Prediction Of Stock Performance in The U.S. Stock Market Based on Logistic Regression Model

Zhaoyang Hong[1, *]

[1]Department of Quantitative Finance, Hong Kong University of Science and Technology, Kowloon, Hong Kong

*Corresponding author: zhongaf@connect.ust.hk

Abstract. How to select stocks as investment targets is a hot issue of concern in the investment market. Some researchers have found that the quintile ranking of payout, valuation, profitability, and historical growth in the market can be used as a screening basis to help investors make initial selections. However, the differences in ranking within the same quintile are still obvious, which does not allow for a fine and clear selection. Therefore, in this paper, by collecting the data profile of more than five thousand US stocks from 2015 to 2021 and fitting a logistic model to the data, the author examines price to book ratio, sales growth rate, price to earnings ratio, gross profit to asset ratio, return on equity, dividend yield and earnings per share to predict whether the stocks will outperform in the future, and further derive the result of whether it is worth investing or not under different decision thresholds to help differently risk-averse investors to screen stocks. The results show that for the prediction of whether stocks would "outperform or not", the final model's prediction was 71.28 percent accurate. This excellent result shows that this logistic model can be applied to the real market to help general investors and investment institutions to select the right stocks.

Keywords: prediction of stock performance, U.S. stock market, financial and accounting ratios, logistic regression model.

1. Introduction

Nowadays, with the development of big data and artificial intelligence, integrating data of listed companies and predicting stock performance is becoming a big hot trend of research in the financial market. Using data mining techniques to predict the stock market is one of the most crucial concerns that has to be looked into. [1].

There are many ways to predict stock performance, one of which is based on financial ratio prediction. For example, the famous Fama-French 3-factor model has pointed out that financial and accounting ratios are crucial instruments for forecasting stock performance, and recently financial ratios have been widely used to predict the future trend of stock prices [2]. Based on this status, more and more financial ratios have been invented and created to represent the current situations and future trends of companies. The financial ratios contain almost all information about listed companies that are engaged in market [3]. All financial ratios can be broadly classified into four categories, which represent four different aspects of a company, namely profitability, payout, valuation, and historical growth.

The popular method of screening based on financial ratios is to consolidate stock data from the market and use quintiles to calculate where a company is in the market for each of the four dimensions. Focusing on growth stocks, if a company's valuation metrics are in the higher quintile, for example, the 4th quintile or 5th quintile, profitability and historical growth are in the top quintile from highest to lowest, and the dividend yield is positive, then the company has a high probability of having a good performance in the future, the stock price is likely to perform well in the future as well. On the other hand, for value stocks, if a company's valuation metrics are in the lower quintile, for example, the 1st and 2nd quintile, profitability and historical growth are in the top quintile sorted from high to low, and the dividend yield is high enough, then the company also has a large chance to outperform in terms of rate of return in the future. This method is a quick and easy way to screen out some of the promising stocks in the future, but it is very rough if investors need to select and judge a particular stock precisely. For example, when still focusing on growth stocks, using a quintile to screen

valuation and payout would be still very simple, but since different companies have different property allocation strategies, even though the profitability and historical growth data are in the same highest quintile, within quintile, the differences in the respective ranking of profitability and historical growth data can be very large, sometimes even affecting the future price direction of the stock and thus interfering with investors' final stock selection, the same is true for value stocks. Addressing this issue is of great interest for stock selection.

The most straightforward way to move from rough financial ratios' quintile prediction to accurate statistical prediction is to derive a forecasting model by fitting the data through machine learning. Most studies in the past have shown that the results obtained using traditional linear statistical forecasting methods, such as the ordinary least square method (OLS), are often discarded due to their low precision, as historical data often have outliers as well non-linear characteristics. Nonlinear statistical methods, on the other hand, can provide advanced inspiration for predicting stock prices more accurately [4]. Within the nonlinear statistical methods, logistic regression model is very suitable for dividing a collection of independent factors into two or even more mutually exclusive categories [5]. So, this study fits a nonlinear binary logistic model based on financial and accounting ratios of more than five thousand US-listed companies from 2015 to 2021 in an attempt to predict whether the future rate of return of stocks will outperform the market and further help individual investors and investment institutions to select good stocks.

This study uses gross profit to asset ratio and return on equity to represent profitability, price-to-book ratio and the price-to-earnings ratio to represent valuation, dividend yield to represent payout, historical earnings per share and revenue growth rate to represent historical growth. The data of all companies are divided into train set and test set, a logistic regression model is built to forecast whether the target stock will outperform in the future, namely, whether the stock's rate of return is exceeding the whole market under different thresholds, i.e., above S&P 500 index. The final outcomes show that for determining whether to expect stocks would "outperform or not", the final model's prediction was 71.28 percent accurate at default threshold and can adjust different thresholds which leads to different accuracy based on diverse risk tolerance of investors. This model can be innovatively applied to investors with different risk preferences, and it shows higher accuracy in investors who are a little bit more risk averse.

Next, the paper will present the study results in the order of literature review, methodology and model framework, results, and conclusions.

2. Literature Review

The strong market growth and various macroeconomic events in recent years have increased investment opportunities and led various distinguished financiers around the world to invest more and more effort in forecasting stock performance in different markets.

In this section of the article, the authors divide some related studies into two parts and use a chronological review of some of the key studies related to this area, not only to see how the models for predicting stock performance have evolved but also to intuitively illustrate the need for and significance of research in this area.

2.1 The Feasibility of Financial and Accounting Ratios in Stock Prediction

In 1992, Fama and French pointed out the existence of a correlation across the board between stock returns and underlying factors for many stocks in the United States based on existing studies, which, were based on the investigation and research of several financial ratios, such as the book to market value, earnings yield, cash flow yield, and business size, while they have become crucial indicators in determining stock performance and even predicting future stock performance [8]. They have become crucial indicators in determining stock performance and even predicting future stock performance.

In 2013, Novy-Marx concluded that in the full universe of stocks, Gross Profit to Assets ratio is the best predictor of returns by examining profitability and the cross-section of expected returns [12]. Novy-Marx's research lays the foundation for this study to use gross profit to assets ratio as representatives of firm profitability.

2.2 Model Building in Stocks Prediction

In 1968, Altman, who is considered a pioneer in stock performance prediction, proposed that different financial ratios could be combined into one group and that multiple discriminate analysis models could be used to predict whether a company would go bankrupt based on these disaggregated data [6]. The accuracy of the model was 94%. The high precision of this model proved the effectiveness of the multivariate data analysis framework in predicting the future growth of companies and further encouraged scientists who wanted to continue their research in this field, which opened the way for stock performance prediction.

In 1974, Abdel-Khalik and Lusk found that linear regression could not be used as a predictor when they studied the regression structures to predict the rate of return of test samples [7]. He believed that there were three reasons behind this phenomenon: the particular choice of financial ratios, the strong multi-co-linearity among different ratios, and using the unadjusted R square to gauge an explanatory power.

In 2007, Lee pointed out that the logistic regression model can be useful in determining what will be present or absent of outcomes or characteristics when the prediction results are based on a set of predictor variable's value [9]. Although the logistic regression model shows similarity to the linear regression model in terms of results, it is more practical to use the logistic regression model, and it is possible to create a multivariate regression in a logistic regression model between the dependent and independent variables.

In 2012, Dutta applied the logistic model in Indian stock market to fit 30 large capitalization corporations' eight financial ratios during a four-year period to predict the rate of return of stocks as good or bad, and the final accuracy of Dutta's model reached 74.6% [10, 11].

In 2018, Alia et al. applied a logistic model to fit 109 listed non-financial companies in the stock market of Pakistan from 2011 to 2015 with 6 financial ratios to predict whether the shareholder's profit of the stock is above the KSE 100 index, and the final model accuracy is 89.77% [4].

Alia et al.'s research, together with Dutta's study, demonstrates that it is perfectly feasible and promising to use logistic models to predict the future performance of stocks with very high accuracy, so this study will fit models based on these experiences and financial ratios to help investors and investment institutions in stock screening.

However, a compilation of past research reveals that the logistic model is not among the most commonly used models for predicting which stocks will perform well in the future, but rather, the logistic model is more commonly used for predicting corporate failures, bubble economies, financial disasters, and certain individual credit defaults. Until now, this model has not been used to predict whether the U.S. stocks will outperform or not in terms of rate of return. Therefore, this paper will proceed to this entry point to predict the rate of return performance of U.S. stocks through financial and accounting ratios using logistic regression model, and innovatively introduce different thresholds to simulate the performance of investors with different risk appetite and to observe the corresponding accuracy of this model.

3. Methodology and Model Framework

3.1Methodology

Binary logistic regression model is a commonly used prediction model, it is very famous in the study of predicting binary dependent variables (0 or 1), for example, in this paper the author wants to predict whether the stock's rate of return will outperform the S&P 500 index, where 1 means that the

stock's 5-year rate of return exceeds the S&P 500 index, and 0 means that the stock's 5-year rate of return is lower than the S&P 500 index. The following is the logistic regression model:

$$pi(x_i)=P\ (y_i =1:\ x_i) =[1+\exp(-X^T\beta]^{-1} \quad (1)$$

In this formula if the rate of return of the stock exceeds the rate of return of the S&P 500 index then y_i is equal to 1. Conversely, if the rate of return of the stock is less than or equal to the rate of return of the S&P 500 index, then y_i is equal to 0.

$$X\beta =\beta 0 +\beta 1x1 +\beta 2x2 +.....+\beta nxn \quad (2)$$

β_0 is intercept. x_1, x_2, ... x_n are independent variables. β_1, β_2, ... β_n are coefficients [4].

According to Kleinbaum et al., Odds ratios are necessary for the logistic regression model [13], in this paper odds ratios represent the likelihood that the stock 5-year rate of return will above or below the S&P 500 index's rate of return. In logistic regression, odds ratios for a variable show how the probabilities alter when that variable is increased by one unit while keeping all other variables constant [4]. In this essay, odds refer to the ratio of the likelihood that a specific result will be higher or lower than the rate of return for the S&P 500 index.

$$pi(x)/pi(1-x) =[\exp(-X\beta]^{-1}=odd \quad (3)$$

Implementing a natural log for both sides of the equation.

$$Log\ [pi(x)/pi(1-x)] =[(-X\beta]^{-1} \quad (4)$$

Logistic transformation is the one mentioned above [4].

3.2 Sampling Framework

In this paper, the author started with more than 5000 listed companies in U.S. stock market from 2015 to 2021. On the other hand, several businesses changed every year. Thus, when all the businesses from 2015 to 2021 are combined, the author can exclude the companies with lost data and outliers, finally, 2959 companies were left with complete and normal data that could be used to build the final model.

Long term rate of return is a valid indicator of stock performances, but different companies have diverse dates of collecting financial data ranging from the beginning of the year to end of the year, and the S&P500 rate of return since the beginning of 2022 does not beat the inflation, so, it is not reasonable to use the S&P 500 5-year rate of return on one specific date as the standard of "outperform or not". Instead, the author chooses the median of S&P 500 5-year rate of return in recent 35 years as the standard of classify 1 and 0.

The paper also tries to find the most influential five financial ratios while estimating stock performance. The two-category division of stock performance is displayed in Board A of Table 1. The seven Independent Variables that the author examines every year from 2015 to 2021 are displayed in Board B of Table 1. Stock performance was divided into two categories, 0 and 1, to represent the companies' performance listed on the U.S. stock market.

Table 1. Summary of variables

Board A: Dependent Variables. Final Results of prediction	
Outperform (coded as 1)	Rate of return above S&P 500
Not outperform (coded as 0)	Rate of return below S&P 500
Board B: Independent Variables (Financial and Accounting Ratios)	
Variables' Names	Corresponding Description
PE	Price to earnings ratio
PB	Price to book ratio
GPAsset	Gross profit to asset ratio
Dividend yield	Cash Dividend per share / Market Price per share
epsfx	Earnings per share
ROE	Return on Equity
Revenue growth rate	Percentage change of total revenue

4. Results

4.1 Logistic Regression Model Results

Utilizing maximum likelihood estimation, the final logistic regression equation is calculated for categorizing stock performance:

$$Z_{it} = \beta_1 + \beta_2 PE2021_{it} + \beta_3 PB2021_{it} + \beta_4 GPAsset2021_{it} + \beta_5 Dividend\ yield2021_{it} + \beta_6 epsfx2021_{it} + \beta_7 ROE2021_{it} + \beta_8 Revenue\ growth\ rate2021_{it} + \beta_9 PE2020_{it} + \ldots\ldots + \upsilon_{it} \quad (5)$$

where:

z= log [p/ (1− p)] and 'p' is the likelihood that the result is "outperform" [4].

4.2 Confusion Matrix

4.2.1 Confusion Matrix Methodology

The author uses confusion matrix to show the performance of the final prediction results.

A confusion matrix often provides details about the present and anticipated categories provided by a categorization system [14]. Statistical metrics can be used to evaluate classification algorithms in terms of correct prediction rate [15]. They can be displayed in a confusion matrix like in table 2:

Table 2. Confusion Matrix

	Predict: Negative	**Predict: Positive**
Observe: Negative	TN	FP
Observe: Positive	FN	TP

The terminologies in confusion matrix can be defined as follows:

True Positives (whose abbreviation is **TP)**: both the forecast and the result are positive. **False Positives (**whose abbreviation is **FP):** when the forecast is positive, but the result is negative. **True Negatives (**whose abbreviation is **TN)**: both the forecast and the result are shown to be negative**. False Negatives (**whose abbreviation is **FN)**: when the forecast is negative, but the result turns out to be positive [15].

Then some indices can be calculated based on confusion matrix to show the performance of final model:

How often a classifier is accurate is measured by **accuracy**. **Sensitivity, Recall,** or **True Positive Rate** can be used to describe how many numbers the model can capture, which is the quantity. It is calculated as the amount of correctly categorized positive cases divided by the entire amount of actual positive cases. **Precision** measures how discerning the model is, which is the quality of one model, by dividing the amount of accurately categorized positive cases by the entire amount of instances that are predicted to be positive.

Machine learning predicts the label based on the decision threshold. By changing the threshold, the prediction labels can be changed towards more "positive" or more "negative". In python, the default threshold of all logistic models is 0.5, which doesn't show any preference towards "positive" or "negative". When decision threshold is adjusted, different confusion matrixes will be determined and thus different precision and recall. For instance, if the prediction threshold of being 1 is increased from 0.5 to 0.7, then an example will be more difficult to be classified as 1, thus leads to different confusion matrix and overall accuracy. Based on this characteristic, Receiver Operating Characteristic, which is also called ROC curve, is created to demonstrate the performance of a binary categorization model over different decision thresholds. Area Under Curve, whose abbreviation is AUC, is invented to measure the full two-dimensional area underneath the complete ROC curve (integral calculus) from (0,0) to the maximum value (0,1). The size of AUC shows a strong positive correlation with the performance of the model.

4.2.2 Confusion Matrix Final Results

The author uses the test set confusion matrix under default threshold (0.5) to obtain the necessary indices as shown in table 3:

Table 3. Confusion matrix and overall performance of test set under default threshold

		Performance Predicted	
		Not Outperform	Outperform
Actual Performance	Not Outperform	226	88
	Outperform	82	196
Overall Accuracy		71.28%	
AUC		0.7124	
Precision		69.01%	
Recall		70.50%	

The final logistic model has both satisfactory precision and recall, which means Many results can be returned, and the majority of the projected labels are accurate. The final result predicts overall 71.28% companies accurately. The AUC of this model is also above 0.5, which shows the validity of model.

4.2.3 Different Thresholds and Corresponding Overall Accuracy

The author also cares about applying this model to investors with different risk preferences. According to the characteristic of confusion matrix and logistic regression model, if the threshold of being 1 is increased, a higher Precision and a lower Recall can be achieved, which means the investors care more about a higher probability of correct predictions among all positive predictions (quality) than the percentage of stocks with high returns that are screened out (quantity), investors would rather miss some of the high return stocks than mistakenly select "not outperform" stocks as "outperform" ones, thus they are more cautious in pursuing an "outperform" stock and risk-averse in this case. So, by adjusting decision threshold, different risk preferences can be stimulated, and different accuracy can be obtained accordingly as shown in table 4:

Table 4. Different overall accuracies according to different decision thresholds

Decision Threshold	Overall Accuracy
0.3	50.34%
0.4	52.87%
0.5	71.28%
0.502	71.45%
0.6	64.86%
0.7	59.29%
0.8	56.76%

It is obvious that the overall accuracy achieves its highest when decision is a little bit more than 0.5, which means this model works best when applied to investors who are slightly risk averse.

4.3.4 Five Most Influential Ratios and Coefficients

This paper also tries to find the most influential 5 financial ratios when forecasting stock performance, The outcomes are displayed in table 5:

Table 5. 5 most influential ratios and their coefficient

Ratios	**Coefficient**
epsfx2021	0.1030
Revenue growth rate 2016	-0.0872
epsfx2018	0.0981
epsfx2017	-0.0903
epsfx2016	-0.1617

According to the results of the above table, it's obvious that the five ratios with the highest absolute value of coefficient, which means that the five ratios that have the greatest impact on the prediction results, are the annual earnings per share and the revenue growth rate. These ratios can be classified

as the profitability and historical growth of one company, respectively, and it can be concluded that these two ratios have a relatively important role in predicting the future performance of the stock.

5. Conclusion

In order to forecast the performance of more than 5000 listed companies in the U.S. stock market using accounting and financial ratios, this study developed a logistic regression model. This paper collected the representative financial and accounting ratios of listed companies and constructed a logistic regression model to these data to forecast whether the stocks will outperform, namely, whether the rate of return of listed companies will exceed the S&P 500 index. As a result, the accuracy of the model is 71.28% under default decision threshold. The accuracy of satisfaction is achieved under different decision thresholds and when the threshold is slightly higher than 0.5, which means that investors are a little bit more cautious in pursuing higher precision, the highest overall accuracy (71.45%) can be obtained. The innovation of this research is that it can not only help investors and investment institutions to make correct investment judgments about stocks' future performance, but also can be used for investors with different risk preferences. This model is extremely helpful for investors to predict whether one stock can outperform the whole market in 5 years' time frame. In order to forecast stock performance, diverse financial ratios i.e. Price to book ratios, price to earnings ratios, gross profit to assets ratios, return on equity, dividend yield, earnings per share, and revenue growth rate were utilized. The final model's accuracy level, according to this paper's findings, was 71.28%.

In addition, this model has some shortcomings as well. Firstly, this paper simply and intuitively used decision threshold to represent investors risk preference. Although the decision thresholds and risk preferences must have a strong positive correlation according to the logic of logistic regression model, the specific algebraic relationship needs more rigorous research to reach a conclusion, and then this relationship can be applied to the existing model. In the future, the mathematical expression of specific investors risk appetite can be obtained by collecting the respective decision thresholds and other related behavioral data of investors with different risk preferences, then how this model can be better applied to investors with different risk preferences could be precisely studied. Secondly, the application of this model requires the use of a large amount of data for each of the years from 2015 to 2021 to accurately predict whether a stock will outperform using the median of S&P 500 index 5-year rate of return as a benchmark, which represents a high demand for accurate predictions on the number of independent variables, and it may be difficult to achieve in the real market. Therefore, in the future, a more detailed study based on variables' coefficient can be conducted to selectively screen these variables to minimize the number of input data required without disturbing the accuracy, in order to improve the model's ability to be applied to the real market.

References

[1] Falinouss, P. (2007). Stock trend prediction using news articles: a text mining approach.

[2] Mubin, M., Iqbal, A., & Hussain, A. (2014). Determinant of return on assets and return on equity and its industry wise effects: evidence from KSE (Karachi Stock Exchange). Research Journal of Finance and Accounting, The International Institute for Science. Technology and Education, 5(5), 148-158.

[3] Wijaya, M., & Yustina, A. I. (2019). The impact of financial ratio toward stock price: evidence from banking companies. JAAF (Journal of Applied Accounting and Finance), 1(1), 27-44.

[4] Ali, S. S., Mubeen, M., Lal, I., & Hussain, A. (2018). Prediction of stock performance by using logistic regression model: evidence from Pakistan Stock Exchange (PSX). Asian Journal of Empirical Research, 8(7), 247-258.

[5] Smita, M. (2021). Logistic regression model for predicting performance of S&P BSE 30 company using IBM SPSS. International Journal of Mathematics Trends and Technology, 6(7), 118-134.

[6] Altman, E. I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. Journal of Finance, 23, 589-609.

[7] Abdel-Khalik, A. R., & Lusk, E. J. (1974). Transfer pricing-a synthesis. The Accounting Review, 49(1), 8-23.

[8] Fama, E., & French, K. (1988). Permanent and temporary components of stock prices. Journal of Political Economy 96, 246-273.

[9] Lee, S. (2004). Application of likelihood ratio and logistic regression models to landslide susceptibility mapping using GIS. Environmental Management, 34(2), 223-232.

[10] Dutta, A., Bandopadhyay, G., & Sengupta, S. (2008). Classification and prediction of stock performance using logistic regression. An Empirical Examination from Indian Stock Market: Redefining Business Horizons: McMillan Advanced Research Series, 46-62.

[11] Dutta, A., Bandopadhyay, G., & Sengupta, S. (2015). Prediction of stock performance in Indian stock market using logistic regression. International Journal of Business and Information, 7(1), 105- 136.

[12] Novy-Marx, R. (2013). The other side of value: The gross profitability premium. Journal of financial economics, 108(1), 1-28.

[13] Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2014). Applied regression analysis and other multivariable methods. Cengage Learning. 5th Edition.

[14] Kohavi, R., & Provost, F. (1998). Confusion matrix. Machine learning, 30(2-3), 271-274.

[15] Maria Navin, J. R., & Pankaja, R. (2016). Performance analysis of text classification algorithms using confusion matrix. International Journal of Engineering and Technical Research (IJETR), 6(4), 75-8.