

US Stocks Market Movements Prediction: Classification of SP-500 Using Machine Learning Technology

Fang Chen^{1, *, †}, Jinglun Gao^{2, *, †}, Zhiwen Zhang^{3, *, †}

¹Department of System Engineering and Engineering Management, Chinese University of Hong Kong, Hong Kong, China

²The College of Liberal Arts & Sciences, University of Illinois at Urbana Champaign, Champaign, Illinois, USA

³Rady School of Management, University of California, San Diego, USA

*Corresponding author: 1155124517@link.cuhk.edu.hk, jinglun3@illinois.edu, zh074@ucsd.edu

[†]These authors contributed equally.

Abstract. In the field of quantified investment, risk quantification and maximum expected return are the problems focused on by the investors. Besides, a powerful toolkit for predicting the stock price movement is also very important for investors. In this paper, five stocks that are components of the SP-500 Index are selected, and the Mean-Variance method is used to optimize the portfolio of the above stocks. Moreover, five machine learning methods are compared to evaluate the performance in the application of stock price movement prediction. The results show that the combination of “AMZN”, “MSFT” and “AAPL” can achieve a good expected return within a low risk. In addition, the Artificial Neural Network method has the highest accuracy in predicting the multiclass stock price movement. Our research has a reference significance for the investors in the application of risk quantification and stock price prediction.

Keywords: component; Stocks Market Movements Prediction; Multiclass Classification; Feature Engineering; Machine Learning

1. Introduction

Trend prediction for the stock is very useful in the field of intelligent measured investment [1]. However, the risk existing in the market would influence the judgement of people [2]. The classical framework of modern portfolio theory assumes that the investor only cares about the first two moments of the return distribution: mean and variance. The variance serves as a measure of risk, and the risk-adjusted portfolio performance is measured by the Sharpe ratio, which the investor wants to maximize. However, a rational investor would not consider all variability as risk, but the only variability below a certain benchmark that depends on his preferences [3]. After minimizing the risk, the return may be maximized, if we can find a suitable model to predict the trend of investment combination.

Many researchers proposed some methods to measure the risk, for example, Sharpe [4, 5] defined the Sharpe ratio (SR) as a ratio of the expected portfolio return to the standard deviation. Please note that the optimal portfolio in the sense of maximizing the SR belongs to the efficient frontier in the case without a risk-free asset, which can be obtained as a solution to Markowitz's optimization problem. Krokmal et al [6] proposed the downside risk measures which depend only on the positive values of the loss function or negative values of the return. The risk theory has been pushed that the quantile-based measures are well-suited functions to quantify risk [7-9].

Many passages did similar things [10-12]. They used some indicators such as moving average, money flow index, related strength index, Stochastic Oscillator, Moving Average Convergence/Divergence [13]. Besides, there were many models: Artificial Neural Network (ANN) [14], Support Vector Machine (SVM) [14], Random Forest [14], Naive-Bayes [14], Decision Stump [15], Linear Regression [15], Boosting [15], Stock Prediction based on Textual Analysis of Financial News Articles [15], Applying Expert Weighting to Stock Prediction [15].

The nature of the stock market is dynamic and unpredictable, it is a very challenging task to predict the actual stock prices. And the previous research results about predicting the binary future stock movements by machine learning techniques did not have a very persuasive accuracy score. Hence, this paper considers the situation when the stock prices are relatively stable and transforms the binary classification into the ternary classification.

In this paper, two steps are proposed to obtain the best combination of investment. The first step is Mean-Variance-based risk optimization; the second step is the machine learning-based trend prediction. Based on our work, it is possible to transfer positions convertible based on our prediction, which is useful to reduce our investment risk. Finally, a comparison of machine learning-based classification for predicting the stock market movements is given to analyze the performance of different methods in applying stock price prediction. In this paper, several machine learning models are implemented, including Logistics Regression (LR), Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machine (SVM), and Artificial Neural Network (ANN). And the model performances are tested using the accuracy score.

The rest of this paper includes that Section 2 introduces the methods; Section 3 introduces the results and discussion; Section 4 introduces the conclusion.

2. Methods

2.1 Data Preparation

In this paper, the data sets of the target stocks and indexes are derived from the Yahoo Finance website [16]. To be more specific, six features (high price, low price, open price, close price, adjusted close price and volume) are extracted directly. In the trend prediction step, these features are translated into fifteen indicators as the inputs of our machine learning models, including “Simple Moving Average”, “Weighted Moving Average”, “Exponential Moving Average”, “Momentum”, “Williams %R”, “Accumulation/Distribution Oscillator”, “Moving Average Convergence Divergence”, “Stochastic Oscillator”, “Relative Strength Index”, “Commodity Channel Index”, “Accumulation/Distribution Index”, “Negative Directional Indicator”, “Positive Directional Indicator”, “Aroon Down Indicator”, “Aroon Up Indicator”.

The formulas of these indicators are given in Table 1.

Table 1. Formulas Of Chosen Indicators

Name of indicators	Formulas
Simple n (5 here)-day Moving Average (SMA)	$\frac{C_t + C_{t-1} + \dots + C_{t-4}}{n}$
Weighted n (5 here)-day Moving Average (WMA)	$\frac{(5)C_t + (4)C_{t-1} + \dots + C_{t-4}}{n + (n-1) + \dots + 1}$
Exponential n (5 here)-day Moving Average (EMA)	$C_t \cdot (1 + \alpha) + EMA_{t-1} \cdot (1 - \alpha)$
Momentum	$\frac{C_t - C_{t-5}}{H_n - L_n} \times 100$
Williams %R	$\frac{H_t - C_{t-1}}{H_t - L_t}$
Accumulation/Distribution Oscillator (ADO)	$MACD(n)_{t-1} + \frac{2}{n+1} \times (EMA(5)_t - EMA(10)_t - MACD(n)_{t-1})$
Moving Average Convergence Divergence (MACD)	$\%K = \frac{C_t - LL_{t-(n-1)}}{HH_{t-(n-1)} - LL_{t-(n-1)}} \times 100$
Stochastic Oscillator (SO)	$\%D = \frac{\sum_{i=0}^{n-1} K_{t-1}}{10} \%$

Relative Strength Index (RSI)	$100 - \frac{100}{1 + \frac{\sum_{i=0}^{n-1} UP_{t-i}/n}{\sum_{i=0}^{n-1} DW_{t-i}/n}}$
Commodity Channel Index (CCI)	$\frac{M_t - SM_t}{0.015D_t}$
Accumulation/Distribution Index (ADI)	$\frac{(C_t - L_t) - (H_t - C_t)}{H_t - L_t} \times Vol$
Negative Directional Indicator (NDI)	$\frac{SMDM}{ATR}$
Positive Directional Indicator (PDI)	$\frac{SPDM}{ATR} \times 100$
Aroon Down Indicator (ADC)	$\frac{5 - \text{Periods Since 5 period Low}}{5} \times 100$
Aroon Up Indicator (AUC)	$\frac{5 - \text{Periods Since 5 period High}}{5} \times 100$

In the above table, C_t means the closing price at time t , H_t represents the high price and L_t is the low price. α means $\frac{\text{Smoothing}}{1+n}$. HH_t and LL_t are highest price and the lowest price in the last t days, respectively. $M_t = \frac{H_t + L_t + C_t}{3}$, $SM_t = \frac{(\sum_{i=1}^n M_{t-i+1})}{n}$ and $D_t = \frac{(\sum_{i=1}^n |M_{t-i+1} - SM_t|)}{n}$. UP_t and DW_t mean upward and downward price change, respectively. $SMDM = \text{Smoothed}(L_t - L_{t-n+1})$ and $SPDM = \text{Smoothed}(H_t - H_{t-n+1})$. ATR means the average true range.

Different from the ten prominent indicators selected by [14], indicators calculated using volume are considered in this paper. Moreover, the previous movement of stocks is another important feature. All the indicators are treated as continuous variables. The movement of the stock is turned into categorical variables. To be more specific, if the next-day close price increases more than 0.5%, the representation is “+1”; if the next-day close price decreases more than 0.5%, the representation is “-1”; and in the other cases, the representation is “0”. All the rows with null values are deleted. Then the data is split into training and testing set using the “train_test_split” function in the package of toolkit sklearn, where 20% of the data is used as a testing set.

2.2 Mean Risk Optimization

In this paper, the best return should be within the limitation of risk criteria. Therefore, the risk quantification is of great importance. For example, the benchmark below which volatility is considered to be downside volatility depends on the investor's preferences, and it does not necessarily coincide with the mean portfolio return. If returns are perfectly symmetrically distributed, the investor requires to target the volatility below the benchmark directly. Finally, the measurement by the downside deviation can be calculated by the formula (1).

$$\sigma_B = \sqrt{\frac{1}{T} \sum_{t=1}^T [\text{Min}(r_t - B, 0)]^2} \quad (1)$$

T means the whole-time series, and r represents the returns; the investor proposes a benchmark B set.

2.3 Machine Learning for Predicting the Movement

In this paper, five classical machine learning-based methods (LR, RF, GBM, SVM, ANN) are used to predict the movement of the stock price.

2.3.1 Logistical Regression

In this paper, the logistic regression is used to fit for those dependent features calculated by Subsection “Data Preparation”. And the stock price movement prediction could be transformed into

a three-classes classification machine learning problem. The result can be categorized into “up (1)”, “neutral (0)” and “down (-1)”. The historical data is used to train the model, which could predict the point in the next timestamp. The probability of the point in the next timestamp can be obtained by formula (2).

$$p = \frac{\exp(c_0 + c_1 x_1 + \dots + c_k x_k)}{1 + \exp(c_0 + c_1 x_1 + \dots + c_k x_k)} \quad (2)$$

where c_i represents the stable variable; x_i represents the i^{th} feature. Natural logarithm operations are performed on the formula (1) obtained through the likelihood function of joint density function with n samples. Therefore the maximum likelihood probability can be calculated by Formula (3).

$$\ln LF = \sum_{i=1}^n y_i (c_0 + \sum_{k=1}^m x_{ki}) - \ln(1 + \exp(c_0 + \sum_{k=1}^m x_{ki})) \quad (3)$$

where the x_{ki} represents the k^{th} feature index variable of the i^{th} time stamp, $k = 0, 1, 2, \dots, m$. Hyperparameter tuning is performed for Logistics Regression. Two penalizations are used, L1 and L2, and different regularization strength values are tried.

2.3.2 Random Forest

In this paper, random forest is used to aggregate a large number of decision trees to solve the inherent defects of a single model or a set of parameters of the model, learn from each other, and avoid limitations.

There are four steps for the random forest:

Step 1 is to do random sampling and train the decision tree. For a sample with a sample size of N , N samples are put back and extracted for N times, and 1 sample is extracted for each time, finally forming N samples. The selected N samples are used to train a decision tree as the samples at the root node of the decision tree.

Step 2 is to select random attributes to make node splitting attributes. When each sample has M attributes, and each node of the decision tree needs to be split, m attributes are randomly selected from these M attributes to meet the condition $m \ll M$. Then some strategy (such as information gain) is used to select one attribute from the m attributes as the split attribute of the node.

Step 3 is to repeat step 2 until the tree can no longer be split.

Step 4 is to repeat steps 1 to 3 to create a forest of decision trees.

Hyperparameter tuning is also performed for Random Forest Classifier. The different number of trees in the forest, the maximum depth of the tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, and the number of features to consider are tried.

2.3.3 Gradient Boosting Machine

In this paper, a gradient boosting machine is used to combine a few simple models iteratively, called “weak learners,” to get “strong learners” with a higher accuracy score.

The best result in multiple classifications can be expressed as formula (4).

$$\gamma_{jkm} = \frac{K-1}{K} \cdot \frac{\sum_{x_i \in R_{jkm}} r_{ik}}{\sum_{x_i \in R_{jkm}} |r_{ik}| \cdot |2 - r_{ik}|} \quad (4)$$

Hyperparameter tuning is also performed for Gradient Boosting Classifier. The different number of boosting stages to perform, the maximum depth of the tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, and the number of features to consider are tried.

2.3.4 Support Vector Machine

In this paper, Support Vector Machine (SVM) is used to maximize the margin in the feature space calculated by Subsection “Data Preparation”. Assume that input vectors $x_i \in \mathbb{R}^n, i = 1, 2 \dots N$ and class labels $y_i \in \{+1, -1\}, i = 1, 2 \dots N$, and there is a resulting decision boundary defined by formula (5).

$$f(x) = \text{sign}(\sum_{i=1}^N y_i \alpha_i \cdot K(x, x_i) + b) \quad (5)$$

Then, the problem can be turned into a quadratic programming problem, which is shown by formula (6) to (8).

$$\text{Maximize } \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \cdot y_i y_j \cdot K(x_i, x_j) \quad (6)$$

$$\text{Subject to } 0 \leq \alpha_i \leq c \quad (7)$$

$$\sum_{i=1}^N \alpha_i y_i = 0, i = 1, 2 \dots N \quad (8)$$

Use Gaussian kernel function as kernel function and classification decision function can be shown as formula (9).

$$f(x) = \text{sign}(\sum_{i=1}^N y_i \alpha_i \cdot \exp(-\frac{\|x - x_i\|^2}{2\sigma^2}) + b) \quad (9)$$

Hyperparameter tuning is also performed for the Support Vector Classifier. Two kernel types, linear and RBF are used, and different strengths of the regularization are tried.

2.3.5 Artificial Neural Network

The artificial neural network is a typical machine learning algorithm designed to imitate human brain processes based on a collection of interconnected nodes or units called neurons. This paper uses a three-layer feed-forward neural network, which contains one input layer, one hidden layer, and one output layer. The input layer has 16 units or neurons which corresponds to 16 technical indicators. The output layer has 3 units or neurons which corresponds to 3 classes and it employed the ‘softmax’ activation function. Hyperparameter tuning for ANN is also used in this paper. The different number of hidden layer neurons or units, different values of epochs, different number of batch size, and three different activation function is tested. The set of parameters that gives the best test accuracy score is using 10 neurons in the hidden layer, 100 as epochs value, 20 as batch size, and is using the ‘softplus’ as activation function.

3. Results and Discussion

In this paper, the risk optimization step is processed to minimize the risk, which could give a satisfying investment strategy for getting the best return. The following shows the results of “risk optimization” and “prediction of stock movement”.

3.1 Results of Risk Optimization

Five stocks (AAPL, AMZN, FB, GOOG, MSFT) which are components of the S&P 500 Index, are chosen to form an optimal portfolio. The stock price during ten years (from 2010-12-31 to 2021-1-1) is extracted from the Yahoo finance website [16] and the corresponding daily returns of these five stocks are calculated. The expected returns and covariance matrix of five stocks are estimated based on the historical data. In the experiment, the risk measure used to optimize the portfolio is the standard deviation, and the objective function is set to maximize the Sharpe ratio of the generated

portfolio. The experiment is based on the toolkit Riskfolio-Lib. The optimal weights are shown in Table 1.

Table 3. Weights of Five Selected Stocks

	AAPL	AMZN	FB	GOOG	MSFT
weights	0.2178	0.4098	0.0668	4.4127e-10	0.3056

In table 1, the results show that the weights of “AMZN”, “MSFT” and “AAPL” are the biggest, however, the weights of “GOOG” and “FB” are the smallest. That means among five stocks Facebook and Google have relatively small weights compared to others. Besides, please note that the summary of “AMZN”, “MSFT” and “AAPL” arrives above 0.9.

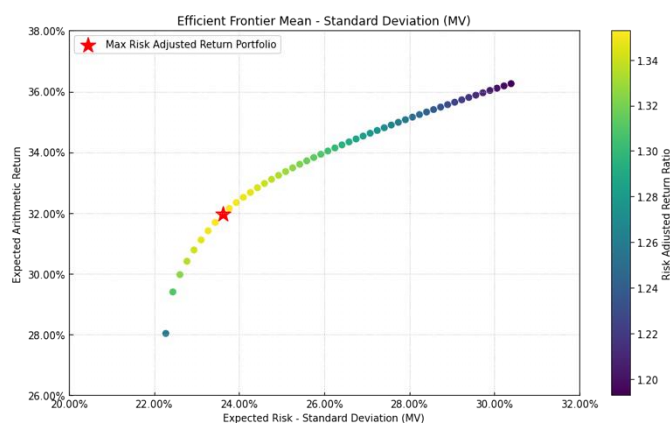


Figure 1. Risk-Adjusted Return Ratio.

In figure 1, the x-axis represents the quantified risk; the y-axis represents the expected return. The red star means the maximum risk-adjusted return shown in the legend. Figure 1 shows that the best-expected return may stay at the 27.8% standard deviation ratio of the risk.

Obviously, the results show that our strategy, which selects the “AMZN”, “MSFT” and “AAPL” can achieve a good expected return within a low risk. This may be because the three companies are the big stable corporations, which can resist the bad influences of the risk from the stock market.

3.2 Results of Movement Prediction

From the results of part A, four stocks that have relatively large weights in the optimal portfolio are selected to be the targets for movement prediction. Besides, the SP-500 Index is also considered in the experiment. The accuracy score is used to evaluate the performance of five machine learning-based methods. The results are shown in Table 3.

Table 3. Results of movement prediction

	LR	RF	GBM	SVM	ANN
AAPL	0.435	0.485	0.490	0.465	0.510
AMZN	0.395	0.460	0.460	0.385	0.415
FB	0.370	0.400	0.365	0.385	0.415
MSFT	0.495	0.520	0.485	0.490	0.530
SP-500	0.635	0.630	0.595	0.620	0.640

Table 3 shows the performance analysis of stock movement prediction based on machine learning. The accuracy is used to evaluate the five methods. The accuracy scores for four stocks are quite low from the results, with most of the scores below 0.500, which is useless in the prediction. This illustrates that the stocks selected has a random characteristic. However, the results for SP-500 are

much higher than those for stocks, with the highest accuracy score of 0.640 reached by the ANN method. To sum up, ANN is the most suitable model in most cases for movement prediction.

According to the combination of risk optimization and movement prediction results, it is reasonable to obtain that the US stocks market has great volatility, which means that there is a high risk when we want to quantify the expected return from the US stocks. However, after our optimization, the combination that “AMZN”, “MSFT” and “AAPL” can help to achieve a good expected return within a low risk. Moreover, the machine learning methods can aid in predicting the stock movement to some extent.

4. Conclusion

This paper processed two steps to quantify the expected return including risk control and stock price movement prediction. Firstly, portfolio optimization is proposed to quantify the risk and keep the risk within a low level. Secondly, five machine learning methods are compared based on US stocks movement prediction.

Our proposed strategy can effectively help the investor obtain their maximum expected return while keeping the risk at the limit of their accepted range. And the prediction results of the ANN method can be a meaningful reference for investors to change their positions. Moreover, the indexes are much suitable for movement prediction than the stocks. In conclusion, our research has a reference significance in the application of quantified investment.

References

- [1] Poh K L. An intelligent decision support system for investment analysis. *Knowledge and Information Systems*, 2000, 2(3): 340-358.
- [2] Campbell J Y. Understanding risk and return. *Journal of Political economy*, 1996, 104(2): 298-345.
- [3] Rigamonti A. Mean-variance optimization is a good choice, but for other reasons than you might think. *Risks*, 2020, 8(1): 29.
- [4] Sharpe W F. Mutual fund performance. *The Journal of business*, 1966, 39(1): 119-138.
- [5] Sharpe W F. The sharpe ratio. *Journal of portfolio management*, 1994, 21(1): 49-58.
- [6] Krokmal P, Zabarankin M, Uryasev S. Modeling and optimization of risk. *Handbook of the fundamentals of financial decision making: Part II*, 2013: 555-600.
- [7] Acerbi C, Tasche D. Expected shortfall: a natural coherent alternative to value at risk. *Economic notes*, 2002, 31(2): 379-388.
- [8] Artzner P, Delbaen F, Eber J M, et al. Coherent measures of risk. *Mathematical finance*, 1999, 9(3): 203-228.
- [9] Acerbi C, Tasche D. Expected shortfall: a natural coherent alternative to value at risk. *Economic notes*, 2002, 31(2): 379-388.
- [10] Xu Y, Cohen S B. Stock movement prediction from tweets and historical prices. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018: 1970-1979.
- [11] Lu H, Han J, Feng L. Stock movement prediction and n-dimensional inter-transaction association rules. *1998 ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, Seattle, WA, USA, ACM, New York, USA. 1998.
- [12] Qiu M, Song Y. Predicting the direction of stock market index movement using an optimized artificial neural network model. *PloS one*, 2016, 11(5): e0155133.
- [13] Coupelon O. Neural network modeling for stock movement prediction A state of the art. *Blaise Pascal University*, 2007: 83.
- [14] Patel J, Shah S, Thakkar P, et al. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, 2015, 42(1): 259-268.

- [15] Shah V H. Machine learning techniques for stock prediction. Foundations of Machine Learning| Spring, 2007, 1(1): 6-12.
- [16] Website: <https://finance.yahoo.com/>