Stock Price Prediction Based on Multiple Linear Regression Model

Runqing Hu[1,*]

[1]College of Engineering, University of California, Santa Barbara, Goleta, California, 93106, USA

*jinhua@ucsb.edu

Abstract. In the modern society, the rise and fall of stocks price have become the most discussed topic among people. It is difficult to make precise stock buying and selling decisions based on personal experience in current stock market. Statistics and programming can effectively solve this problem. In machine learning field, there are many models that can be utilized to predict stock price like Recurrent Neural Network (RNNs), LSTMS, and regression. This article explores the utilization of the multiple linear regression model in prediction the stock price and use the Alphabet company as the example. All the data is extracted from Yahoo Finance. Initially, processing all the data by python pandas, numpy, and statsmodels library. Then, visualize the prediction result by matplotlib. In the end, the article obtained relatively accurate prediction results that much higher than the original expectation. There is a small difference between the prediction price with the actual price. Users can also use this model to predict other parameters while making discussions.

Keywords: stock price, machine learning, multiple linear regression, finance, statistics

1 Introduction

With the rapid and vigorous development of the global economy, the income and living conditions of people around the world are increasing year by year [1]. Accompanied with the increase in income, people's financial-management awareness is gradually strengthening. In addition to fixed income, people will choose other investments. Thus, more and more people choose to invest their assets in stocks and other financial derivatives in their spare time. The rise and fall of stock price have become one of the hottest topics discussed today.

Many beginners only buy stocks based on their own experience or simply look at the K-line chart. However, stock prices are constantly changing, making it difficult to judge the trend of stocks based on simple analysis. From a financial perspective, the rise and fall of a company's stock are influenced by both external and internal factors. The external influencing factors of enterprises include market factors, political factors, macroeconomic factors, industry factors, etc. The internal influencing factors of a company

A. Bhunia et al. (eds.), *Proceedings of the 2023 International Conference on Finance, Trade and Business Management (FTBM 2023)*, Advances in Economics, Business and Management Research 264,
https://doi.org/10.2991/978-94-6463-298-9_48

include its management capabilities, organizational structure, financial information, etc. But from a statistical perspective, quantifying existing indicators can help visualize future trends. External and internal factors are difficult to quantify. In this case, machine learning can effectively help to solve this problem.

Machine learning is one of the branches in the field of artificial intelligence to process existing datasets for future prediction. There are many models that could predict the stock price such as Recurrent Neural Networks (RNNs), LSTMS, and regression [2].

This paper will use a multiple linear regression model to analyze the relationship between the highest price tomorrow between six independent variables. The rest of the paper is arranged as follows: section II will introduce the concept of Multiple linear regression, provide the dataset, and finally evaluate its realizability. Section III is the conclusion that will summarize the application of this model and future approvement.

2 Method

2.1 Multiple linear regression introduction

In statistical area, MLR is defined as a technique that uses several independent variables to make prediction for the outcome of responsive dependent variables [3]. The mathematical formula is expressed shown below as (1):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_i x_i + \varepsilon_i \quad (1)$$

In this formula, β_o is the population Y-intercept. The β_1 to β_i are population slopes. The last term ε_i is referred to as the random variable.

2.2 Source of data

All the data was taken from Yahoo Finance. To ensure sufficient sample size, this paper chose the data from the past five years from August 27, 2018, to March 27, 2023. The chosen company is Alphabet Inc (GOOG). The total size of the data is 1256×6.

Dependent variable. There are six columns within the dataset, which will be regarded as six different independent variables: Open, High, Low, Close, Adj Close, and Volume.

Independent variable. In the contradictory, this article will set the corresponding highest price on the second day as the independent variable to explore the linear relationship with six dependent variables mentioned in “Dependent variable”.

2.3 Data processing

Based on the Yahoo Finance database, and the multiple linear regression model discussed in the last part, this paper will use the Python library to finish the data processing. Initially, going to Yahoo Finance website to search the Alphabet company. It will automatically generate CSV files that have six independent pieces of information. The CSV document needs to be modified to have an additional line called "high_tmr" as the dependent variable. The first 10 rows of processed csv file are shown below in Table 1.

Table 1. Google stock price from 2018/5/29 to 2018/6/11

Date	Open	High	Low	Close	Adj Close	Volume	High_tmr
2018/5/29	53.24	53.67	52.76	53.02	53.02	3.73 E+07	53.46
2018/5/30	53.15	53.46	52.84	53.39	53.39	2.28 E+07	54.86
2018/5/31	53.38	54.86	53.38	54.25	54.25	6.18 E+07	56
2018/6/1	54.97	56	54.92	55.97	55.97	4.84 E+07	57.09
2018/6/4	56.12	57.09	56.1	56.96	56.96	3.78 E+07	57.29
2018/6/5	57.05	57.29	56.66	56.98	56.98	3.36 E+07	57.15
2018/6/6	57.11	57.15	56.29	56.84	56.84	3.40 E+07	56.79
2018/6/7	56.57	56.79	55.83	56.19	56.19	3.04 E+07	56.33
2018/6/8	55.91	56.33	55.61	56.04	56.04	2.58 E+07	56.86
2018/6/11	55.93	56.86	55.93	56.5	56.5	2.16 E+07	56.99

2.4 Machine learning models

Initially, it will utilize the numpy and pandas libraries to convert CSV files into Pandas data frames indexed by date [4]. Next, define the six independent and dependent variables in the data frame named as $x_1, x_2, x_3, x_4, x_5, x_6, y$. Next, use the OLS model under the statsmoodels library in Python to find the most fitting model and derive the multiple regression equation. Moving to the next step, using the python sklearn library to split the dataset into two groups: a training set and a test set [5]. Finally, use the matplotlib to plot the prediction and actual stock price in the same diagram.

2.5 Evaluation metric

Coefficient of Multiple Determination. This parameter is utilized to determine the proportion of Y variation "explained" by all X variables [6]. The formula can be mathematically expressed below as (2)

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2} \tag{2}$$

For the value of the R^2, closer the value approaches 1, the better the fit of the entire model.

P test. The P value within the multiple linear regression model could determine whether the corresponding independent variable has a strong relationship with the corresponding independent variables. If the value of p is smaller than 5%, it will propose that the null hypothesis can be rejected [7]. In other words, if there is a p-value greater than 0.05, the deduced entire multiple linear regression equation should be reshaped and eliminate the independent variables, which are not significantly related to y linearly [8].

2.6 Study procedure

Initially, using the python special library to derive the MLR equation. Consequently, making sure all the p value in the derived model is smaller than 0.05 and evaluating the R^2 value. The next step is to eliminate non-related variables to repeat the entire process until it passes the test. Finally, establish the correlation matrix and use the matplotlib to visualize the prediction vs the actual highest stock price.

3 Result

3.1 Regression results

Table 2. OLS Regression Results of model information

Dep. Variables:	y	R-Squared:	0.998
Model:	OLS	Adj.R-squared:	0.998
Method:	Least Squares	F-statistic:	1.12E+05
Date:	Sun,28 May 2023	Prob(F-statistic):	0
Time:	23:35:37	Log-Likelihood:	-2241.2
No.Observations:	1257	AIC:	4494
Df Redisuals:	1251	BIC:	4525
Df Models:	5		
Covariance Type	nonrobust		

From Table 2 to Table 4, should focus on the column: “coef” initially. As this article mentioned in 2.1, the coef corresponds to the β_i, which is the population slopes and y-intercept. Based on the coef shown above, the multiple linear prediction formula could be derived as (3) shown below:

$$y = 0.0057 - 0.0575x1 + 0.3433x2 - 0.0673x3 + 0.3936x4 + 0.3936x5 + 3.675e^{-9}x6 \quad (3)$$

Moving to the coefficient of multiple determination, the $R^2 = 0.998$, which is extremely close to 1. High R^2 the value represents most y can be represented by x variables.

The next part is to verify the p-test. As discussed in the p-test. Any independent variables value greater than 0.05 in the column P>| t | should be removed. P values for x1(Open) and x3(Low) are much higher than 0.05, which means the original hypothesis will not be rejected. The entire regression model is rejected.

Table 3. OLS Regression Results of parameters information

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0057	0.199	0.029	0.977	-0.385	0.396
x1	-0.0575	0.065	-8.79E-01	0.379	-0.186	0.071
x2	0.3433	0.072	4.737	0	0.201	0.485
x3	-0.0673	0.077	-0.879	0.38	-0.217	0.083
x4	0.3936	0.032	12.169	0	0.33	0.457
x5	0.3936	0.032	12.169	0	0.33	0.457
x6	3.68E-09	3.86E-09	0.951	0.342	-3.91E-09	1.13E-08

Table 4. OLS Regression Results of testing information

Omnibus:	459.647	Durbin-Watson	2.03
Prob(Omnibus):	0	Jarque-Bera(JB)	5302.338
Skew:	1.355	Prob(JB)	0.00E+00
Kurtosis:	12.69	Cond.No.	1.97E+21

3.2 Regression model modification

According to the above section, since the original hypothesis will not be rejected, the MLR equation verification could not get through. It is necessary to modify the independent variables to make the P-values of multiple linear regression all less than 0.05: three independent variables x1, x3, and x6 should be eliminated. The revision dataset only has three variables: High, close and adjust close. The first 10 rows of the new dataset are shown in Table 5.

Next, it should move to fit the modified dataset with dependent variables using the OLS module. The summary of the regression model is shown in Table 5- Table 8:

Table 5. First 10 rows of modified dataset

	High	Close	Adj Close	High_tmr
0	53.67	53.02	53.02	53.46
1	53.46	53.39	53.39	54.86
2	54.86	54.25	54.2495	56
3	56	55.97	55.97	57.09
4	57.09	56.96	56.96	57.29
5	57.29	56.98	56.98	57.15
6	57.15	56.84	56.84	56.79
7	56.79	56.19	56.19	56.33
8	56.33	56.04	56.04	56.86
9	56.86	56.50	56.50	56.99

Table 6. Modified OLS Regression Results of model information

Dep. Variables:	y	R-Squared:	0.998
Model:	OLS	Adj.R-squared:	0.998
Method:	Least Squares	F-statistic:	2.776E+05
Date:	Mon,29 May 2023	Prob(F-statistic):	0.00
Time:	00:07:29	Log-Likelihood:	-2246.4
No.Observations:	1257	AIC:	4499.
Df Redisuals:	1254	BIC:	4514.
Df Models:	2		
Covariance Type	nonrobust		

Table 7. Modified OLS Regression Results of parameters information

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1635	0.128	1.274	0.203	-0.088	0.415
x2	0.2712	0.041	6.543	0	0.190	0.352
x4	0.3678	0.021	17.559	0	0.327	0.409
x6	0.3678	0.021	17.559	0	0.327	0.409

Table 8. Modified OLS Regression Results of testing information

Omnibus:	480.689	Durbin-Watson	1.931
Prob(Omnibus):	0.000	Jarque-Bera(JB)	5328.187
Skew:	1.448	Prob(JB)	0.00
Kurtosis:	12.662	Cond.No.	6.22E+15

After the overview information of the multiple linear regression model was readjusted, the P value of these three independent variables in the new model all pass the significance test. Simultaneously, the $R^2 = 0.998$ is a relatively high value. The final multiple linear prediction equation is expressed below as (4):

$$y = 0.1635 + 0.2712x2 + 0.3678x4 + 0.3678x5 \quad (4)$$

3.3 Correlation matrix visualization

To further verify the linearity between these three independent variables, the next step is to establish the correlation matrix. The correlation matrix will explore the linear correlation between dependent variable with independent variables [9]. The closer it is to 1, the better the linearity exists [10]. The matrix for these 4 elements in this model is presented in Fig. 1 below, it could derive the conclusion that the relationship between the highest stock price with the three independent variables is perfectly positively correlated since all 16 correlation values are equal to 1.

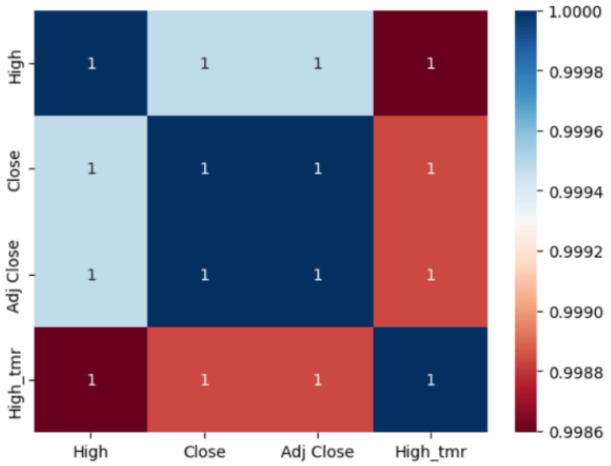

Fig. 1. Correlation metrics for modified dataset (Photo credit: Original)

3.4 Prediction verification

After completing multiple linear regression analyses and removing irrelevant factors from these variables, the next step should move to visualize the prediction result. As discussed in the section 2.4, the sklearn python library can finish the linear-regression process and make a comparison with the real stock price and the prediction result. The sklearn library can split all the 5 years of data into two parts: train and test. Fig. 2 below shows the visualization of the prediction vs real stock price, and Table 9 below shows the first 10 exact values of Actual stock price and predicted price.

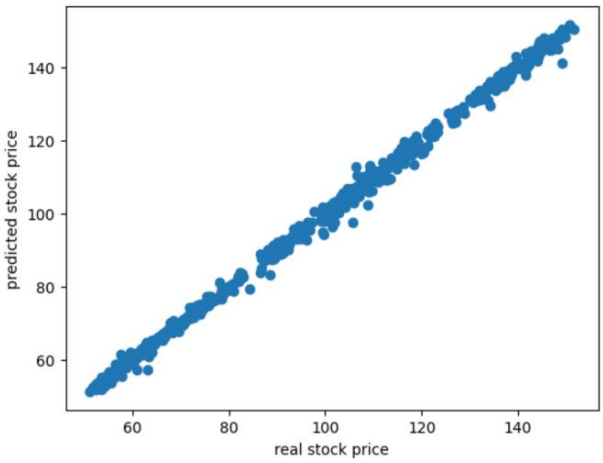

Fig. 2. Predicted stock price vs real stock price (Photo credit: Original)

Table 9. First 10 rows for the Actual and predicted stock price

	Actual	Predicted
0	57.15	57.60841
1	68.724	68.45108
2	62.463650	62.56676
3	122.75	124.2903
4	55.9755	56.12264
5	57.1955	58.41692
6	59.3445	60.10213
7	73.71295	73.621
8	66.187	65.21909
9	85.4857	87.23959

From Fig. 2 above, it can clearly see that the predicted and actual values are basically distributed on the ideal expression of y=x, and the overall trend is in a straight line. This indicates that our prediction model was very successful in this experiment. Through the prediction model, it can accurately predict stock prices, and the predicted stock prices will not differ significantly from the actual values.

4 Conclusion

In summary, this paper investigates the application of multiple linear regression in stock-price prediction by analyzing Alphabet as an example. Obtaining unknown dependent variables from the six known independent variables in the database. From section 3.4, it could come to the conclusion that this MLR model is successful:

This set of models can not only be applied to estimate tomorrow's highest price but also to estimate tomorrow's opening price. Just change the dependent variable and do the same process as shown in part II.

Besides the multiple regression model, there remain many other models that can make the prediction such as Recurrent Neural Network (RNNs), LSTMS. In future research, a new topic is the comparison of the effectiveness for different models in stock prices prediction, to help contemporary investors to make more accurate predictions and judgments.

Reference

1. Hoynes, H., Rothstein, J.: Universal basic income in the United States and advanced countries. Annual Review of Economics, 11, 929-958 (2019).
2. Mehtab, S., Sen, J., Dutta, A.: Stock price prediction using machine learning and LSTM-based deep learning models. In Machine Learning and Metaheuristics Algorithms, and Applications: Second Symposium, SoMMA 2020, Chennai, India, October 14–17, 2020, Revised Selected Papers 2 (pp. 88-106). Springer Singapore (2021).
3. Roback, P., Legler, J.: Beyond multiple linear regression: applied generalized linear models and multilevel models in R. CRC Press (2021).

4. Stojiljkovic, M.: Linear regression in Python. Real Python. https://realpython. com/linear-regression-in-python/. Accessed, 8 (2021).
5. Hao, J., Ho, T. K.: Machine learning made easy: a review of scikit-learn package in python programming language. Journal of Educational and Behavioral Statistics, 44(3), 348-361 (2019).
6. Chicco, D., Warrens, M. J., Jurman, G.: The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623 (2021).
7. Dunkler, D., Haller, M., Oberbauer, R., Heinze, G.: To test or to estimate? P-values versus effect sizes. Transplant International, 33(1), 50-55 (2020).
8. Maulud, D., Abdulazeez, A. M.: A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), 140-147 (2020).
9. Archakov, I., Hansen, P. R.: A new parametrization of correlation matrices. Econometrica, 89(4), 1699-1715 (2021).
10. Senthilnathan, S.: Usefulness of correlation analysis. Available at SSRN 3416918 (2019).