Communications for Statistical Applications and Methods
2024, Vol. 31, No. 6, 677–701

https://doi.org/10.29220/CSAM.2024.31.6.677
Print ISSN 2287-7843 / Online ISSN 2383-4757

# Optimal stock investment strategy using prediction models

Jimin Kim[a], Jongwoo Song[1,a]

[a]Department of Statistics, Ewha Womans University, Korea

**Abstract**

Stock price prediction has traditionally been known as a challenging task. However, recent advancements in machine learning and deep learning models have spurred extensive research in predicting stock returns. This study applies these predictive models to U.S. stock data to forecast stock returns and develop investment strategies based on these forecasts. Additionally, the performance of the model-based investment strategy was compared with that of a widely recognized method, market capitalization-weighted investing. The results indicate that, overall, market capitalization-weighted investing outperformed model-based investing. However, the highest returns were observed in the model-based strategy. It was also found that model-based investing exhibits higher volatility in returns, with significant disparities between years of high and low returns. While investing through machine learning methodologies may be attractive to investors seeking high risk and high return, market capitalization-weighted investing is likely more suitable for those desiring stable returns.

Keywords: asset pricing, machine learning, investment strategy, prediction model, dimension reduction

## 1. Introduction

According to a survey conducted by Jeffrey (2023), the percentage of American adults who have invested in the stock market reached 61%, the highest level since 2008. As interest in the stock market continues to rise, the use of predictive models for stock investment analysis has emerged as a crucial research topic. Predictive models, which utilize historical stock price data and various variables, serve as tools to forecast market directions and provide valuable information for making investment decisions.

Research on predicting the stock market has been conducted in various forms historically. One of the most conventional techniques is the ARIMA model, a statistical method for time series forecasting that is particularly effective for short-term predictions (Ariyo *et al*., 2014). Text mining techniques have also been applied to stock price prediction. Fung *et al*. (2003) proposed using text documents to implement predictive models, based on the premise that news articles indirectly impact the stock market. Additionally, Mittal and Goel (2012) attempted sentiment analysis using social media data to gauge public opinion and analyze stock market movements.

The introduction of machine learning techniques has marked a significant advancement in the financial sector. Gu *et al*. (2020) utilized various machine learning techniques such as linear regression, random forest, and neural networks for predicting stock returns. Shen *et al*. (2012) and Dey *et al*.

[1] Corresponding author: Department of Statistics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: josong@ewha.ac.kr

Published 30 November 2024 / journal homepage: http://csam.or.kr
© 2024 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

(2016) respectively applied support vector machine (SVM) and Xtreme gradient boost (XGBoost) methods to create predictive models. Unsupervised learning techniques have also been applied to stock price prediction; Bini and Mathew (2016) used clustering methods such as $K$-means and EM algorithm.

Research involving deep learning models is also actively ongoing. Chen *et al*. (2024) attempted stock price prediction using deep learning models like feedforward network and generative adversarial network, enhancing prediction performance by extracting hidden states from macroeconomic data using long short-term memory (LSTM) methods. Ding *et al*. (2015) employed five predictive models based on convolutional neural networks (CNN). Some studies have combined multiple methodologies; Kim and Han (2000) proposed an approach that applies genetic algorithms (GA) to artificial neural networks (ANN) to reduce complex dimensions and noise.

This paper aims to present a method for predicting stock returns using predictive models. By using firm-specific data and macroeconomic data, and applying appropriate dimensionality reduction techniques to each, we aim to enhance the performance of predictive models. Machine learning algorithms such as CatBoost, XGBoost, and LightGBM will be implemented to create models, with the best-performing model selected for return prediction. Additionally, this study will construct a portfolio of companies with the highest predicted monthly returns to identify the optimal investment strategy. This research is expected to play a significant role in developing efficient investment strategies in the future stock market.

Chapter 2 describes the data, explaining the firm-specific and macroeconomic data and how these datasets are combined to construct a monthly dataset. Chapter 3 explains predictive models for monthly stock returns, applying PCA for dimensionality reduction on firm-specific data and LSTM algorithms for macroeconomic data. CatBoost, XGBoost, and LightGBM machine learning algorithms are used to implement predictive models and forecast returns. Chapter 4 describes the construction of a portfolio based on companies with high predicted returns and presents various investment strategies and outcomes. Finally, Chapter 5 summarizes our findings and provides conclusions.

## 2. Data description

Typically, data used for stock price prediction include firm specific data, which provides characteristics of individual companies, and Macroeconomic data, which reflects the overall market conditions. There have been attempts to analyze stock prices using other types of data as well. For instance, Jiang (2021) collected information about companies through text data from web search results and image data from CCTV footage. However, this paper aims to construct a model for predicting stock returns using firm specific data and macroeconomic data.

Given our goal of predicting monthly stock returns, all data will be aggregated on a monthly basis. However, the observed periods for the collected macroeconomic data and firm specific data vary, including monthly, quarterly, semiannual, and annual frequencies. To align all variables to a monthly frequency, we used the most recent value for variables that are not available monthly. Furthermore, to enhance the training speed and performance of the model, all variables were standardized.

We collected 113 Macroeconomic data variables from the FRED-MD database, referencing McCracken and Ng (2016). Variables that were discontinued or had incomplete data from 1997 to 2021 were excluded. The Macroeconomic variables are categorized into eight groups: Output and income, labor market, housing, consumption, orders and inventories, money and credit, interest and exchange rates, prices, and stock market. Detailed descriptions of the macroeconomic data can be found in Appendix A.

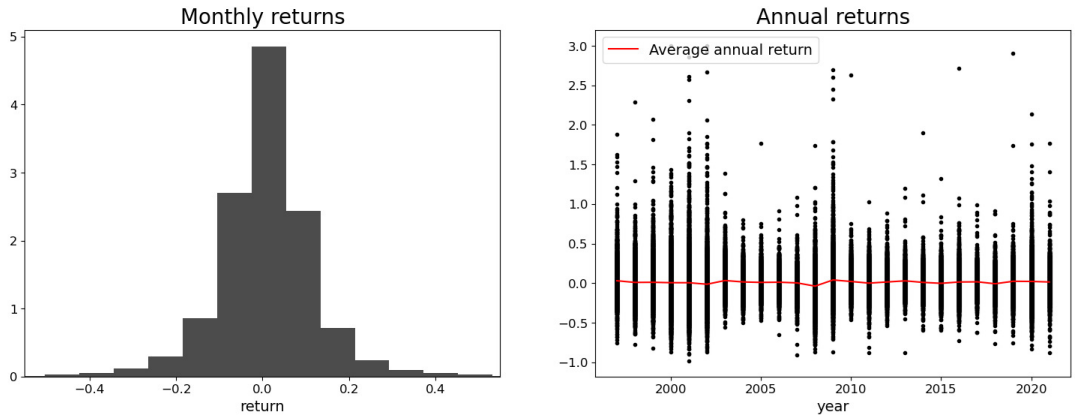


Figure 1: *Monthly and annual returns of the stocks.*

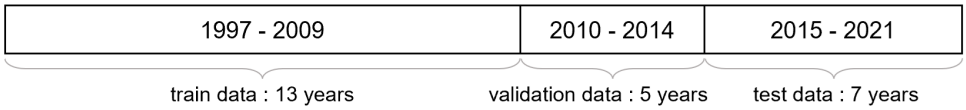


Figure 2: *Train, validation, and test data construction.*

Firm Specific data was obtained from the CRSP and Compustat databases. The data used in the analysis spans 25 years, from January 1997 to December 2021, and includes only stocks of companies listed on the NYSE, AMEX, and NASDAQ for more than one year. Variables regarding company information were referenced from the studies by Green *et al*. (2017) and Cong *et al*. (2022). There are 51 firm-specific characteristics, categorized into six groups: Investment, intangibles, profitability, value-versus-growth, momentum, and frictions. A list and detailed descriptions of these variables are provided in Appendix B.

Since our ultimate goal is to develop an investment strategy, companies with excessively low market capitalization, which are less reliable for investment, were excluded. Therefore, we collected data only for the top 1000 companies by market capitalization each year over the 25-year period. The resulting dataset includes approximately 4200 companies. Among these are companies like GOOG and AMZN, which entered the top 1000 by annual market capitalization in the 2000s and have maintained high market capitalization through 2021. Conversely, companies like ENE and LEH, which fell out of the top 1000 by annual market capitalization in the 2000s, are also included.

The response variable is the monthly return of each company's stock price compared to the previous month. The monthly return can be calculated as follows:

$$r_{k,t} = \frac{S_{k,t} - S_{k,t-1}}{S_{k,t-1}}, \tag{2.1}$$

where,

$r_{k,t}$ : Return on stock $k$ at time $t$

$S_{k,t-1}$ : Price of stock $k$ at time $t-1$

$S_{k,t}$ : Price of stock $k$ at time $t$

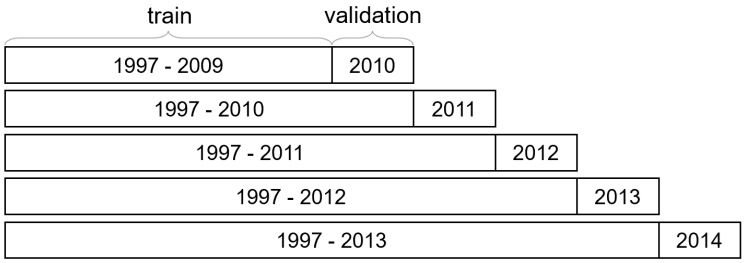


Figure 3: *Train and validation sets.*

The left graph in Figure 1 presents a histogram of returns for the entire dataset. Most of the return values are relatively small, primarily ranging between $-0.2$ and 0.2. The right graph is a scatter plot of yearly returns, with the red line representing the annual average return. The mean value of returns converges to nearly zero, with little variation across years.

Next, we combine the firm specific data and Macroeconomic data based on each month to create a unified dataset. The complete 25-year dataset is divided into 13 years of training data, 5 years of validation data, and 7 years of test data, as illustrated in Figure 2. This division is utilized for the development and evaluation of the predictive model.

## 3. Models

This section describes the preprocessing of variables and the various models used for stock price prediction. As previously mentioned, the dataset contains numerous explanatory variables, so appropriate dimensionality reduction is expected to improve model performance. Subsequently, we will perform stock prediction analysis using various machine learning models.

### 3.1. Dimension reduction

Economic indicators and firm-specific metrics can be highly correlated. Additionally, an excessive number of explanatory variables can impair the performance of prediction models. Dimensionality reduction of explanatory variables is an effective way to address these issues. Therefore, we will reduce the dimensions of the 113 Macroeconomic data variables and the 51 firm specific data variables and compare the performance of the models with and without dimensionality reduction.

First, we reduce the Macroeconomic data. Since Macroeconomic data is time-series data, we apply the method proposed by Chen *et al*. (2024). We use the long short-term memory (LSTM) method to extract hidden states from the time-series data. In contrast, firm specific data does not have continuous values for all companies as it consists of the top 1000 companies' data collected annually. This means the LSTM method cannot be applied. Therefore, we use the most traditional dimensionality reduction method, principal component analysis (PCA), as referenced by Zhong and Enke (2017).

It is unclear how many components should be retained to achieve the best performance for the prediction model. Therefore, we establish prediction models using both the full dataset and datasets reduced to specific numbers of components. The Macroeconomic data is reduced to 4, 10, 20, and 50 components from the original 113 variables. The firm specific data is reduced to 4, 10, and 20 components from the original 51 variables. Additionally, we include the case where no dimensionality reduction is applied to both datasets, resulting in a total of 20 different dataset combinations.

Table 1: R-squared values for the machine learning models

| Firm specific data dimension | Macroeconomic data dimension | Machine learning methodology | | |
|---|---|---|---|---|
| | | CatBoost | XGBoost | LightGBM |
| 4 | 4 | 0.0403 | 0.0389 | 0.0253 |
| | 10 | 0.0280 | 0.0199 | 0.0230 |
| | 20 | 0.0344 | 0.0405 | 0.0027 |
| | 50 | 0.0336 | 0.0104 | −0.0384 |
| | 113 | 0.0372 | 0.0227 | 0.0224 |
| 10 | 4 | 0.0354 | **0.0476** | 0.0249 |
| | 10 | 0.0269 | 0.0211 | 0.0206 |
| | 20 | 0.0366 | 0.0379 | −0.0144 |
| | 50 | 0.0327 | 0.0109 | −0.0747 |
| | 113 | 0.0385 | 0.0177 | 0.0223 |
| 20 | 4 | 0.0367 | 0.0388 | 0.0261 |
| | 10 | 0.0273 | 0.0211 | 0.0237 |
| | 20 | 0.0336 | 0.0369 | −0.0283 |
| | 50 | 0.0295 | 0.0104 | −0.0607 |
| | 113 | 0.0313 | 0.0145 | 0.0227 |
| 51 | 4 | **0.0439** | **0.0420** | 0.0231 |
| | 10 | 0.0311 | 0.0225 | 0.0205 |
| | 20 | 0.0309 | 0.0382 | 0.0004 |
| | 50 | 0.0294 | 0.0188 | −0.0380 |
| | 113 | 0.0312 | 0.0084 | −0.0221 |

Table 2: R-squared values for the better machine learning models

| Candidate 1 | | Macroeconomic data dimension | | | | |
|---|---|---|---|---|---|---|
| Method : XGBoost | | 2 | 3 | 4 | 5 | 6 |
| Firm specific data dimension | 8 | **0.0500** | 0.0293 | 0.0456 | 0.0251 | 0.0348 |
| | 9 | 0.0389 | 0.0285 | 0.0445 | 0.0252 | 0.0352 |
| | 10 | 0.0486 | 0.0329 | 0.0476 | 0.0249 | 0.0333 |
| | 11 | 0.0390 | 0.0287 | 0.0377 | 0.0265 | 0.0311 |
| | 12 | 0.0413 | 0.0283 | 0.0426 | 0.0249 | 0.0344 |
| Candidate 2 | | Macroeconomic data dimension | | | | |
| Method : CatBoost | | 2 | 3 | 4 | 5 | 6 |
| Firm specific data dimension | 49 | 0.0507 | 0.0470 | 0.0365 | 0.0384 | 0.0418 |
| | 50 | **0.0533** | 0.0478 | 0.0393 | 0.0410 | 0.0382 |
| | 51 | 0.0516 | 0.0495 | 0.0439 | 0.0341 | 0.0437 |
| Candidate 3 | | Macroeconomic data dimension | | | | |
| Method : XGBoost | | 2 | 3 | 4 | 5 | 6 |
| Firm specific data dimension | 49 | 0.0442 | 0.0309 | 0.0411 | 0.0237 | 0.0333 |
| | 50 | 0.0437 | 0.0318 | 0.0402 | 0.0236 | 0.0324 |
| | 51 | 0.0340 | **0.0446** | 0.0420 | 0.0228 | 0.0285 |

## 3.2. Machine learning models

There are a total of 20 explanatory variable set combinations. The machine learning methodologies employed include three boosting algorithms: CatBoost (Prokhorenkova *et al*., 2018), XGBoost (Chen and Guestrin, 2016), and LightGBM (Ke *et al*., 2017). Boosting algorithms enhance predictive accuracy by combining multiple weak learners and are effective in capturing complex data patterns.

For each set combination, a machine learning model is applied, and the R-Square value is calculated by iteratively training and validating the model on the train and validation datasets. The method used to create the train and validation sets during model training is illustrated in Figure 3. Since the stock market data is time series data, cross-validation is not performed as with general datasets.

Table 3: Candidate models

| Result | Methodology | Firm specific data dimension | Macroeconomic data dimension | R-squared value |
|---|---|---|---|---|
| Candidate 1 | XGBoost | 8 | 2 | 0.0500 |
| Candidate 2 | CatBoost | 50 | 2 | 0.0533 |
| Candidate 3 | XGBoost | 51 | 3 | 0.0446 |

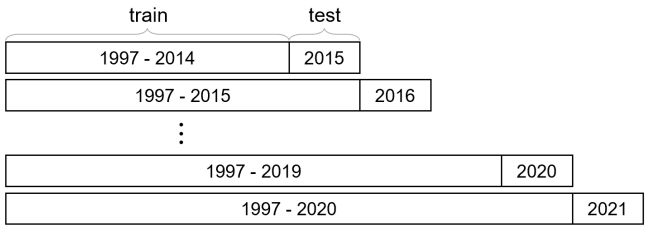


Figure 4: *Train and test sets.*

Table 4: R-squared values for test data

| | Candidate 1 | Candidate 2 | Candidate 3 |
|---|---|---|---|
| R-squared value | −0.0025 | −0.0132 | 0.0071 |

Instead, the train data precedes the validation data to reflect the temporal order, with the train data representing the past and the validation data representing the future.

For each individual set and model, we identify the optimal tuning parameters that yield the best performance on the validation data. We fit the model to the optimal parameters annually and calculate the R-squared values over a five-year period. We then compute the average to obtain the final R-squared values for each set and model. The training results are presented in Table 1.

Here are the results of applying three machine learning methodologies–CatBoost, XGBoost, and LightGBM–to a total of 20 combinations of datasets. The LightGBM algorithm showed relatively lower model performance compared to the other two algorithms. While the performance differences for firm specific data were not significant across different numbers of reduced components, the macroeconomic data generally performed better with fewer reduced components. The highlighted sections in the table represent the top three R-squared values, all of which are in the 0.04 range.

The current number of components used was arbitrarily selected, so we cannot be certain that these are the optimal results. Therefore, we will treat the three highlighted combinations and their corresponding models as the primary candidates. We will then adjust the number of reduced components slightly and attempt model fitting again to identify the combination with the best performance.

Table 2 presents the results of the second attempt for the three candidate models. Performance improvements were observed across all candidate models, with the best-performing scenario for each model detailed in Table 3. These three candidates will be used to fit the test data.

## 3.3. Selections of the portfolio

The model fitting method is the same as the one used in model validation. As illustrated in Figure 4, the data is split into training and test sets. After predicting one year of data, the model is updated, and the training data is extended to forecast the next year (Gu *et al.* 2020). Using this method, we will predict the monthly returns of companies for a total of seven years, from 2015 to 2021, which

Table 5: Annual returns for the top 10 companies by market capitalization

| Date | Equal weight investment returns | Capital weighted investment returns | Equal weight investment return rate | Capital weighted investment return rate |
|---|---|---|---|---|
| 2015-12 | 10319.99 | 10409.57 | 3.20% | 4.10% |
| 2016-12 | 10857.53 | 10833.66 | 5.21% | 4.07% |
| 2017-12 | 14024.62 | 14416.59 | 29.17% | 33.07% |
| 2018-12 | 15114.26 | 15961.53 | 7.77% | 10.72% |
| 2019-12 | 17786.10 | 19509.37 | 17.68% | 22.23% |
| 2020-12 | 22993.54 | 27974.38 | 29.28% | 43.39% |
| 2021-12 | 32914.64 | 39329.08 | 43.15% | 40.59% |
| Average | - | - | 19.35% | 22.60% |

Table 6: Annual returns for the top 10 companies by model

| Date | Equal weight investment returns | Capital weighted investment returns | Equal weight investment return rate | Capital weighted investment return rate |
|---|---|---|---|---|
| 2015-12 | 9260.85 | 11145.88 | −7.39% | 11.46% |
| 2016-12 | 19520.83 | 13714.30 | 110.79% | 23.04% |
| 2017-12 | 24676.16 | 16597.54 | 26.41% | 21.02% |
| 2018-12 | 20143.74 | 16007.01 | −18.37% | −3.56% |
| 2019-12 | 20490.08 | 14868.16 | 1.72% | −7.11% |
| 2020-12 | 22512.25 | 17214.71 | 9.87% | 15.78% |
| 2021-12 | 32405.99 | 30540.53 | 43.95% | 77.41% |
| Average | - | - | 23.85% | 19.72% |

Table 7: Annual returns for the S&P 500 index

| Date | Investment returns | Investment return rate |
|---|---|---|
| 2015-12 | 10245.36 | 2.45% |
| 2016-12 | 11222.26 | 9.54% |
| 2017-12 | 13401.62 | 19.42% |
| 2018-12 | 12565.73 | −6.24% |
| 2019-12 | 16194.47 | 28.88% |
| 2020-12 | 18827.51 | 16.26% |
| 2021-12 | 23890.75 | 26.89% |
| Average | - | 13.89% |

corresponds to the test data period.

The predictive results are presented in Table 4. Overall, it is evident that the performance on the test data is low. We have selected the dataset and methodology of Candidate 3, which demonstrated the best performance, as our final dataset and model.

## 4. Investment strategies

When constructing an investment portfolio, two key decisions must be made: Which companies to invest in, and how much to invest in each company. Consequently, if investments are made in $n$ companies each month, the portfolio will be structured as follows.

$$P = w_1 x_1 + \cdots + w_n x_n, \tag{4.1}$$

where,

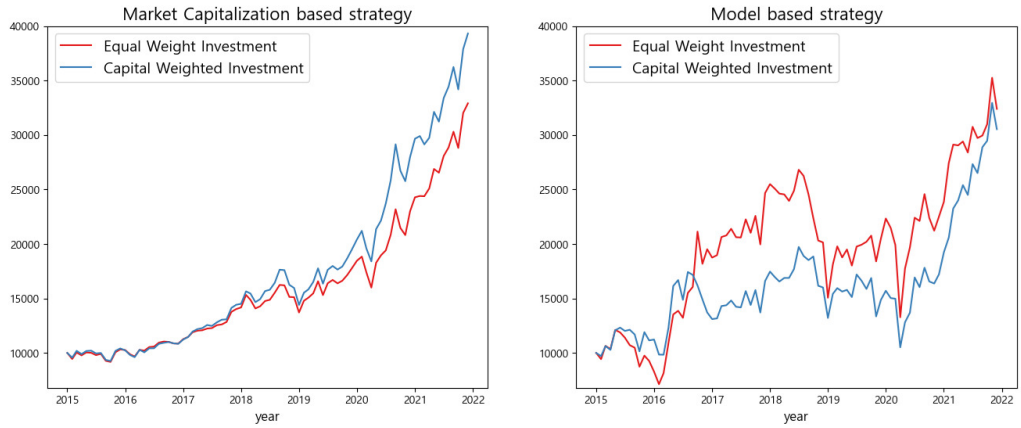


Figure 5: *Returns for the market capitalization based strategy and model based strategy.*

$P$ : current value of the portfolio.

$w_1, \ldots, w_n$ : weight ($\sum_{i=1}^{n} w_i = 1$).

$x_1, \ldots, x_n$ : stocks (companies).

The selection of companies to invest in each month is determined by two methods:

1. **Model based**: Using a return prediction model, select the top $n$ companies based on their forecasted returns for the month.

2. **Market capitalization based**: Select the top $n$ companies each month based on their market capitalization.

Next, two investment strategies are determined.

1. **Equal weight investment**:

$$w_i = \frac{1}{n}, \quad (i = 1, \ldots, n).$$

Allocate an equal amount of investment to each company. For example, if the initial investment is \$10,000 and $n = 10$, we invest \$1,000 in the stocks of each company.

2. **Capital weighted investment**:

$$w_i \propto \text{cap}(x_i).$$

Allocate investment funds in proportion to each company's market capitalization.

Each month, we select $x_1, \ldots, x_n$ companies using a return prediction model and market capitalization. Then, we determine the weight of each company using either the equal weight investment strategy or the capital weighted investment strategy.

Table 8: Final amounts and annual average returns for the market capitalization based strategy and model based strategy

| Data | Top $n$ | Equal weight investment final amount | Capital weighted investment final amount | Equal weight investment average return | Capital weighted investment average return | Equal weight investment SD of returns | Capital weighted investment SD of returns |
|---|---|---|---|---|---|---|---|
| Market capitalization based | 1 | 46204.58 | 46204.58 | 26.88% | 26.88% | 25.44% | 25.44% |
| | 3 | **63910.36** | 60704.41 | **31.78%** | 30.93% | 19.60% | 20.16% |
| | 5 | 44657.99 | 48642.81 | 24.87% | 26.61% | 16.27% | 18.01% |
| | 10 | 32914.64 | 39329.08 | 19.35% | 22.60% | 13.93% | 15.55% |
| | 20 | 26253.41 | 32795.00 | 15.07% | 19.01% | 8.12% | 11.20% |
| | 30 | 26035.60 | 31661.98 | 14.81% | 18.27% | 6.14% | 9.34% |
| | 50 | 25130.34 | 29672.48 | 14.31% | 17.20% | 7.43% | 9.61% |
| | 100 | 22996.76 | 27466.15 | 12.90% | 15.90% | 7.79% | 9.24% |
| Model based | 1 | 8175.87 | 8175.87 | 160.95% | 160.95% | 426.37% | 426.37% |
| | 3 | 15219.06 | 50287.84 | 23.19% | 44.85% | 66.21% | 75.75% |
| | 5 | 30305.08 | **69790.91** | 24.96% | **34.68%** | 47.91% | 27.82% |
| | 10 | 32405.99 | 30540.53 | 23.85% | 19.72% | 40.38% | 25.87% |
| | 20 | 25208.99 | 27150.40 | 16.16% | 16.76% | 22.27% | 18.65% |
| | 30 | 23101.52 | 29851.02 | 13.80% | 17.72% | 15.99% | 13.95% |
| | 50 | 23126.76 | 31240.41 | 13.52% | 18.43% | 13.56% | 13.67% |
| | 100 | 25642.19 | 26167.31 | 15.18% | 15.44% | 13.38% | 12.86% |

## 4.1. Model based investment versus market capitalization based investment

An investment portfolio was constructed by investing in 10 companies each month from January 2015 to December 2021. The performance of the prediction model was evaluated by comparing the results of investments in model based companies with those of investments in market capitalization based companies and the S&P 500 index. The outcomes of each investment strategy are presented in Tables 5 through 7. The initial investment amount was uniformly set at $10,000 for all comparisons.

Investing in market capitalization based companies yielded higher returns under the capital weighted investment strategy. In this case, the average annual return was 22.6%, with a final investment amount of $39,329.08. When investing in the S&P 500 index, the average annual return was 13.89%, with a final investment amount of $23,890.75. When investing in model based companies, the highest returns were achieved with an equal weight investment strategy, resulting in an average annual return of 23.85% and a final investment amount of $32,405.99.

The simulation graph in Figure 5 shows the comparison of investing in market capitalization based companies and model based companies, with investments made in 10 companies each month. For investments in market capitalization based companies, a description of the companies comprising the portfolio each year can be found in Table C.1 of Appendix C. For investments in model based companies, a description of the companies comprising the portfolio each year can be found in Table C.2 of Appendix C.

## 4.2. Optimal portfolio

Subsequently, we modified the number of companies invested in each month and analyzed the investment returns. The methodology remained consistent with the approach used to examine the investment returns for 10 companies each month. We summarized the investment results from 2015 to 2021 for the top 1, 3, 5, 20, 30, 50, and 100 companies each month, as shown in Table 8. We present the final amounts, the averages and the standard deviations of the annual returns. The annual investment results according to the number of companies invested in each month are detailed in Tables D.1 through D.8 of Appendix D.

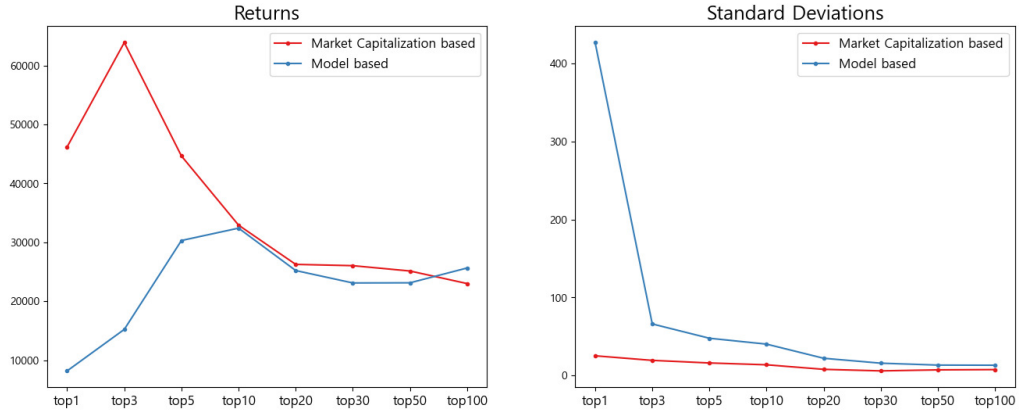


Figure 6: *Return and standard deviation for equal weight investment.*

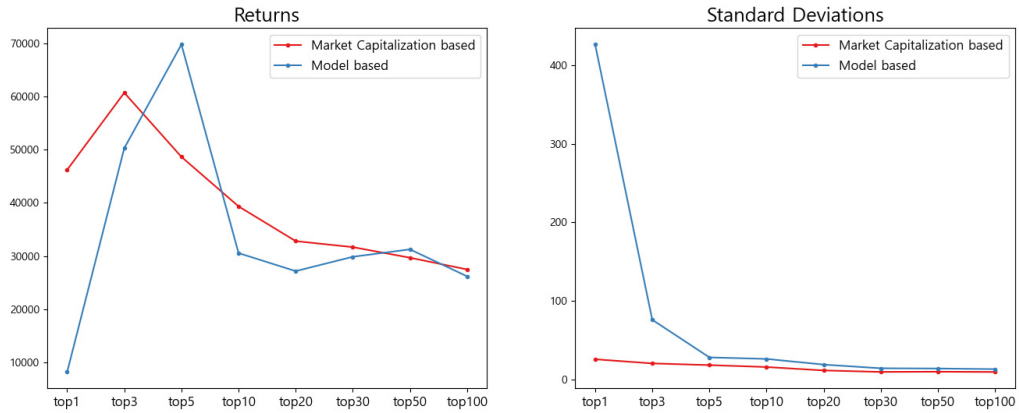


Figure 7: *Return and standard deviation for capital weighted investment.*

A closer examination of the table reveals that for investments in model based companies, the average annual return was highest at 160.95% when $n = 1$. However, the final amount was \$8,175.75, resulting in a loss. This outcome can be attributed to the high volatility observed in Appendix D, Table D.1, where the return in a particular year surged by 1,195.03%, while in other years, significant losses of −80.86% and −60.11% were recorded. This indicates that with the model-based strategy, a smaller number of companies invested in each month increases volatility, leading to a scenario where the average annual return may not correspond to the final asset value.

For investments in market capitalization based companies, the best result was achieved by investing in 3 companies each month with equal weight investment, resulting in a final amount of \$63,910.36 and an average return of 31.78%. For investments in model based companies, the best result was achieved by investing in 5 companies each month with capital weighted investment, resulting in a final amount of \$69,790.91 and an average return of 74.14%. Overall, as the number of companies invested in each month increased, the average returns decreased, and higher investment gains were achieved when investing in 10 or fewer companies.

We have illustrated the final investment amounts and the standard deviation of annual returns

Table 9: Most frequent companies selected by the best machine learning model

| Ticker | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | Total |
|---|---|---|---|---|---|---|---|---|
| RAD | - | - | 4 | 6 | - | - | - | 10 |
| ALT | - | - | - | 8 | - | - | - | 8 |
| WLL | 3 | 2 | - | - | 3 | - | - | 8 |
| CVEO | 7 | - | - | - | - | - | - | 7 |
| SPWR | - | 2 | 5 | - | - | - | - | 7 |
| ENDP | - | - | 7 | - | - | - | - | 7 |
| CLVS | - | 5 | - | 2 | - | - | - | 7 |

Table 10: Final amounts and annual average returns for market capitalization based strategy from 1997 to 2021

| Data | Top $n$ | Equal weight investment final amount | Capital weighted investment final amount | Equal weight investment average return | Capital weighted investment average return | Equal weight investment SD of returns | Capital weighted investment SD of returns |
|---|---|---|---|---|---|---|---|
| Market capitalization based | 1 | 76495.97 | 76495.97 | 12.26% | 12.26% | 28.86% | 28.86% |
| | 3 | **205294.49** | 189251.76 | **15.07%** | 14.76% | 21.91% | 22.30% |
| | 5 | 102493.32 | 116881.17 | 11.20% | 11.98% | 17.57% | 18.93% |
| | 10 | 77878.61 | 95571.05 | 9.93% | 10.97% | 16.93% | 17.95% |
| | 20 | 84077.43 | 95911.31 | 10.13% | 10.85% | 15.58% | 16.67% |
| | 30 | 91909.38 | 98398.47 | 10.37% | 10.83% | 14.68% | 15.81% |
| | 50 | 104769.52 | 101913.88 | 11.02% | 10.99% | 15.01% | 15.77% |
| | 100 | 89541.47 | 96040.86 | 10.42% | 10.74% | 15.39% | 15.64% |

based on the number of companies invested in each month.

Figure 6 illustrates the results for the equal weight investment strategy. When investing in market capitalization based companies, the highest returns were achieved when the number of firms invested in per month was three; however, as the number of firms increases, the investment returns decrease. In contrast, investing in Model based companies generally yields lower returns.

Figure 7 presents the results for the capital weighted investment strategy. For both market capitalization based and model based companies, there is a trend of diminishing final investment amounts once the number of firms invested in per month exceeds three to five.

The analysis of both investment strategies indicates that investing in market capitalization based companies exhibited a generally lower and more stable standard deviation of annual returns. Conversely, investing in model based companies showed relatively higher volatility, especially when the number of companies invested in per month was smaller.

Overall, investing in market capitalization based companies is suitable for conservative investors due to its stable final investment amounts and low volatility. In contrast, investing in Model based companies is more appropriate for aggressive investors, as it achieved the highest returns with approximately a sevenfold increase in the final investment amount when investing in five companies per month, despite the higher volatility.

As shown in Table 8, investing in five model based companies each month yielded the highest returns. We analyzed the companies comprising the portfolio each year under this strategy. Table 9 shows the companies that were included in the portfolio more than seven times during the test data period and the frequency of their appearance in the annual portfolios. We examined the status of these companies as of the start of the test data period in January 2015 and as of December 2023.

Firstly, RAD and WLL filed for bankruptcy in October 2023 and April 2020, respectively. ALT was listed in May 2017, with its stock price at approximately \$133.20 at the time of listing; however, by December 2023, its stock price had fallen by 91.55% to about \$11.25. CVEO and SPWR had stock prices of \$46.44 and \$16.84 in January 2015, respectively, but by December 2023, their prices had

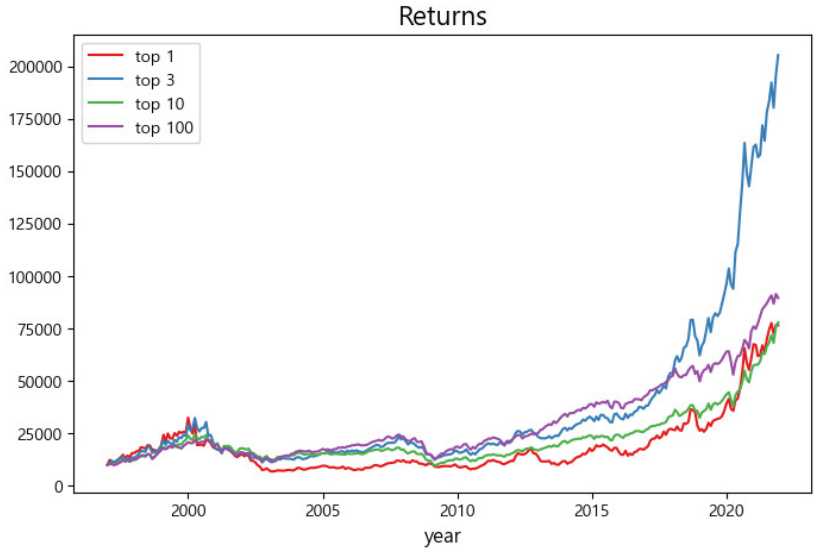


Figure 8: *Long term returns for the market capitalization based investment.*

declined by 50.80% and 71.32% to $22.85 and $4.83, respectively. Additionally, ENDP was delisted in August 2022, and CLVS filed for bankruptcy in December 2022. Such instances of bankruptcy or sharp stock price declines highlight the risks associated with Model based Investment.

### 4.3. Long term results for market capitalization weighted investment

When investing in market capitalization based companies, the previously calculated average annual returns were exceptionally high, reaching up to 30%. However, the period from 2015 to 2021 was characterized by significant growth in the U.S. stock market, which may naturally account for these high returns. Therefore, we also compared periods when the stock market did not exhibit such growth. We extended the investment period to 1997–2021 to analyze long-term investments. Since predicting returns for this extended period is not feasible due to the lack of available data, we only examined the results of investing in Market Capitalization based companies. The initial investment amount was consistently set at $10,000. For the period from 1997 to 2021, the final investment results as of December 2021 are summarized in Table 10. Investing in three companies per month resulted in the highest final amount.

Next, we analyzed the results of investing in the top 1, 3, 10, and 100 market capitalization stocks using the Equal Weight Investment strategy, which showed the highest returns. Figure 8 illustrates the simulation graphs of these four investment strategies over time. Until the early 2010s, the performance differences among the four investment strategies were not significant. However, from the late 2010s, the strategy of investing in the top 3 market capitalization stocks exhibited a sharp increase in returns. This suggests that the top 3 large-cap stocks played a leading role in the market.

Additionally, we analyzed investments in market capitalization based companies during the period from 1997 to 2014, when the stock market did not experience significant growth. The final investment outcomes for this period are presented in Table 11. During this period, investing in a larger number of companies, specifically more than 50, resulted in the highest final investment amount.

Table 11: Final amounts and annual average returns for market capitalization based strategy from 1997 to 2014

| Data | Top $n$ | Equal weight investment final amount | Capital weighted investment final amount | Equal weight investment average return | Capital weighted investment average return | Equal weight investment SD of returns | Capital weighted investment SD of returns |
|---|---|---|---|---|---|---|---|
| Market capitalization based | 1 | 17838.32 | 17838.32 | 7.02% | 7.02% | 28.08% | 28.08% |
| | 3 | 32993.71 | 32343.99 | 8.73% | 8.68% | 19.12% | 19.74% |
| | 5 | 23452.62 | 24771.45 | 6.00% | 6.46% | 14.97% | 15.97% |
| | 10 | 24108.34 | 24910.54 | 6.38% | 6.59% | 16.55% | 16.72% |
| | 20 | 32385.16 | 29748.90 | 8.28% | 7.77% | 17.25% | 17.34% |
| | 30 | 35775.47 | 31627.28 | 8.73% | 8.05% | 16.55% | 16.83% |
| | 50 | **41910.76** | 34764.28 | **9.77%** | 8.65% | 16.89% | 16.97% |
| | 100 | 38893.14 | 35205.79 | 9.45% | 8.77% | 17.38% | 17.08% |

## 5. Conclusion

In this study, our goal was to establish an investment strategy in the stock market using predictive models. First, we constructed a model to predict stock returns. Firm Specific data and Macroeconomic data were used, and these data were reduced into several components to generate 20 combinations of datasets. Machine learning algorithms were applied to implement the prediction model, and monthly returns were predicted using the model with best performance.

Subsequently, we developed a stock investment strategy. Companies for monthly investment were selected based on the return prediction model and market capitalization. Next, we determined the investment amount for each company.

A portfolio was constructed by investing in 10 companies each month from 2015 to 2021. When investing in Model based companies, the strategy with the highest returns achieved an average annual return of 23.85%, with a final investment amount of $32,405.99. For Market Capitalization based companies, the average annual return was 22.6%, with a final investment amount of $39,329.08. When investing in the S&P 500 Index, the average annual return was 13.89%, and the final investment amount was $23,890.75. Overall, investing in Model based companies outperformed investing in the S&P 500 Index but was slightly less effective than investing in Market Capitalization based companies.

The subsequent analysis focused on selecting the optimal number of companies to invest in each month to achieve higher final amounts. Investments in Market Capitalization based companies showed stable returns and low volatility, while investments in Model based companies showed higher returns but with increased volatility. Thus, conservative investors are advised to invest in Market Capitalization based companies, while aggressive investors may prefer investing in Model based companies. However, it is crucial to note the higher risk of adverse events such as bankruptcy or sharp stock price declines when investing in Model based companies.

Additionally, we analyzed the investment outcomes for Market Capitalization based companies by changing the overall investment period. During extended periods of stock market growth, investing in fewer than 10 companies each month led to higher returns. In contrast, during periods without significant market growth, investing in more than 50 companies per month was preferable and provided more stable returns.

These findings are expected to offer valuable insights to investors interested in the stock market. Future research could enhance the accuracy and reliability of stock return predictions by considering a wider array of variables and machine learning algorithms.

## Appendix A: Macroeconomic variables

| Fred | Description |
|---|---|
| RPI | Real Personal Income |
| W875RX1 | Real personal income excluding current transfer receipts |
| INDPRO | Industrial Production: Total Index |
| IPFPNSS | Industrial Production: Final Products and Nonindustrial Supplies |
| IPFINAL | Industrial Production: Final Products |
| IPCONGD | Industrial Production: Consumer Goods |
| IPDCONGD | Industrial Production: Durable Consumer Goods |
| IPNCONGD | Industrial Production: Non-Durable Consumer Goods |
| IPBUSEQ | Industrial Production: Equipment: Business Equipment |
| IPMAT | Industrial Production: Materials |
| IPDMAT | Industrial Production: Durable Goods Materials |
| IPNMAT | Industrial Production: Non-Durable Goods Materials |
| IPMANSICS | Industrial Production: Manufacturing |
| IPB51222S | Industrial Production: Non-Durable Consumer Energy Products: Residential Utilities |
| IPFUELS | Industrial Production: Non-Durable Consumer Energy Products: Fuels |
| CUMFNS | Capacity Utilization: Manufacturing |
| CLF16OV | Civilian Labor Force Level |
| CE16OV | Employment Level |
| UNRATE | Unemployment Rate |
| UEMPMEAN | Average Weeks Unemployed |
| UEMPLT5 | Number Unemployed for Less Than 5 Weeks |
| UEMP5TO14 | Number Unemployed for 5-14 Weeks |
| UEMP15OV | Number Unemployed for 15 Weeks and over |
| UEMP15T26 | Number Unemployed for 15-26 Weeks |
| UEMP27OV | Number Unemployed for 27 Weeks and over |
| ICSA | Initial Claims |
| PAYEMS | All Employees, Total Nonfarm |
| USGOOD | All Employees, Goods-Producing |
| CES1021000001 | All Employees, Mining, Quarrying, and Oil and Gas Extraction |
| USCONS | All Employees, Construction |
| MANEMP | All Employees, Manufacturing |
| DMANEMP | All Employees, Durable Goods |
| NDMANEMP | All Employees, Nondurable Goods |
| SRVPRD | All Employees, Service-Providing |
| USTPU | All Employees, Trade, Transportation, and Utilities |
| USWTRADE | All Employees, Wholesale Trade |
| USTRADE | All Employees, Retail Trade |
| USFIRE | All Employees, Financial Activities |
| USGOVT | All Employees, Government |
| CES0600000007 | Average Weekly Hours of Production and Employees, Goods-Producing |
| AWOTMAN | Average Weekly Overtime Hours of Production and Employees, Manufacturing |
| AWHMAN | Average Weekly Hours of Production and Employees, Manufacturing |
| CES0600000008 | Average Hourly Earnings of Production and Nonsupervisory Employees, Goods-Producing |
| CES2000000008 | Average Hourly Earnings of Production and Nonsupervisory Employees, Construction |
| CES3000000008 | Average Hourly Earnings of Production and Nonsupervisory Employees, Manufacturing |
| HOUST | Housing Units Started: Total |
| HOUSTNE | Housing Units Started: Northeast Census Region |
| HOUSTMW | Housing Units Started: Midwest Census Region |
| HOUSTS | Housing Units Started: South Census Region |
| HOUSTW | Housing Units Started: West Census Region |

| Fred | Description |
|---|---|
| PERMIT | Housing Units Authorized in Permit-Issuing Places: Total |
| PERMITNE | Housing Units Authorized in Permit-Issuing Places: Northeast Census Region |
| PERMITMW | Housing Units Authorized in Permit-Issuing Places: Midwest Census Region |
| PERMITS | Housing Units Authorized in Permit-Issuing Places: South Census Region |
| PERMITW | Housing Units Authorized in Permit-Issuing Places: West Census Region |
| DPCERA3M086SBEA | Real personal consumption expenditures |
| CMRMTSPL | Real Manufacturing and Trade Industries Sales |
| MRTSSM44X72USS | Retail Sales: Retail Trade and Food Services |
| DGORDER | Manufacturers' New Orders: Durable Goods |
| ACOGNO | Manufacturers' New Orders: Consumer Goods |
| ANDENO | Manufacturers' New Orders: Nondefense Capital Goods |
| AMDMUO | Manufacturers' Unfilled Orders: Durable Goods |
| BUSINV | Total Business Inventories |
| ISRATIO | Total Business: Inventories to Sales Ratio |
| M1SL | M1 Money Stock |
| M2SL | M2 Money Stock |
| M2REAL | Real M2 Money Stock |
| TOTRESNS | Total Reserves of Depository Institutions |
| NONBORRES | Reserves of Depository Institutions |
| BUSLOANS | Commercial and Industrial Loans, All Commercial Banks |
| REALLN | Real Estate Loans, All Commercial Banks |
| NONREVSL | Nonrevolving Consumer Credit Owned and Securitized |
| DTCOLNVHFNM | Consumer Motor Vehicle Loans Owned by Finance Companies, Level |
| DTCTHFNM | Total Consumer Loans and Leases Owned and Securitized by Finance Companies, Level |
| SBCACBW027SBOG | Securities in Bank Credit, All Commercial Banks |
| FEDFUNDS | Federal Funds Effective Rate |
| TB3MS | 3-Month Treasury Bill Secondary Market Rate, Discount Basis |
| TB6MS | 6-Month Treasury Bill Secondary Market Rate, Discount Basis |
| GS1 | Market Yield on U.S. Treasury Securities at 1-Year Constant Maturity |
| GS5 | Market Yield on U.S. Treasury Securities at 5-Year Constant Maturity |
| GS10 | Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity |
| AAA | Moody's Seasoned Aaa Corporate Bond Yield |
| BAA | Moody's Seasoned Baa Corporate Bond Yield |
| TB3SMFFM | 3-Month Treasury Bill Minus Federal Funds Rate |
| TB6SMFFM | 6-Month Treasury Bill Minus Federal Funds Rate |
| T1YFFM | 1-Year Treasury Constant Maturity Minus Federal Funds Rate |
| T5YFFM | 5-Year Treasury Constant Maturity Minus Federal Funds Rate |
| T10YFFM | 10-Year Treasury Constant Maturity Minus Federal Funds Rate |
| AAAFFM | Moody's Seasoned Aaa Corporate Bond Minus Federal Funds Rate |
| BAAFFM | Moody's Seasoned Baa Corporate Bond Minus Federal Funds Rate |
| EXSZUS | Swiss Francs to U.S. Dollar Spot Exchange Rate |
| EXJPUS | Japanese Yen to U.S. Dollar Spot Exchange Rate |
| EXUSUK | U.S. Dollars to U.K. Pound Sterling Spot Exchange Rate |
| EXCAUS | Canadian Dollars to U.S. Dollar Spot Exchange Rate |
| WPSFD49207 | Producer Price Index by Commodity: Final Demand: Finished Goods |
| WPSFD49502 | Producer Price Index by Commodity: Final Demand: Finished Consumer Goods |
| WTISPLC | Spot Crude Oil Price: West Texas Intermediate (WTI) |
| PPICMM | Producer Price Index by Commodity: Metals and Metal Products: Primary Nonferrous Metals |

| Fred | Description |
|---|---|
| CPIAUCSL | Consumer Price Index for All Urban Consumers: All Items |
| CPIAPPSL | Consumer Price Index for All Urban Consumers: Apparel |
| CPITRNSL | Consumer Price Index for All Urban Consumers: Transportation |
| CPIMEDSL | Consumer Price Index for All Urban Consumers: Medical Care |
| CUSR0000SAC | Consumer Price Index for All Urban Consumers: Commodities |
| CUSR0000SAD | Consumer Price Index for All Urban Consumers: Durables |
| CUSR0000SAS | Consumer Price Index for All Urban Consumers: Services |
| CPIULFSL | Consumer Price Index for All Urban Consumers: All Items Less Food |
| CUSR0000SA0L2 | Consumer Price Index for All Urban Consumers: All Items Less Shelter |
| CUSR0000SA0L5 | Consumer Price Index for All Urban Consumers: All Items Less Medical Care |
| PCEPI | Personal Consumption Expenditures: Chain-type Price Index |
| DDURRG3M086SBEA | Personal consumption expenditures: Durable goods |
| DNDGRG3M086SBEA | Personal consumption expenditures: Nondurable goods |
| DSERRG3M086SBEA | Personal consumption expenditures: Services |
| S&P 500 | S&P 500 Index |

## Appendix B: Firm specific variables

| Acronym | Description | Category |
|---|---|---|
| acc | Operating Accruals | Investment |
| adm | Advertising Expense-to-market | Intangibles |
| agr | Asset growth | Investment |
| alm | Quarterly Asset Liquidity | Intangibles |
| ato | Asset Turnover | Profitability |
| bm | Book-to-market equity | Value-versus-growth |
| bm_ia | Industry-adjusted book to market | Value-versus-growth |
| cash | Cash holdings | Value-versus-growth |
| cashdebt | Cash to debt | Value-versus-growth |
| cfp | Cashflow to price | Value-versus-growth |
| chcsho | Change in shares outstanding | Investment |
| chpm | Change in profit margin | Profitability |
| chtx | Change in tax expense | Momentum |
| cinvest | Corporate investment | Investment |
| depr | Depreciation / PP&E | Momentum |
| dolvol | Dollar trading volume | Frictions |
| dy | Dividend yield | Value-versus-growth |
| ep | Earnings-to-price | Value-versus-growth |
| gma | Gross profitability | Investment |
| grltnoa | Growth in long-term net operating assets | Investment |
| herf | Industry sales concentration | Intangibles |
| hire | Employee growth rate | Intangibles |
| ill | Illiquidity rolling (3 months) | Frictions |
| lev | Leverage | Value-versus-growth |
| lgr | Growth in long-term debt | Investment |
| maxret | Maximum daily returns (3 months) | Frictions |
| me | Market equity | Frictions |
| me_ia | Industry-adjusted size | Frictions |
| mom1m | Previous month return | Momentum |
| mom6m | Cumulative Returns in the past (2-6) months | Momentum |
| mom12m | Cumulative Returns in the past (2-12) months | Momentum |
| mom36m | Cumulative Returns in the past (13-35) months | Momentum |
| mom60m | Cumulative Returns in the past (13-60) months | Momentum |
| ni | Net Equity Issue | Investment |
| nincr | Number of earnings increases | Momentum |
| noa | Net Operating Assets | Investment |
| op | Operating profitability | Profitability |
| pctacc | Percent operating accruals | Investment |
| pm | Profit margin | Profitability |
| pscore | Performance Score | Profitability |
| rd_sale | R&D to sales | Intangibles |
| rdm | R&D to market | Intangibles |
| rna | Return on Net Operating Assets | Profitability |
| roa | Return on Assets | Profitability |
| roe | Return on Equity | Profitability |
| rsup | Revenue surprise | Momentum |
| sgr | Sales growth | Value-versus-growth |
| sp | Sales-to-price | Value-versus-growth |
| std_dolvol | Std of dollar trading volume (3 months) | Frictions |
| std_turn | Std of Share turnover (3 months) | Frictions |
| turn | Shares turnover | Frictions |

## Appendix C: Companies comprising the portfolio each year

Table C.1: Investment in market capitalization based companies

| Rank | Ticker | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | Total |
|---|---|---|---|---|---|---|---|---|---|
| 1 | AAPL | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 84 |
| 2 | MSFT | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 84 |
| 3 | JNJ | 12 | 12 | 12 | 12 | 12 | 12 | 11 | 83 |
| 4 | AMZN | 5 | 12 | 12 | 12 | 12 | 12 | 12 | 77 |
| 5 | FB | 3 | 12 | 12 | 12 | 12 | 12 | 12 | 75 |
| 6 | JPM | 12 | 8 | 12 | 12 | 12 | 7 | 12 | 75 |
| 7 | GOOG | 2 | 4 | 8 | 12 | 12 | 12 | 12 | 62 |
| 8 | XOM | 12 | 12 | 12 | 12 | 9 | - | - | 57 |
| 9 | WMT | 10 | - | 3 | 3 | 12 | 12 | 8 | 48 |
| 10 | WFC | 12 | 10 | 11 | 6 | - | - | - | 39 |
| 11 | GE | 12 | 12 | 5 | - | - | - | - | 29 |
| 12 | V | - | - | - | - | 6 | 11 | 7 | 24 |
| 13 | PG | 7 | 3 | - | - | 3 | 6 | - | 19 |
| 14 | BAC | - | - | 5 | 10 | 1 | 1 | - | 17 |
| 15 | TSLA | - | - | - | - | - | 4 | 12 | 16 |
| 16 | T | 1 | 10 | 4 | - | - | - | - | 16 |
| 17 | BRK | - | - | - | 5 | 5 | - | - | 10 |
| 18 | NVDA | - | - | - | - | - | 1 | 7 | 8 |
| 19 | PFE | 6 | - | - | - | - | - | - | 6 |
| 20 | MA | - | - | - | - | - | 6 | - | 6 |
| 21 | CVX | 2 | - | - | - | - | - | - | 2 |
| 22 | UNH | - | - | - | - | - | - | 2 | 2 |
| 23 | VZ | - | 1 | - | - | - | - | - | 1 |
| 24 | HD | - | - | - | - | - | - | 1 | 1 |

Table C.2: Investment in model based companies (companies included in the portfolio at least 5 times)

| Rank | Ticker | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | Total |
|---|---|---|---|---|---|---|---|---|---|
| 1 | PCG | - | - | - | 2 | 6 | 2 | 5 | 15 |
| 2 | RAD | - | - | 6 | 9 | - | - | - | 15 |
| 3 | BBBY | - | - | 2 | 10 | - | - | - | 12 |
| 4 | ENDP | - | 1 | 9 | - | - | - | - | 10 |
| 5 | SWN | - | 1 | 3 | 5 | - | - | - | 9 |
| 6 | WLL | 3 | 2 | - | - | 4 | - | - | 9 |
| 7 | BLUE | - | - | - | - | - | - | 8 | 8 |
| 8 | NAV | 4 | - | - | - | 1 | 2 | 1 | 8 |
| 9 | TMBR | - | - | - | - | - | - | 8 | 8 |
| 10 | CVEO | 8 | - | - | - | - | - | - | 8 |
| 11 | JOY | 5 | 3 | - | - | - | - | - | 8 |
| 12 | ALT | - | - | - | 8 | - | - | - | 8 |
| 13 | ODP | - | 3 | 4 | - | - | - | - | 7 |
| 14 | TLRY | - | - | - | - | 4 | 3 | - | 7 |
| 15 | VMW | 6 | 1 | - | - | - | - | - | 7 |
| 16 | DFBG | - | - | 7 | - | - | - | - | 7 |
| 17 | SPWR | - | 2 | 5 | - | - | - | - | 7 |
| 18 | UFS | 3 | 2 | 2 | - | - | - | - | 7 |
| 19 | CLVS | - | 5 | - | 2 | - | - | - | 7 |
| 20 | UA | - | - | 3 | 2 | 1 | 1 | - | 7 |
| 21 | NBR | 1 | - | 5 | - | - | - | - | 6 |
| 22 | GPRO | - | 6 | - | - | - | - | - | 6 |
| 23 | PBYI | 1 | 5 | - | - | - | - | - | 6 |
| 24 | CNX | 1 | 1 | 1 | - | 3 | - | - | 6 |
| 25 | NKTR | - | - | - | - | 1 | - | 5 | 6 |
| 26 | RRC | - | - | 2 | 2 | 2 | - | - | 6 |
| 27 | MDT | 5 | 1 | - | - | - | - | - | 6 |
| 28 | SGMS | - | - | - | 2 | 3 | - | - | 5 |
| 29 | HTZ | - | - | 5 | - | - | - | - | 5 |
| 30 | DO | - | - | 5 | - | - | - | - | 5 |
| 31 | LNCO | 5 | - | - | - | - | - | - | 5 |
| 32 | OCN | 5 | - | - | - | - | - | - | 5 |
| 33 | QEP | - | - | 5 | - | - | - | - | 5 |
| 34 | SAM | - | - | - | - | - | 2 | 3 | 5 |
| 35 | SM | 3 | 2 | - | - | - | - | - | 5 |
| 36 | COTY | - | - | - | 2 | 1 | 2 | - | 5 |
| 37 | RLGY | - | - | - | 1 | 4 | - | - | 5 |
| 38 | GEMP | - | - | - | - | 5 | - | - | 5 |
| 39 | CIE | - | 5 | - | - | - | - | - | 5 |

## Appendix D: Investment results based on the number of companies

Table D.1: Monthly investment results (top 1 companies)

| Data | Date | Equal weight investment returns | Capital weighted investment returns | Equal weight investment return rate | Capital weighted investment return rate |
|---|---|---|---|---|---|
| Market capitalization based | 2015-12 | 10897.17 | 10897.17 | 8.97% | 8.97% |
| | 2016-12 | 10402.35 | 10402.35 | −4.54% | −4.54% |
| | 2017-12 | 16431.80 | 16431.80 | 57.96% | 57.96% |
| | 2018-12 | 17328.35 | 17328.35 | 5.46% | 5.46% |
| | 2019-12 | 21619.75 | 21619.75 | 24.77% | 24.77% |
| | 2020-12 | 36585.71 | 36585.71 | 69.22% | 69.22% |
| | 2021-12 | 46204.58 | 46204.58 | 26.29% | 26.29% |
| | Average | - | - | 26.88% | 26.88% |
| | SD | - | - | 25.44% | 25.44% |
| Model based | 2015-12 | 5373.04 | 5373.04 | −46.27% | −46.27% |
| | 2016-12 | 69582.24 | 69582.24 | 1195.03% | 1195.03% |
| | 2017-12 | 119741.35 | 119741.35 | 72.09% | 72.09% |
| | 2018-12 | 15730.43 | 15730.43 | −86.86% | −86.86% |
| | 2019-12 | 6274.84 | 6274.84 | −60.11% | −60.11% |
| | 2020-12 | 11332.90 | 11332.90 | 80.61% | 80.61% |
| | 2021-12 | 8175.87 | 8175.87 | −27.86% | −27.86% |
| | Average | - | - | 160.95% | 160.95% |
| | SD | - | - | 426.37% | 426.37% |

Table D.2: Monthly investment results (top 3 companies)

| Data | Date | Equal weight investment returns | Capital weighted investment returns | Equal weight investment return rate | Capital weighted investment return rate |
|---|---|---|---|---|---|
| Market capitalization based | 2015-12 | 10729.35 | 10709.19 | 7.29% | 7.09% |
| | 2016-12 | 11428.19 | 11103.76 | 6.51% | 3.68% |
| | 2017-12 | 16749.25 | 16410.09 | 46.56% | 47.79% |
| | 2018-12 | 21594.43 | 20493.83 | 28.93% | 24.89% |
| | 2019-12 | 28447.64 | 27039.53 | 31.74% | 31.94% |
| | 2020-12 | 47400.96 | 44727.61 | 66.63% | 65.42% |
| | 2021-12 | 63910.36 | 60704.41 | 34.83% | 35.72% |
| | Average | - | - | 31.78% | 30.93% |
| | SD | - | - | 19.60% | 20.16% |
| Model based | 2015-12 | 6512.95 | 8032.03 | −34.87% | −19.68% |
| | 2016-12 | 16396.60 | 21754.41 | 151.75% | 170.85% |
| | 2017-12 | 26125.78 | 32569.47 | 59.34% | 49.71% |
| | 2018-12 | 12286.96 | 26058.45 | −52.97% | −19.99% |
| | 2019-12 | 8364.49 | 15611.70 | −31.92% | −40.09% |
| | 2020-12 | 10248.65 | 21166.04 | 22.53% | 35.58% |
| | 2021-12 | 15219.06 | 50287.84 | 48.50% | 137.59% |
| | Average | - | - | 23.19% | 44.85% |
| | SD | - | - | 66.21% | 75.75% |

Table D.3: Monthly investment results (top 5 companies)

| Data | Date | Equal weight investment returns | Capital weighted investment returns | Equal weight investment return rate | Capital weighted investment return rate |
|---|---|---|---|---|---|
| Market capitalization based | 2015-12 | 10494.19 | 10538.39 | 4.94% | 5.38% |
| | 2016-12 | 11338.52 | 11105.33 | 8.05% | 5.38% |
| | 2017-12 | 15634.09 | 15647.53 | 37.88% | 40.90% |
| | 2018-12 | 17898.29 | 18163.48 | 14.48% | 16.08% |
| | 2019-12 | 21547.53 | 22771.89 | 20.39% | 25.37% |
| | 2020-12 | 32800.61 | 35952.35 | 52.22% | 57.88% |
| | 2021-12 | 44657.99 | 48642.81 | 36.15% | 35.30% |
| | Average | - | - | 24.87% | 26.61% |
| | SD | - | - | 16.27% | 18.01% |
| Model based | 2015-12 | 8326.57 | 12072.51 | −16.73% | 20.73% |
| | 2016-12 | 19175.35 | 21574.85 | 130.29% | 78.71% |
| | 2017-12 | 23574.01 | 27012.75 | 22.94% | 25.20% |
| | 2018-12 | 16494.60 | 29874.90 | −30.03% | 10.60% |
| | 2019-12 | 18414.56 | 30588.61 | 11.64% | 2.39% |
| | 2020-12 | 24048.20 | 40078.14 | 30.59% | 31.02% |
| | 2021-12 | 30305.08 | 69790.91 | 26.02% | 74.14% |
| | Average | - | - | 24.96% | 34.68% |
| | SD | - | - | 47.91% | 27.82% |

Table D.4: Monthly investment results (top 10 companies)

| Data | Date | Equal weight investment returns | Capital weighted investment returns | Equal weight investment return rate | Capital weighted investment return rate |
|---|---|---|---|---|---|
| Market capitalization based | 2015-12 | 10319.99 | 10409.57 | 3.20% | 4.10% |
| | 2016-12 | 10857.53 | 10833.66 | 5.21% | 4.07% |
| | 2017-12 | 14024.62 | 14416.59 | 29.17% | 33.07% |
| | 2018-12 | 15114.26 | 15961.53 | 7.77% | 10.72% |
| | 2019-12 | 17786.10 | 19509.37 | 17.68% | 22.23% |
| | 2020-12 | 22993.54 | 27974.38 | 29.28% | 43.39% |
| | 2021-12 | 32914.64 | 39329.08 | 43.15% | 40.59% |
| | Average | - | - | 19.35% | 22.60% |
| | SD | - | - | 13.93% | 15.55% |
| Model based | 2015-12 | 9260.85 | 11145.88 | −7.39% | 11.46% |
| | 2016-12 | 19520.83 | 13714.30 | 110.79% | 23.04% |
| | 2017-12 | 24676.16 | 16597.54 | 26.41% | 21.02% |
| | 2018-12 | 20143.74 | 16007.01 | −18.37% | −3.56% |
| | 2019-12 | 20490.08 | 14868.16 | 1.72% | −7.11% |
| | 2020-12 | 22512.25 | 17214.71 | 9.87% | 15.78% |
| | 2021-12 | 32405.99 | 30540.53 | 43.95% | 77.41% |
| | Average | - | - | 23.85% | 19.72% |
| | SD | - | - | 40.38% | 25.87% |

Table D.5: Monthly investment results (top 20 companies)

| Data | Date | Equal weight investment returns | Capital weighted investment returns | Equal weight investment return rate | Capital weighted investment return rate |
|---|---|---|---|---|---|
| Market capitalization based | 2015-12 | 10432.31 | 10463.56 | 4.32% | 4.64% |
| | 2016-12 | 11490.47 | 11328.18 | 10.14% | 8.26% |
| | 2017-12 | 13824.89 | 14191.11 | 20.32% | 25.27% |
| | 2018-12 | 14986.30 | 15613.28 | 8.40% | 10.02% |
| | 2019-12 | 16937.09 | 18345.90 | 13.02% | 17.50% |
| | 2020-12 | 20158.13 | 24377.54 | 19.02% | 32.88% |
| | 2021-12 | 26253.41 | 32795.00 | 30.24% | 34.53% |
| | Average | - | - | 15.07% | 19.01% |
| | SD | - | - | 8.12% | 11.20% |
| Model based | 2015-12 | 8949.53 | 9061.23 | −10.50% | −9.39% |
| | 2016-12 | 13894.95 | 10692.25 | 55.26% | 18.00% |
| | 2017-12 | 16643.35 | 12194.99 | 19.78% | 14.05% |
| | 2018-12 | 16011.86 | 12941.04 | −3.79% | 6.12% |
| | 2019-12 | 15912.60 | 13611.00 | −0.62% | 5.18% |
| | 2020-12 | 18104.44 | 20812.65 | 13.77% | 52.91% |
| | 2021-12 | 25208.99 | 27150.40 | 39.24% | 30.45% |
| | Average | - | - | 16.16% | 16.76% |
| | SD | - | - | 22.27% | 18.65% |

Table D.6: Monthly investment results (top 30 companies)

| Data | Date | Equal weight investment returns | Capital weighted investment returns | Equal weight investment return rate | Capital weighted investment return rate |
|---|---|---|---|---|---|
| Market capitalization based | 2015-12 | 10762.75 | 10676.81 | 7.63% | 6.77% |
| | 2016-12 | 11778.92 | 11531.43 | 9.44% | 8.00% |
| | 2017-12 | 14364.15 | 14449.68 | 21.95% | 25.31% |
| | 2018-12 | 15544.33 | 15869.49 | 8.22% | 9.83% |
| | 2019-12 | 18208.01 | 18912.37 | 17.14% | 19.17% |
| | 2020-12 | 21008.51 | 24208.84 | 15.38% | 28.01% |
| | 2021-12 | 26035.60 | 31661.98 | 23.93% | 30.79% |
| | Average | - | - | 14.81% | 18.27% |
| | SD | - | - | 6.14% | 9.34% |
| Model based | 2015-12 | 9819.30 | 10653.07 | −1.81% | 6.53% |
| | 2016-12 | 13721.08 | 11447.72 | 39.74% | 7.46% |
| | 2017-12 | 16236.70 | 14529.70 | 18.33% | 26.92% |
| | 2018-12 | 15597.37 | 14383.09 | −3.94% | −1.01% |
| | 2019-12 | 15608.99 | 16602.83 | 0.07% | 15.43% |
| | 2020-12 | 17454.21 | 20923.09 | 11.82% | 26.02% |
| | 2021-12 | 23101.52 | 29851.02 | 32.36% | 42.67% |
| | Average | - | - | 13.80% | 17.72% |
| | SD | - | - | 15.99% | 13.95% |

Table D.7: Monthly investment results (top 50 companies)

| Data | Date | Equal weight investment returns | Capital weighted investment returns | Equal weight investment return rate | Capital weighted investment return rate |
|---|---|---|---|---|---|
| Market capitalization based | 2015-12 | 10508.22 | 10535.80 | 5.08% | 5.36% |
| | 2016-12 | 11150.30 | 11205.94 | 6.11% | 6.36% |
| | 2017-12 | 13528.81 | 13957.40 | 21.33% | 24.55% |
| | 2018-12 | 14621.04 | 15265.68 | 8.07% | 9.37% |
| | 2019-12 | 16758.26 | 17878.98 | 14.62% | 17.12% |
| | 2020-12 | 20119.80 | 22814.40 | 20.06% | 27.60% |
| | 2021-12 | 25130.34 | 29672.48 | 24.90% | 30.06% |
| | Average | - | - | 14.31% | 17.20% |
| | SD | - | - | 7.43% | 9.61% |
| Model based | 2015-12 | 10010.17 | 10809.84 | 0.10% | 8.10% |
| | 2016-12 | 12524.31 | 11616.08 | 25.12% | 7.46% |
| | 2017-12 | 14118.47 | 14143.25 | 12.73% | 21.76% |
| | 2018-12 | 13922.43 | 14251.36 | −1.39% | 0.76% |
| | 2019-12 | 13961.39 | 17783.27 | 0.28% | 24.78% |
| | 2020-12 | 17004.93 | 21530.80 | 21.80% | 21.07% |
| | 2021-12 | 23126.76 | 31240.41 | 36.00% | 45.10% |
| | Average | - | - | 13.52% | 18.43% |
| | SD | - | - | 13.56% | 13.67% |

Table D.8: Monthly investment results (top 100 companies)

| Data | Date | Equal weight investment returns | Capital weighted investment returns | Equal weight investment return rate | Capital weighted investment return rate |
|---|---|---|---|---|---|
| Market capitalization based | 2015-12 | 10289.10 | 10418.92 | 2.89% | 4.19% |
| | 2016-12 | 10836.18 | 11035.79 | 5.32% | 5.92% |
| | 2017-12 | 13343.34 | 13762.54 | 23.14% | 24.71% |
| | 2018-12 | 14053.04 | 14836.10 | 5.32% | 7.80% |
| | 2019-12 | 15996.69 | 17238.55 | 13.83% | 16.19% |
| | 2020-12 | 18910.57 | 21518.54 | 18.22% | 24.83% |
| | 2021-12 | 22996.76 | 27466.15 | 21.61% | 27.64% |
| | Average | - | - | 12.90% | 15.90% |
| | SD | - | - | 7.79% | 9.24% |
| Model based | 2015-12 | 10008.57 | 10566.58 | 0.09% | 5.67% |
| | 2016-12 | 11913.03 | 10456.98 | 19.03% | −1.04% |
| | 2017-12 | 14125.65 | 12268.53 | 18.57% | 17.32% |
| | 2018-12 | 13753.69 | 12662.22 | −2.63% | 3.21% |
| | 2019-12 | 14606.21 | 14758.90 | 6.20% | 16.56% |
| | 2020-12 | 19227.08 | 19935.36 | 31.64% | 35.07% |
| | 2021-12 | 25642.19 | 26167.31 | 33.36% | 31.26% |
| | Average | - | - | 15.18% | 15.44% |
| | SD | - | - | 13.38% | 12.86% |

## References


Ariyo AA, Adewumi AO, and Ayo CK (2014). Stock price prediction using the ARIMA model, In *Proceedings of 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, Cambridge, 106–112. IEEE.

Bini BS and Mathew T (2016). Clustering and regression techniques for stock prediction, *Procedia Technology*, **24**, 1248–1255.

Chen T and Guestrin C (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen L, Pelger M, and Zhu J (2024). Deep learning in asset pricing, *Management Science*, **70**, 714–750.

Cong LW, Feng G, He J, and He X (2022). Growing the efficient frontier on panel trees, *NBER Working Paper*, w30805, Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4316289

Dey S, Kumar Y, Saha S, and Basak S (2016). Forecasting to classification: Predicting the direction of stock market price using Xtreme gradient boosting, *PESIT South Campus*, 1–10.

Ding X, Zhang Y, Liu T, and Duan J (2015). Deep learning for event-driven stock prediction, *Twenty-fourth International Joint Conference on Artificial Intelligence*, 2327–2333.

Fung GPC, Yu JX, and Lam W (2003). Stock prediction: Integrating text mining approach using real-time news. In *Proceedings of 2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003. Proceedings*, Hong Kong, 395–402. IEEE.

Green J, Hand JR, and Zhang XF (2017). The characteristics that provide independent information about average US monthly stock returns, *The Review of Financial Studies*, **30**, 4389–4436.

Gu S, Kelly B, and Xiu D (2020). Empirical asset pricing via machine learning, *The Review of Financial Studies*, **33**, 2223–2273.

Jeffrey MJ (2023). U.S. Stock Ownership Highest Since 2008, Gallup, Retrieved May 24, 2023, Available from: https://news.gallup.com/poll/506303/stock-ownership-highest-2008.aspx

Jiang W (2021). Applications of deep learning in stock market prediction: Recent progress, *Expert Systems with Applications*, **184**, 115537.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, and Liu TY (2017). Lightgbm: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, **30**, Available from: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Kim KJ and Han I (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, *Expert Systems with Applications*, **19**, 125–132.

McCracken MW and Ng S (2016). FRED-MD: A monthly database for macroeconomic research, *Journal of Business & Economic Statistics*, **34**, 574–589.

Mittal A and Goel A (2012). Stock prediction using twitter sentiment analysis, *Standford University, CS229*, **15**, 2352, Available from: https://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, and Gulin, A (2018). CatBoost: Unbiased boosting with categorical features, *Advances in Neural Information Processing Systems*, **31**, Available from: https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

Shen S, Jiang H, and Zhang T (2012). Stock market forecasting using machine learning algorithms, *Department of Electrical Engineering, Stanford University, Stanford, CA*, 1–5.

Zhong X and Enke D (2017). Forecasting daily stock market return using dimensionality reduction, *Expert Systems with Applications*, **67**, 126–139.

*Received July 18, 2024; Revised August 24, 2024; Accepted August 24, 2024*