasso regression algorithm to construct a prediction model and identify the dissimilarity factor [12]. In order to capture some nonlinear characteristics of stock markets, some scholars have applied random forest and neural network methods to the study of predicting stock returns. Patel and Shah et al. (2015) predicted the direction of stock movement and stock price index in the Indian stock market using four models: support vector machine, artificial neural network, naive Bayes, and random forest [13]. The comparison revealed that the overall performance of the random forest was better than the other three prediction models. Adil Moghar et al. (2020) used LSTM recurrent neural network (RNN) to predict the future stock market value and studied the accuracy of the algorithm prediction [14]. As for some emerging machine learning algorithms, Xiao-dan Zhang et al. (2016) proposed a new status box feature method to predict stock trends and used a new hybrid classifier integrating Adaboost algorithm, genetic algorithm, and support vector machine to solve the problem of imbalance in the classification of stock turning points [15].

The key point of this paper is to apply different machine learning methods to the study and prediction of stocks in the New York Stock Exchange. The prediction results of different modeling approaches were compared and analyzed. The validity of the results was studied and one of the most effective machine learning methods was selected to provide a reference for investors. In addition, based on traditional machine learning methods, this paper tries to apply the new emerging machine learning methods to the market. Based on the forecasting results, whether the use of emerging models in the real environment of NYSE outperforms traditional models were discerned. It was also demonstrated whether each model can adapt to the complex environment of multiple constraints in the stock market in order to determine the applicability of the model in the real market.

3.PORTFOLIO STRATEGY APPROACH

A portfolio is a combination of various types of assets such as stocks, bonds, and cash that an investor allocates, taking into account investment risks, expected investment returns, and other factors. The objective is to minimize the risk of the investment while guaranteeing a certain expected return. In this paper, the portfolio was selected based on the assessment of the future performance of assets by machine learning methods and the proportion of each asset is allocated in equal proportions.

3.1. Experimental Steps

The experiment in this paper simulated an investor picking a portfolio and holding it for the long term (bought in January 2021 and sold at the end of June 2021) based on publicly available data for stocks from June 2016 to December 2020. First, the obtained public data were preprocessed, which included missing value processing, depolarization, and data transformation. The purpose of this is to transform all the data into data that can be received as input by the models. Second, this experiment used six algorithms: Random Forest (RF), Support Vector Regression (SVR), Neural Network (NN), Decision Tree (Tree), Adaboost Regression, and Bayesian Ridge Regression to learn and predict the processed data, and focuses on SVM and NN. The sample data from June 2016 to December 2020 were selected as the training set for the construction of the model and optimization of the parameters. The parameters to be optimized include the number of neurons in the hidden layers of the neural network, the number of decision trees in the random forest, and the number of split features in each decision tree. In addition, the period from January 2021 to June 2021 was selected as the model backtesting period, and the model backtesting period was used to predict and rank the increases and decreases of stock price using the December 2020 factor data to obtain the 10 stocks with the highest predicted returns. By repeating the above training and backtesting operation several times, the 10 stocks with the highest number of occurrences among the 10 stocks with the highest predicted returns were selected as the final portfolio. The actual return of the portfolio was calculated based on the actual performance and compared to the Dow Jones Index (DJIA).

3.2. Support Vector Regression Model

Given a training sample $\mathrm{D} = \{(x_1, y_1),(x_2, y_2), \ldots ,(x_m, y_m)\}, y_i \in \mathrm{R}$, it was desired to obtain a regression model such that $f(x)$ was as close as possible to y . The ω and b were the model parameters to be determined. For sample (x, y), traditional regression models usually calculated the loss directly based on the difference between the model output $f(x)$ and the true output. The loss was zero when and only when $f(x)$ is exactly the same as y . In contrast, the support vector regression assumption tolerates at most the deviation between r and d. In contrast, the support vector regression assumption tolerated a deviation of at most ε between $f(x)$ and y . Thus a loss function with a regular term could be obtained:

$$\frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2 + \frac{\lambda}{2}\| \omega \|^2$$
（1）

It could be seen that the latter part of this error function was similar to the objective function in the SVM. We then replaced the quadratic error function with an -insensitive error function, so the SVR problem form could be reduced to:

$$C\sum_{n=1}^{N}E_{\varepsilon}\left(y(x)-t_n\right)+\frac{1}{2}\| \omega \|^2$$
（2）

The -insensitive error function was as follows.

$$E_{\varepsilon}\left(y(x)-t\right)=\begin{cases} 0, & if\ \left|y(x)-t\right|<\varepsilon \\ \left|y(x)-t\right|-\varepsilon, & otherwise \end{cases}$$
（3）

After introducing the slack variables we obtained:

$$\begin{aligned} t_n &\le y\left(x_n\right)+\varepsilon+\xi_n \\ t_n &\ge y\left(x_n\right)-\varepsilon-\hat{\xi}_n \end{aligned}$$
（4）

Thus, the optimization problem for the support vector regression machine could be written as:

$$\min_{w,\, b,\, \xi_n,\, \hat{\xi}_n} C\sum_{n=1}^{N}\left(\xi_n+\hat{\xi}_n\right)+\frac{1}{2}\| w \|^2$$
（5）

$$\begin{aligned} s.t. \quad & t_n \le y\left(x_n\right)+\varepsilon+\xi_n \\ & t_n \ge y\left(x_n\right)-\varepsilon-\hat{\xi}_n \\ & \xi_n \ge 0,\ \hat{\xi}_n \ge 0,\ n=1,\ \ldots,N \end{aligned}$$
（6）

Similarly, the constraints were brought into the objective function using the Lagrangian multiplier method:

$$\begin{aligned} L = {} & C\sum_{n=1}^{N}\left(\xi_n+\hat{\xi}_n\right)+\frac{1}{2}\|\omega\|^2-\sum_{n=1}^{N}\left(\mu_n\xi_n+\hat{\mu}_n\hat{\xi}_n\right) \\ & -\sum_{n=1}^{N}a_n\left(\varepsilon+\xi_n+y_n-t_n\right)-\sum_{n=1}^{N}\hat{a}_n\left(\varepsilon+\hat{\xi}_n-y_n+t_n\right) \end{aligned}$$
（7）

Next, using the idea of solving the pairwise problem in SVM, we obtained：

$$\frac{\partial L}{\partial \omega}=0 \Rightarrow \omega=\sum_{n=1}^{N}\left(a_n-\hat{a}_n\right)\phi\left(x_n\right)$$
（8）

$$\frac{\partial L}{\partial b}=0 \Rightarrow \sum_{n=1}^{N}\left(a_n-\hat{a}_n\right)=0$$
（9）

$$\frac{\partial L}{\partial b}=0 \Rightarrow \sum_{n=1}^{N}\left(a_n-\hat{a}_n\right)=0$$
（10）

$$\frac{\partial L}{\partial b}=0 \Rightarrow \sum_{n=1}^{N}\left(a_n-\hat{a}_n\right)=0$$
（11）

Using these results to eliminate the corresponding variables in the Lagrangian function, the pairwise problem for SVR was obtained as follows:

$$\max_{\alpha,\ \hat{\alpha}} \tilde{L}(\alpha,\ \hat{\alpha}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(x_n,\ x_m) - \varepsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) + \sum_{n=1}^{N}(a_n - \hat{a}_n)t_n \quad （12）$$

$$s.t.\quad \sum_{n=1}^{N}(a_n - \hat{a}_n) = 0,\ where\ 0 \le a_n,\ \hat{a}_n \le C \quad （13）$$

The final SVR regression model obtained was:

$$y(x) = \sum_{n=1}^{N}(a_n - \hat{a}_n)k(x,\ x_n) + b \quad （14）$$

This experiment used radial basis function (RBF) kernel in the support vector regression method, set 0.001 as the tolerance when stopping the judgment, set the value of epsilon to 0.1, and removes the hard limit of iteration within the solver.

3.3.Neural Network Models

The neural network model consists of three parts, the input layer for the original independent variables, the hidden layer for the transformation of intermediate variables, and the output layer for the prediction results. Neurons between layers in a neural network are interconnected, and the number of layers of hidden layers is not fixed. The neural network model is continuously optimized under the stimulation of external inputs and calibrated outputs, so that the output is constantly close to the desired output. Three hidden layers with 32, 16 and 8 neurons were set up in this experiment. In terms of the nonlinear activation function, this experiment instead used the same activation function on all nodes, called the rectified linear unit (ReLU), defined as:

$$\mathrm{ReLU}(x) = \begin{cases} 0, & if\ x < 0 \\ x, & otherwise \end{cases} \quad （15）$$

Compared with the sigmoid function and tanh function, the ReLu activation function supports faster derivative calculation, which can make the network training faster. In addition, increasing the nonlinearity of the network can prevent the gradient from vanishing at the same time. And the network was trained with an alpha of 0.01 and a stochastic gradient descent based optimizer (adam) proposed by Kingma, Diederik and Jimmy Ba. BP neural networks are processed in a parallel distributed manner, which makes them fault tolerant and adaptive. The network can implement complex nonlinear mapping functions and converge to arbitrary nonlinear continuous functions with arbitrary accuracy. In contrast, the pattern of stock market returns tends to be nonlinear, making it suitable for exploring the underlying patterns within the stock market.

4.EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Data

4.1.1. Data source

This experiment studied more than 1,000 stocks on the New York Stock Exchange (NYSE) from June 2016 to June 2021. The experimental predictors were divided into two categories: (1) financial data of listed companies and (2) market data of companies. The financial data included ROE (deduction/dilution), ROA, net profit/gross operating income, operating profit/total operating income, EBIT/gross operating income, debt-to-assets ratio, current ratio, quick ratio, total assets turnover, and market data include monthly average volume, monthly average amplitude, and monthly average turnover rate. This experiment explored the relationship between the above 12 predictors and the next month's stock returns and compares the predicted portfolio returns to the Dow Jones Index (DJIA). We presented the full names, abbreviations, and categories of all predictors and indices in Table 1. The data used in the experiments are obtained from the Wind financial database.

Table 1. Name and classification of indicators

Category	Full name of indicators	Abbreviation for indicators
Fundamental factors	Return on equity(deduction/dilution)	roe_exdiluted
	Return on total assets	roa
	Net profit to gross operating income ratio	profittogr
	Operating profit to total operating income ratio	optogr
	Ebit to gross operating income ratio	ebittogr
	Debt-to-assets ratio	debttoassets
	Current ratio	current
	Quick ratio	quick
	Total assets turnover	assetsturn
Market factors	Monthly average volume	avgvolpmon
	Monthly average amplitude	avgamppmon
	Monthly average turnover rate	avgturnpmon
Index	Dow Jones Industria Average	DJIA

4.1.2. Data pre-processing

In the data pre-processing stage, this experiment processed the experimental data in three steps.

The first step was missing value processing. Since the period of the stock experiment data was from June 2016 to June 2021, and some companies listed on NYSE after June 2016, thus resulting in a data gap at the pre-IPO time point, this gap needed to be ignored when training the model. Also, as financial data some companies could not be updated monthly on the Wind database, there were gaps in the financial data for some months. Therefore, the financial data from the last update needed to be carried forward unchanged to the month in which the data was missing. Stocks with missing factor values need to be removed.

The second step was the depolarization process. In this experiment, the quantile depolarization method was used to remove values greater than the top 5% of the fractional loci as well as values less than the bottom 5%.

The third step was data conversion. This experiment used the Min-Max normalization process to normalize all factor data to interval $^{[0,1]}$. In addition, this experiment mapped the return of the stock of next month to an exponential function form

$$f : x \mapsto e^{x} \qquad (16)$$

4.2. Comparison of predictive power of machine learning algorithms

In this section, this experiment analyzed and compared the future performance of the portfolios selected by each machine learning algorithm. All the data from June 2016 to December 2020 was used as the training set and all the data from January 2021 to June 2021 was used as the prediction set, the following results were obtained. The graph on the left shows the actual future performance of the portfolios finally selected by various methods versus the actual rise and fall of the Dow Jones Index (DJIA) over the same period in the future, and the graph on the right shows the occurrence of the 10 best-performing stocks counted after multiple selections to train the model.

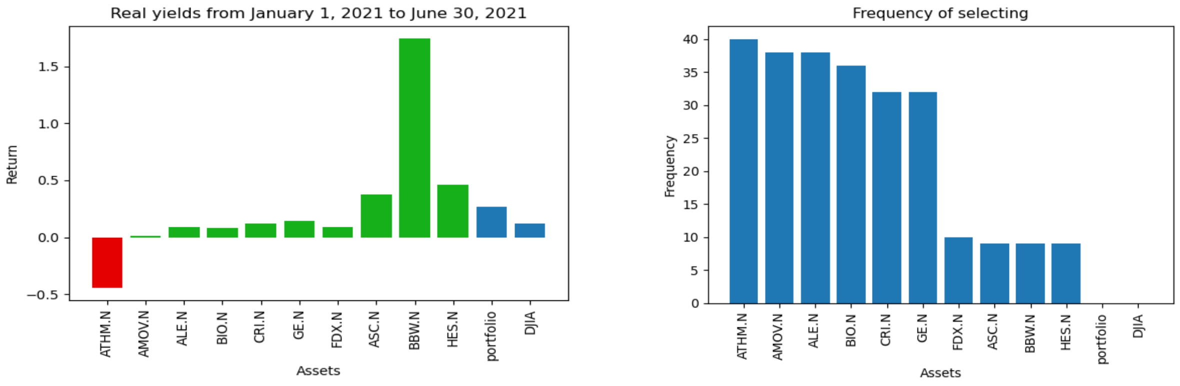


Figure 1. Portfolio prediction results using Random Forest (RF) and actual results

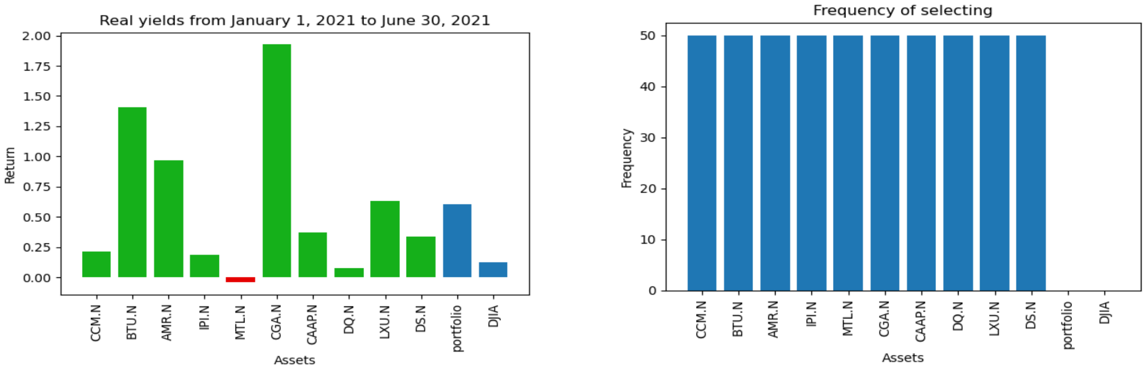


Figure 2. Portfolio prediction results using support vector regression (SVR) and actual results

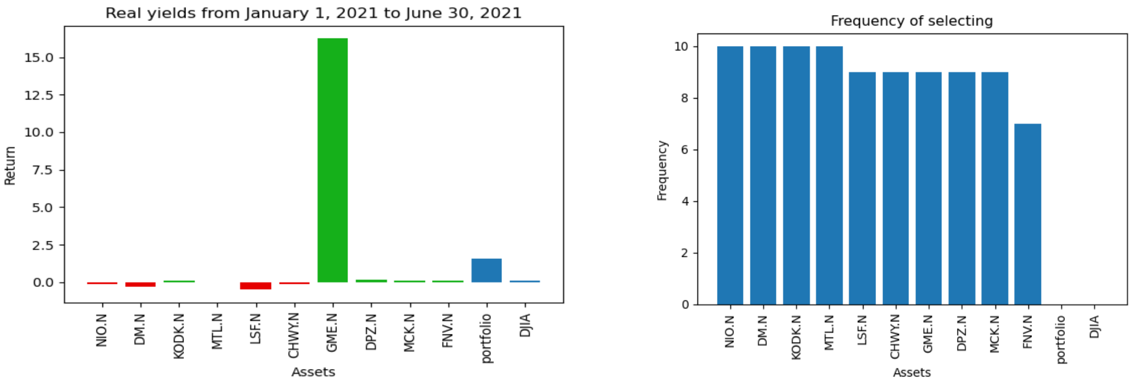


Figure 3. Portfolio prediction results using neural networks (NN) and actual results

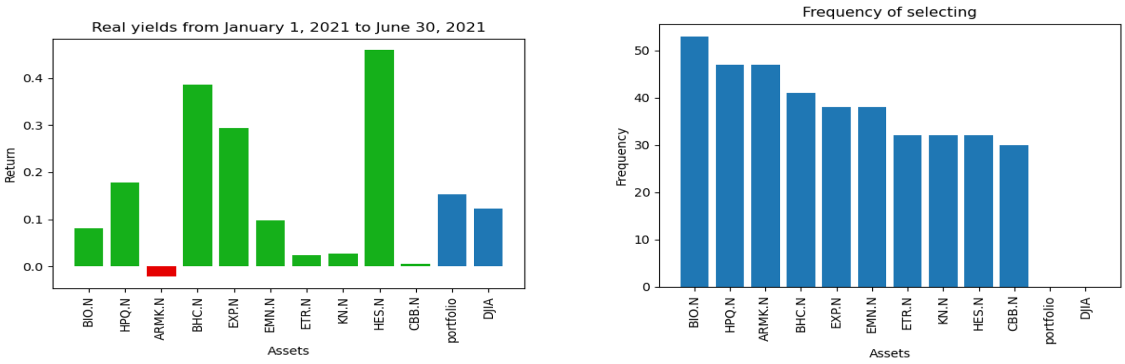


Figure 4. Portfolio forecast results using Decision Tree and actual results

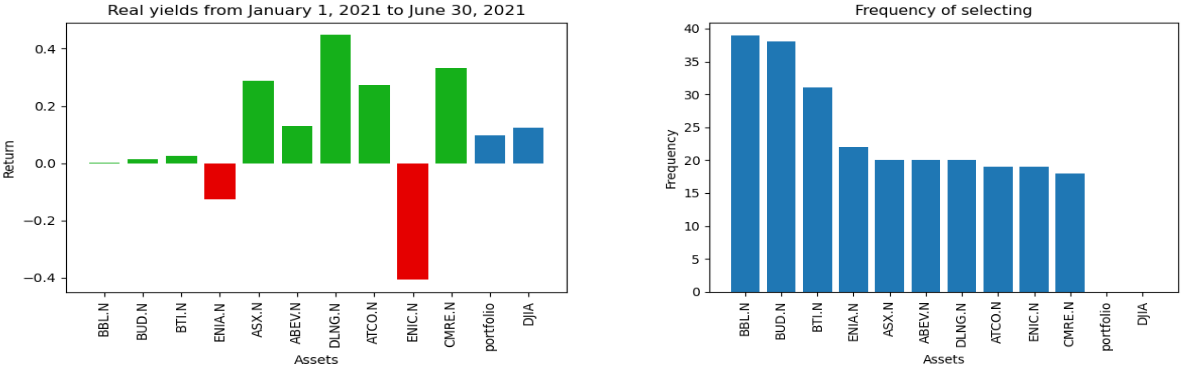


Figure 5. Portfolio prediction results using Adaboost regression and actual results

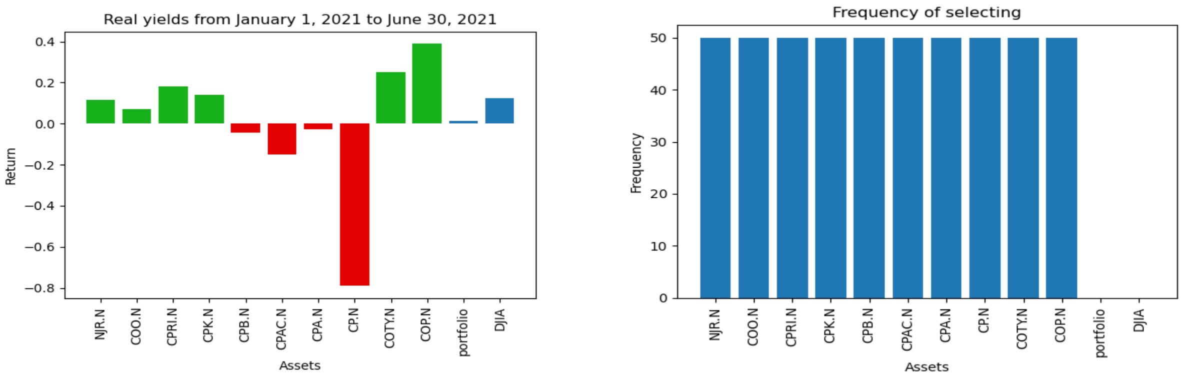
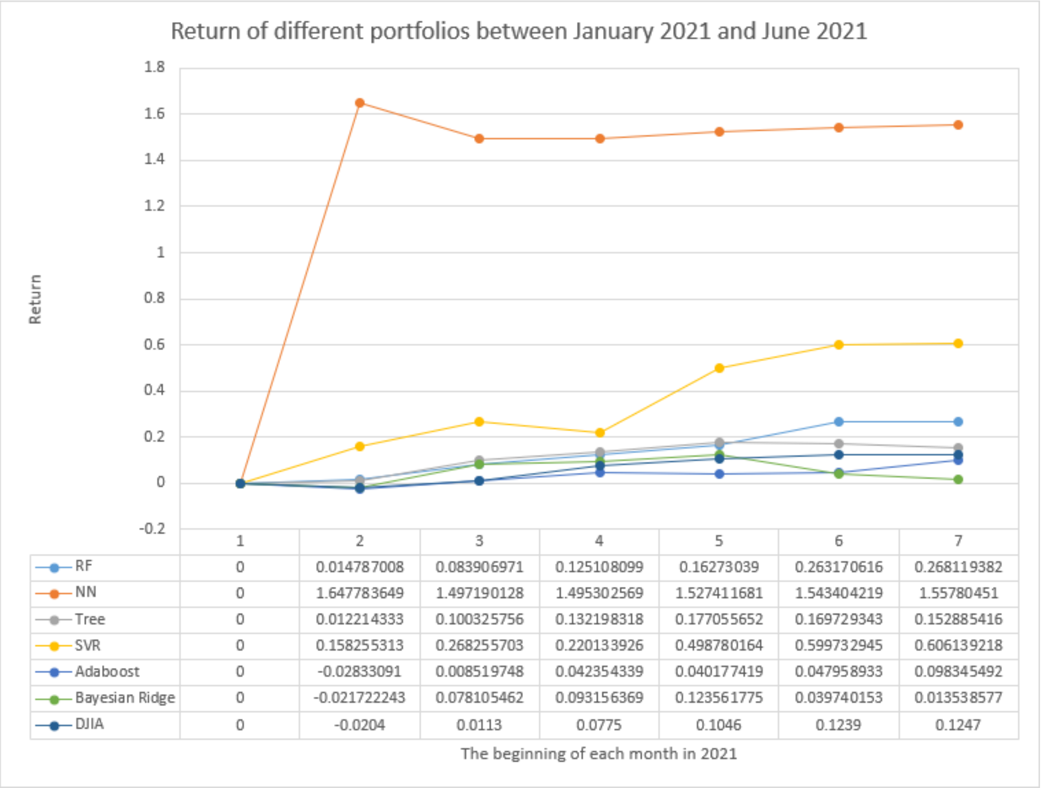


Figure 6. Portfolio prediction results using Bayesian Ridge regression and actual results



	1	2	3	4	5	6	7
RF	0	0.014787008	0.083906971	0.125108099	0.16273039	0.263170616	0.268119382
NN	0	1.647783649	1.497190128	1.495302569	1.527411681	1.543404219	1.55780451
Tree	0	0.012214333	0.100325756	0.132198318	0.177055652	0.169729343	0.152885416
SVR	0	0.158255313	0.268255703	0.220133926	0.498780164	0.599732945	0.606139218
Adaboost	0	-0.02833091	0.008519748	0.042354339	0.040177419	0.047958933	0.098345492
Bayesian Ridge	0	-0.021722243	0.078105462	0.093156369	0.123561775	0.039740153	0.013538577
DJIA	0	-0.0204	0.0113	0.0775	0.1046	0.1239	0.1247

Figure 7. Actual performance of the portfolios in the first half of 2021

The Dow rose 12.31% in the first half of 2021. The above graph shows that the portfolios selected by Random Forest, Support Vector Regression, Neural Network, and Decision Tree outperformed the Dow, reaching 26.81%, 60.61%, 155.78%, and 15.29% real returns, respectively, while the portfolios selected by Adaboost Regression and Bayesian Ridge Regression performed worse, with real returns of 9.83% and 1.35%.

In terms of actual performance, the neural network (NN) selected the portfolio with the highest actual return. The ten stocks with the highest returns all appeared between seven and ten times in the ten replicate experiments of the neural network model, which indicates the strong convergence of the network. One stock in the selected portfolio had a return of 1,626.65%, while the others had an average performance and even four stocks had a negative return. This indicated that the neural network may fall into local extremes while learning and fail to achieve optimal results. In addition, due to the "sawtooth phenomenon" of the algorithm and the complexity of the objective function to be optimized, the algorithm ran more inefficiently than other algorithms.

Eight of the stocks in the portfolio selected by the support vector regression outperformed the Dow and had better overall return performance. Compared to other methods, support vector regression gives better results than other algorithms on small sample training sets. In addition, the optimization of the support vector regression reduced the structed risk, thus avoiding the overfitting problem. The results of the other four algorithms were worse compared to the above two algorithms. For loud-noise real market environments, the random forest algorithm and decision tree algorithm training models may be over-fitted, which in turn leads to model instability, higher randomness of portfolio selection, and therefore higher investment risk.

In contrast, two emerging machine learning algorithms, Adaboost regression, and Bayesian ridge regression performed poorly. The actual portfolio returns only reached 9.83% and 1.35%, and both ended the period with lower returns than the Dow Jones Index. Looking at the entire first half of the year, Adaboost's return was lower than the actual return of the Dow Index for the entire period. The Bayesian Ridge regression portfolio had a yield curve that was significantly higher than the index in the short term (at the end of February, March, and April) but had a significant drop after May.

5.CONCLUSIONS

From the results of the study, machine learning methods were promising for predicting stock returns in the FinTech space. It could help investors solve the practical investment problems of stock selection, timing, and risk management. The portfolios selected by two traditional machine learning models, support vector regression and neural networks, were effective in the recent NYSE market environment. They outperformed the Dow and the results of model training are more stable, which was superior in practice. In contrast, for random forest and decision tree methods, the training results were more unstable, more random, and more biased due to the presence of loud noise in the real market and the accompanying phenomenon of overfitting of the model. Therefore, the investment risk was also higher. The two emerging machine learning algorithms, Adaboost and Bayesian Ridge were more difficult to find a set of effective, context-adapted parameters because the parameters were more complex compared to traditional models. The parameters used in this experiment failed to meet the requirements well, resulting in poor model results. In addition, this paper leaves an important issue: when applied in practice, transaction costs, dynamic environment, etc. need to be considered. The adjustment period and criteria for parameter types and values are difficult to generalize using a particular method. Feature capture and model optimization is still an important direction for future supplementary research.

REFERENCES

[1] Markowitz, H.M. (March 1952). "Portfolio Selection". The Journal of Finance. 7 (1): 77–91. doi:10.2307/2975974. JSTOR 2975974.

[2] Sharpe, William F. (1964). "Capital Asset Prices – A Theory of Market Equilibrium Under Conditions of Risk". Journal of Finance. XIX (3): 425–442. doi:10.2307/2977928. hdl:10.1111/j.1540-6261.1964.tb02865.x. JSTOR 2977928.

[3] The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets, John Lintner, 1965, Review of Economics and Statistics. 47:1, pp. 13–37.

[4] "Equilibrium in a Capital Asset Market", Econometrica, 34, 1966, pp. 768–783.

[5] “A fast and elitist multiobjective genetic algorithm: NSGA-IIK”, Deb, A Pratap, S Agarwal, T MeyarivanIEEE transactions on evolutionary computation 6 (2), 182-197, 2002

[6] Anagnostopoulos, K.P. & Mamanis, Georgios. (2010). A portfolio optimization model with three objectives and discrete variables. Computers & Operations Research. 37. 1285-1297. 10.1016/j.cor.2009.09.009.

[7] Ye Wang, Bo Wang, Xinyang Zhang,A New Application of the Support Vector Regression on the Construction of Financial Conditions Index to CPI Prediction, Procedia Computer Science, Volume 9, 2012, Pages 1263-1272,

[8] Felipe Dias Paiva, Rodrigo Tomás Nogueira Cardoso, Gustavo Peixoto Hanaoka, Wendel Moreira Duarte, Decision-making for financial trading: A fusion approach of machine learning and portfolio selection, Expert Systems with Applications, Volume 115, 2019,Pages 635-655.

[9] Rapach, David & Zhou, Guofu, 2013."Forecasting Stock Returns," Handbook of Economic Forecasting, in: G. Elliott & C. Granger & A. Timmermann (ed.),Handbook of Economic Forecasting, edition 1, volume 2, chapter 0, pages 328-383.

[10] RAPACH, D. E., STRAUSS, J. K., & ZHOU, G. (2013). International Stock Return Predictability: What Is the Role of the United States? The Journal of Finance, 68(4), 1633–1662.

[11] Harvey, Campbell R. and Liu, Yan, Detecting Repeatable Performance (January 21, 2018).

[12] Li Bin, SHAO Xinyue, LI Yueyang. Research on fundamental quantitative investment driven by machine learning [J]. China Industrial Economics,2019(08):61-79.

[13] Jigar Patel, Sahil Shah, Priyank Thakkar, K Kotecha, Predicting stock market index using fusion of machine learning techniques, Expert Systems with Applications, Volume 42, Issue 4,2015, Pages 2162-2172

[14] Adil Moghar, Mhamed Hamiche, Stock Market Prediction Using LSTM Recurrent Neural Network, Procedia Computer Science, Volume 170, 2020, Pages 1168-1173, ISSN 1877-0509.

[15] Xiao-dan Zhang, Ang Li, Ran Pan, Stock trend prediction based on a new status box method and AdaBoost probabilistic support vector machine, Applied Soft Computing, Volume 49, 2016, Pages 385-398, ISSN 1568-4946.