Stock Price Prediction Using Stepwise Regression and Improved with Factor Analysis

Qiang Dai[1,a,*], Yantong Liu[2,b], Kaiyin Cai[3,c], and Chunlin Jia[4,d]

[1]School of International Business and Economics, University of International Business and Economics, Peking, 100029, China
[2]School of Business, City University of Macau, Macau, 411105, China
[3]School of Overseas Education, University of Jimei, Xiamen, 361000, China
[4]School of International Business, University of Shandong Technology and Business, Yantai, 264005, China
a. Roronoa37@163.com, b. 3293452102@qq.com,
c. 1161016739@qq.com, d. 1799861872@qq.com
**corresponding author*
Qiang Dai and Yantong Liu should be considered as co-first authors, and
Kaiyin Cai and Chunlin Jia should be considered as co-second authors.

Abstract: Owing to volatility in stock markets, it is quite elusive to forecast stock prices. Albeit, sometimes regular patterns are manifested in stock prices and a variety of factors are proved to be competent to determine stock prices partly. Hence, using stepwise regression on historical stock price data, this paper proposes determining similar patterns in stock prices and exploring potential rules to select the main factors that can affect stock prices significantly while taking all factors into account. Difference analysis is also employed to probe possible correlations in the data. Eventually, this paper tries to improve stock price prediction using factor analysis and manages to achieve higher accuracy.

Keywords: stock price prediction, stepwise regression, comparison analysis, factor analysis

1. Introduction

As a type of marketable securities, stock has become one of prevalent ways of financial management, and attracts more and more investors to trade stocks. Investing in stocks is accompanied with high risks and high return simultaneously. Among those risks, systematic risks refer to the loss caused by changes in the state of the whole system, which are incurred by variation in the structure, functions or environment of the system itself. Normally, investors can only control risks but are not able to determine the probability of risks. The stock market is a complex and highly uncertain system, where it is difficult to invest stocks safely and profitably. However, researchers still try to use a variety of methods to predict stock prices. Most of existing empirical studies on the effectiveness of stock markets use statistical test methods under the condition of linear paradigm.

Frank A. Fetter's [1] book The Nature of Capital and Income systematically examined the connection between capital and income. Williams [2] followed up on Fisher's research with his own model of dividend discounting, arguing that investors speculate on stocks in order to earn future dividends. He argued the intrinsic value of a stock is equal to the present value of all the combined expected dividends. There are also scholars who have expanded on Williams' theory. For example, Gordon [3] published an article entitled "Dividends, Earnings and Stock Prices", which proposed a dividend discount model with fixed dividend growth, followed by a two-stage growth model. The theory of capital valuation was refined step by step. Pratt, Sereno S [4] proposed that the reason for the existence of stock value is that investors want to receive expected dividend income. And there are many factors that affect the share price. These include the dividend, the company's earnings, the quality of the manager, and the company's earning power, among others. She also pointed out that in theory, the intrinsic value of a stock should be in line with the market price. Graham and Dodd [5] have systematically elaborated on the intrinsic value of stocks, and they both point out that the intrinsic value of a stock is determined by the future profitability expectations of the company.

Wang Zhaodong [6] used a multi-factor model based on factor ranking in "An empirical analysis of multi-factor stock selection models in the Chinese stock market". The main idea is to select the top-ranked stocks to build a portfolio by ranking multiple factors separately according to their sizes. The model assumes that there is a correlation between the size of the large subsets and the future returns of the stocks. Whether this correlation is positive or negative, as long as the correlation is confirmed, it can be used as a basis for stock selection.

The six-factor quantitative stock selection model proposed by Zhengfeng Cao, Hong Ji and Bangchang Xie [7] in "Using Random Forest Algorithm to Achieve Quality Stock Selection" is constructed from six indicators: P/E ratio, P/E ratio, ROA, stock return in the previous month, EPS consensus expectation change rate and EPS consensus expectation. The first two indicators are value factors, which are commonly used indicators to measure the reasonableness of the market and stocks, and are retained by Cao Zhengfeng. The last four indicators reflect the concept of growth factors and are excluded due to the poor stability of stock returns in the previous month, retaining the three indicators of ROA, EPS Consensus Expected Rate of Change and EPS-Consistent Expectations. Five indicators were given and they or a linear combination of them can be frequently found in stock selection models.

In this work, we employ stepwise regression to predict stock prices and evaluate the accuracy. Besides, using the regression outcome and comparison analysis, we explore potential rules for selecting the main factors that significantly affect stock prices. Eventually, factor analysis is conducted and we test whether this could improve stock price prediction.

This paper is organized as follows: Section 2 briefly introduces the dataset and the methodology used in this paper; Section 3 presents the regression result and discussions; Section 4 provides some further exploratory analyses into the dataset using comparison analysis; Section 5 uses factor analysis to improve the stock price prediction for higher accuracy; and, at last, the main conclusions and contributions of this paper are summarized in Section 6.

2. Data and Methodology

2.1. Dataset

The dataset used in this paper is NY Stock Price Prediction RNN LSTM GRU from Kaggle and close price is chosen as the dependent variable. After eliminating missing values, there are 1016 observations left and 75 dependent variables

2.2. Methodology

Firstly, this paper uses regression for stock price prediction. Specifically, the regression method used is stepwise and the entrance and exit criterion of p-value are 0.05 and 0.10 respectively throughout this paper. Then, t-test, Levene’s variance and chi-square test are used for comparison analysis.

Finally, factor analysis is used to improve the stock price prediction. All of the tests are conducted in SPSS software.

2.3. Descriptive Statistics

Table 1 shows the descriptive statistics about the independent variable, price, and a part of the dependent variables selected by stepwise regression. Overall, these variables have an extensive range. Besides, the data is reasonable and matches well with the reality.

Table 1: Descriptive statistics.

Variables	Number of cases	Average value	Standard deviation	Median	Minimum value	Maximum value
Price	1357.000	73.883	82.026	56.320	4.300	1274.900
Earnings Per Share	1274.000	3.322	4.683	2.850	-61.200	50.100
Operating Margin	1357.000	19.030	21.799	16.000	0.000	437.000
Cash Ratio	1085.000	69.800	94.058	39.000	0.000	1041.000
For Year	1357.000	2013.770	1.132	2014.000	2012.000	2016.000
Short-Term Investments	1357.000	961172411.200	5179007604.707	0.000	0.000	107000000000.000
Income Tax	1357.000	656711414.890	1850764574.413	264000000.000	-8013000000.000	31045000000.000
Profit Margin	1357.000	14.520	18.075	11.000	0.000	369.000
Long-Term Investments	1357.000	28899093921.890	158571729628.345	201000000.000	0.000	1650000000000.000

3. Regression Analysis

This part conducts a stepwise regression on raw data for stock price prediction and the outcome is shown in Table 2. And then we evaluate the stock price prediction and explore potential rules existing in the outcome.

Table 2: Regression on raw data.

Adjusted R-squared	0.564								
Sig.	0.000								
Std. Error of the Estimate	60.440								
N	1016								
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
	B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF

Table 2: (continued).

(Constant)	-12860.242	3437.625		-3.741	0	-19606.037	-6114.447		
EarningsPerShare	16.089	0.49	0.892	32.812	0	15.126	17.051	0.581	1.721
OperatingMargin	1.705	0.221	0.425	7.71	0	1.271	2.139	0.141	7.071
NetIncome	-1.356E-08	0	-0.541	-4.936	0	0	0	0.036	27.99
GrossProfit	2.671E-09	0	0.397	7.951	0	0	0	0.172	5.818
Liabilities	1.806E-08	0	0.201	8.706	0	0	0	0.802	1.247
RetainedEarnings	6.784E-10	0	0.219	4.857	0	0	0	0.211	4.73
CashRatio	0.09	0.022	0.092	4.116	0	0.047	0.133	0.864	1.157
ForYear	6.385	1.707	0.079	3.741	0	3.036	9.735	0.973	1.028
ShortTermInvestments	8.93E-10	0	0.056	2.09	0.037	0	0	0.596	1.679
Investments	-4.716E-09	0	-0.128	-3.956	0	0	0	0.407	2.457
IncomeTax	1.183E-08	0	0.247	3.298	0.001	0	0	0.076	13.107
OtherInvestingActivities	-2.51E-09	0	-0.097	-2.825	0.005	0	0	0.366	2.732
EarningsBeforeInterestandTax	-7.458E-09	0	-0.456	-2.849	0.004	0	0	0.017	59.712
ProfitMargin	-0.752	0.272	-0.154	-2.77	0.006	-1.285	-0.219	0.139	7.172
ShortTermDebtCurrentPortionofLongTermDebt	-2.678E-09	0	-0.081	-2.706	0.007	0	0	0.475	2.107
LongTermInvestments	1.409E-09	0	0.071	2.026	0.043	0	0	0.349	2.865
a Dependent Variable: Price									

3.1. ANOVA

From Table 2, F statistic is 0.000, which is statistically significant and indicates that the regression is statistically meaningful. Besides, adjusted R-squared is 0.564 and shows that the stock price prediction has a relatively good accuracy wholistically.

3.2. Collinearity Test

Table 2 shows that only 3 of 16 independent variables has a VIF larger than 10, which suggests that there exists little collinearity problem in the regression overall. On the other hand, to some extent, the three independent variables with a VIF larger than 10 indicate that factor analysis stands a chance to improve the stock price prediction by eliminating collinearity problems.

3.3. Explanation on Coefficients

We try to explore potential rules existing in the regression in Table 2 and explanation to some coefficients of dependent variables is given as follows.

For *Earnings Per Share* (EPS), the sign is positive in the regression. EPS is the portion of a company's profit allocated to each outstanding share of common stock, serving as a profitability indicator. High EPS indicates an excellent capacity to make high profit per unit of capital and shows that the company is equipped with outstanding competence in product marketing, technical skills, management skills and so on that allow the company to generate high profit with finite resources. Earnings per share usually represents the amount of dividends that can be distributed during the year.

For *Operating Margin*, the sign in the regression is positive. Operating margin is the ratio of an enterprise’s operating profit to its operating revenue. A high operating margin indicates that the enterprise earns more profits, has strong profitability and is promising to future growth. Moreover, a high operating margin will make stocks more valuable to invest, and hence a high stock price.

For *Net Income*, the sign in the regression is positive. Net income is the total profit of an enterprise minus its income tax, where income tax is the tax calculated and paid by an enterprise to the state on the total realized profit according to the standards stipulated in the income tax law. It is a deduction from the total profit of an enterprise [8]. The higher the net income, the better the company performs, which will attract the market investors to pay more attention to the company’s stock, endowing the stock price with the potential to rise.

For *Gross profit*, it is positively correlated to stock prices. Gross profit is defined as the accounting revenue from sales minus the direct cost of the main business. A company's gross profit is usually reflected through gross profit margin, which is positively correlated. Besides, gross profit is the basis of net profit, and the level of gross profit of a company directly determines the profitability and affects the stock price.

For *Cash Ratio*, it is positively correlated to stock prices. The cash ratio is a measurement of a company's liquidity. It specifically calculates the ratio of a company's total cash and cash equivalents to its current liabilities. The metric evaluates a company's ability to repay its short-term debt with cash or near-cash resources, such as easily marketable securities. This information is useful to creditors when they decide how much money, if any, they would be willing to loan a company [9]. The higher the cash ratio, the more liquid the assets are and the better the ability to repay in short term. Besides, the amount of cash capacity is also indicative of a company's operating conditions. In addition, inventory turnover ratio and accounts receivable turnover ratio can also be used as indicators to supplement a company's short-term solvency [10]. Overall, higher cash ratio somehow ensures investors that the company is operated better, resulting in a higher stock price.

For *Short-Term Investments*, the sign in the regression is positive. Short-term investments are various marketable securities that can be easily realized and held within a year, as well as other investments that do not last over one year [11]. When companies have a temporary surplus of cash, they will use some funds as short-term investments and choose to buy more liquid stocks and bonds to invest for more return. Hence, short-term investments have a positive effect on stock prices.

4. Comparison Analysis

This part is using comparison analysis to explore potential correlations between variables.

4.1. Earnings Per Share and Stock Price

From Table 3, the mean difference of T-Test test of variance is -55.31 (group 0 - group 1), sig=0.000<0.001, so the mean difference of Earnings Per Share between the two groups is significant.

Table 3: EPS and Price' s T-Test.

Group statistics								
Earnings Per Share_C		Number of cases	Average	Standard Deviation	Standard Mean error			
Price	0	825.000	52.197	42.625	1.484			
	1	532.000	107.514	111.796	4.847			
Independent sample test								
		Levene's test for equivalence of variances		Mean equivalent T-Test				
		F	Significance	t	Freedom	Sig. (bobtail)	Mean Difference	Standard Error Deviation
Price	Assuming equal variance	58.870	0.000	-12.840	1355.000	0.000	-55.317	4.308
	Does not assume equal variance			-10.913	631.640	0.000	-55.317	5.069

As can be seen from the general form of EPS, earnings per share are affected by the following indicators: net asset value per share, equity multiplier. It can also introduce marginal contribution rate, operating leverage and financial leverage into EPS. The effect of EPS is actually the result of the comprehensive effect of various factors. For example, a large equity multiplier indicates that an enterprise which has a large equity multiplier, which generally result in a greater financial risk, decreasing its stock price. When the enterprise’s ROA is higher than the cost of capital rate, the borrowed funds will produce tax avoidance effect and then improve enterprise’ s EPS. Meanwhile, not only the increase of leverage ratio will increase the enterprise value with the increasing debt, but also it will increase the potential of bankruptcies of enterprises, which will reduce the market value of companies, thus decreasing the stock price.

4.2. Earnings Per Share and Cash Ratio

From Table 4, the asymptotic significance was calculated from the chi-square test = 0.009<0.01, so the difference is significant and Cash Ratio and Earnings Per Share are significantly correlated.

Table 4: EPS and Cash Ratio.

Counting					
		Cash Ratio_C		Total	
		0	1		
Earnings Per Share_C	0	598	227	825	
	1	419	113	532	
Total		1017	340	1357	
chi-square test					
	Value	Freedom	Progressive saliency (bilateral)	Precise Significance (Bilateral)	Precise Significance (unilateral)
Pearson Cardinal	6.781[a]	1	0.009		
Continuity correction[b]	6.451	1	0.011		
Seemingly more than	6.879	1	0.009		
Fisher Precision Test				0.01	0.005
Linear correlation	6.776	1	0.009		
Number of active cases	1357				

A strong cash flow of a company is largely an indication that the company's products are competitive and that its main business has potential for growth. Cash ratios are usually higher than EPS because it is reduced by depreciation.

The cash ratio is only a supplementary measure of earnings per share. Although a higher cash ratio indicates a more solvent business, it does not mean that a higher indicator is more favorable for the development of the business. We measure the solvency of a business. As long as a business is able to pay its debts which fall due, this will not result in debt maturity and the business will be able to liquidate its capital funds. If the cash ratio of a business is too high, it can result in a situation where idle funds are not generating income, thus decreasing earnings per share.

4.3. Short-term Investment Ratio and Stock Price

From Table 5, the difference in means was calculated from the Levene's variance matrix equivalence test as 19.22 (Group 0 - Group 1), sig=0.000<0.001, so the difference between the means of two groups is significant.

Short-term investment is based on market price fluctuations. It is based on trend opportunities arising from the volatility of the market itself. For short-term investors, it doesn't matter whether a company's performance is good or bad, or whether the dividend is high or low.

Table 5: Short-term investment and price.

Group statistics								
Short-Term Investments _C		Number of cases	Average	Standard Deviation	Standard Mean error			
Price	0	1177	71.424	60.068	1.751			
	1	180	89.964	164.207	12.239			
Independent sample test								
		Levene's test for equivalence of variances		Mean equivalent T-Test				
		F	Significance	t	Freedom	Sig. (bobtail)	Mean Difference	Standard Error Deviation
Price	Assuming equal variance	28.081	0	-2.831	1355	0.005	-18.54	6.548
	Does not assume equal variance			-1.5	186.389	0.135	-18.54	12.364

Although short-term investing is highly speculative, investment experts often regard it as a major factor in stock price movements. Investing for the long term, we buy stocks just sitting on the dividends. We're not going to have much of a rise and fall in stock prices which are disguised bonds.

4.4. Retained Earnings and Stock Price

From Table 6, we calculated asymptotic significance = 0.003<0.01 from the chi-square test, so the difference is significant.

The reason why evaluating retained earnings is important is that shareholders' earnings include dividends and retained earnings, and retained earnings not only account for a large proportion of earnings, but also have a long payback period.

When a company has retained earnings, management is usually faced with several options:

1. use the money to continue to expand investments in order to obtain a higher ROE than the existing assets.

2. distribute cash dividends to shareholders, who have better access to investment and better returns.

3. Neither invest nor distribute the dividends, but rather keep a book of accounts, or manage the money.

Table 6: Price and RE.

Cross-tabulation					
Price_C * Retained Earnings_C Cross-tabulation					
Counting					
		Retained Earnings_C		Total	
		0	1		
Price_C	0	700.000	193.000	893.000	
	1	330.000	134.000	464.000	
Total		1030.000	327.000	1357.000	
Chi-Square Test					
	Value	Freedom	Progressive saliency (bilateral)	Precise Significance (Bilateral)	Precise Significance (unilateral)
Pearson Cardinal	8.815a	1.000	0.003		
Continuity correctionb	8.423	1.000	0.004		
Seemingly more than	8.656	1.000	0.003		
Fisher Precision Test				0.003	0.002
Linear correlation	8.809	1.000	0.003		
Number of active cases	1357.000				
a. 0 cells (.0%) have an expected count of less than 5. The minimum expected count is 111.81.					
b. Calculation for 2x2 tables only					

The best options for shareholders and management are as follows:

If they choose to continue investing, then new cash flows will be generated:

If it is not a good investment, it will be a drag on existing ROE, resulting in lower cash flow per share (lower equity value) and ultimately reflected in the share price, to the detriment of shareholders' equity.

A good investment, in addition to increasing return on equity, can also increase cash flow per share (i.e. increase in equity value), as Warren Buffett sets a standard: for every dollar of retained earnings, there is at least a dollar of market value.

5. Factor Analysis

This part is trying to use the method of factor analysis to improve the fitness of stepwise regression for higher prediction accuracy.

5.1. KMO and Bartlett’s Test

Table 7: KMO and Bartlett’s test.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.725
Bartlett's Test of Sphericity	Approx. Chi-Square	208651.929
	df	2628
	Sig.	.000

As Table 7 shows, the KMO statistic is larger than 0.7, indicating great correlation among variables. Besides, the result of Bartletts’ test rejects the null hypothesis that variables are independent with each other. These two tests both show that factor analysis is applicable to the dataset in this paper.

5.2. Factor Extraction

Table 8: Total Variance explained.

Initial Eigenvalues				Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	%of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	27.093	36.124	36.124	27.093	36.124	36.124	22.092	29.457	29.457
……	……	……	……	……	……	……	……	……	……
21	0.722	0.963	88.509	0.722	0.963	88.509	1.119	1.492	77.638
……	……	……	……	……	……	……	……	……	……
41	0.140	0.187	98.608	0.140	0.187	98.608	0.255	0.339	97.562
……	……	……	……	……	……	……	……	……	……
58	0.012	0.016	99.981	0.012	0.016	99.981	0.018	0.024	99.981
59	0.004	0.005	99.986						
……	……	……	……						
75	-4.40E-16	-5.87E-16	100						

Table 9: Communalities.

	Initial	Extraction
AccountsPayable	1.000	1.000
Goodwill	1.000	1.000
……	1.000	>=.998
OperatingIncome	1.000	.999
EstimatedSharesOutstanding	1.000	1.000

The extraction method used here is principal component analysis and the criterion of factor extraction is that eigenvalues have to be larger than 0.01, whose outcome is shown in Table 8 and Table 9. As Table 8 indicates, eventually 58 factors are extracted out of all 75 independent variables and these 58 factors are able to explain 99.981% of primitive independent variables. Besides, Table 10 illustrates that the extracted factors are capable of extracting more than 99.8% out of each and every of dependent variables. All of these outcomes prove that the extracted factors have an excellent capacity of explanation for primitive dependent variables both wholistically and respectively, which means the extracted factors are undoubtedly competent to replace previous independent variables for regression.

5.3. Regression on the Extracted Factors

We conduct a stepwise regression using the extracted factors and the outcome is given in Table 10, which shows that 34 dependent variables are included this time, compared to 16 dependent variables included in the previous regression using the raw data. The noticeable variation of the number of dependent variables included is probably incurred by the factor extraction, through which the collinearity of previous data is eliminated partly and hence the extracted factors with less internal correlation have more opportunities to be preserved in the stepwise. Besides, Table 10 also shows a larger adjusted-R2 of 56.7 % compared with previous 56.4%, which is attributed to the elimination of the redundant variables and overlapped part of dependent variables, with the prerequisite of conserving most information of data. Therefore, after extracting factors of previous dependent variables, the stepwise regression outperforms for a more accurate prediction.

Table 10: Regression on extracted factors.

Adjusted R-squared	0.567						
Sig.	0						
Std. Error of the Estimate	60.259						
N	1016						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	77.767	1.891		41.136	0	74.057	81.477

Table 10: (continued).

REGR factor score 26 for analysis 1	33.645	1.891	0.367	17.788	0	29.933	37.357
REGR factor score 27 for analysis 1	25.287	1.891	0.276	13.369	0	21.575	28.999
……	……	……	……	……	……	……	……
REGR factor score 16 for analysis 1	4.171	1.891	0.046	2.205	0.028	0.459	7.883
REGR factor score 41 for analysis 1	4.139	1.891	0.045	2.188	0.029	0.427	7.851

6. Conclusion

This paper uses stepwise regression to predict stock price and achieve a relatively good accuracy. Through the regression outcome and difference analysis, we explore and verify some potential rules for selecting the main factors that significantly affect stock prices. And eventually, we prove that using factor analysis, stock price prediction can be improved for a higher accuracy, which is a useful method that can be referred in following studies.

References

[1] *Fetter, F. A. (1907) The nature of capital and income. Journal of Political Economy, 15(3), 129-148.*

[2] *Williams, J. B. (1938) The Theory of Investment Value. Harvard University Press, Cambridge, MA.*

[3] *Gordon, M. J. (1959) Dividends, earnings, and stock prices. The review of economics and statistics, 99-105.*

[4] *PRATT, S. (1903) The Work of Wall Street. D. Appleton, New York.*

[5] *Graham, B., Dodd, D. L. F., Cottle, S., Murray, R. F., & Block, F. E. (1999) Security analysis. Hainan Publishing House.*

[6] *Wang Zhaodong (2014) An empirical analysis of multi-factor stock selection model in Chinese stock market. (Doctoral dissertation, Shandong University).*

[7] *Cao Zhengfeng, Ji Hong, & Xie Bangchang. (2014) The random forest algorithm is used to select high quality stocks. Journal of Capital University of Economics and Business, 16(02):21-27.*

[8] *Shen Tao - Accounting for Logistics Enterprises - Lixin Accounting Press – 2005*

[9] *Will Kenton (2022): https://www.investopedia.com/terms/c/cash-ratio.asp*

[10] *Hua Wei Real estate finance Fudan University Press, 2004. Jan:76*

[11] *Liao Hong Principles of Accounting first edition, Wuhan University Press·2002*