

Improvements of Stock Price Prediction Based on LSTM by Fama-French Five Factors on Market Capitalization

Chuyue Yang *

Faculty of Business Administration, University of Macau, Macau, China

*Corresponding author: bb90754@um.edu.mo

Abstract. Stock price prediction refers to determining the trend of a budgetary trade's stock projection. The precise prognosis of offer value growth will urge greater profit for speculators. The research based on previously obtained massive data with the use of AI techniques is applicable to a variety of disciplines. The primary goal is to create an analysis for the purpose of producing well-quality information that will be beneficial for decision-making. There are several machine-learning approaches with varying degrees of accuracy. However, choosing of the optimal technique to attain the maximum level of precision is an important requirement. The LSTM techniques are supervised learning-based and will require training to increase accuracy. This paper adopts LSTM algorithms combining with the Fama-French five-factors (FF5F) daily to the daily close price prediction of 10 companies represent as large and small market capitalizations from October 31, 2012 till October 28, 2022. It determines the relationship of the proposed LSTM model and market capitalization of companies. The results reveal that the LSTM model combined with FF5F for the low market capitalization companies outperforms the large market capitalization on stock price prediction.

Keywords: LSTM; Fama-French Five-Factor Model; Market Capitalization.

1. Introduction

In the stock market, a variety of instruments are used, including transactions, exchange, and so on. These instruments are supported by intelligence in order to make price predictions that will maximize stock profit [1]. To forecast stock values, several methods of machine learning and others are utilized which are thrust for several researchers. Researchers from diverse fields emphasize prediction using novel machine-learning methods which distinguish between supervised and unsupervised. Several machine learning-related technologies are readily available and have greater prediction accuracy depending on the factors that matter most [2]. Numerous pieces of data on prior stock price levels and the corresponding relative values of various stock price-influencing factors have been gathered. For determining the relation which impacts on stock prices, the factors are either employed in an unsupervised relationship or in a supervised relationship. The forecasting algorithm constructed by learning from the current data contributes to a more accurate forecast of stock prices [3].

The anticipated return is calculated using the CAPM as the amount of the profit in term of the risk-free ratio and the systematic risk [4, 5]. Furthermore, a sizable body of published research currently shows that beta does not adequately account for the return premium of small-size businesses [6] and market-value businesses with high value [7]. In response to mounting scientific evidence that the CAPM failed miserably to explain realized returns, the Fama-French Three-Factor Model (FF3FM) was created. However, it is core problem to realize the FF3FM was created and evaluated mostly on data, with most stocks based on industrially intensive companies. Hence, the resulting findings may only be applicable to certain markets [8]. Later developments by Fama and French result in the introduction of a five-factor model by adding two simulated elements to the three-factor model to reflect return premiums related to profitability and investment. They discovered that a five-factor model outperforms a three-factor model in clarifying the cross-section of returns of stocks [9].

Stock market forecasting has drawn interest from industry to academia [10, 11]. RNNs are powerful tools for analyzing sequential data including voice, longitudinal data, even textual processing [12]. Some RNN designs are utilized for price forecasting [13, 14]. As the efficient model of RNN, networks of LSTM were able to efficiently connect memories and transmit them in time

using memory cells, enabling them to grasp the data structure dynamically throughout time and increase prediction ability [15, 16].

The LSTM, first introduced by Hochreiter and Schmidhuber, is an advanced type of RNN [17]. It can learn long-term dependencies, so it is frequently used to predict sequencing data. In comparison to ordinary RNN, LSTM introduces a computational unit that substitutes neurons in the concealed layers with 3 gates inside these cells [18]. The capacity of LSTM cell to add or withdraw information from its state is controlled by these gates [19]. With these memory cells, the network can efficiently associate memories and input long-distance information in time, making it suited for dynamically grasping the data structure over time. The prediction for stock prices has been a difficult problem in the financial field for a long time. Due to the nature of the LSTM mentioned above, many scholars have tried to use the LSTM for forecast real stock prices or stock trends. An attention-based LSTM model employs a series of data, including price and indications of technological significance, to forecast the stock price trend the following day. This model is successfully applied in making stock trending strategy decisions [20]. The SVR, and random forest models were contrasted with an LSTM model proposed by Nikou [21]. Nelson et al. used stock price and technical analysis data along with LSTMs to predict future stock price patterns. In experiments, the proposed LSTM model demonstrated higher precision than other algorithms including random forests, multilayer perceptrons, and so on [22].

To enhance the accuracy of short-term stock price prediction, this paper investigates a LSTM model based on Fama-French Five Factors (FF5F) and determine the utility on various market capitalization stocks. Specifically, the proposed model has a significant utility on small market capitalization stocks. The remainder of the paper is structured as follows. Sec. 2 will interpret the data sets, variables, and model frameworks. Sec. 3 will discuss the factors correlations, forecasting results and limitations. Eventually, a conclude remarks of the whole study will be presented in Sec. 4.

2. Data & Method

2.1 Data

The data set includes daily stock prices for the top 5 largest American corporations by market capitalization (AAPL, MSFT, GOOG, AMZN, BRK-B), the 5 American companies with relatively low market capitalization that \$2 billion (ACRX, SBGI, MARK, JAN, STAR), and the FF5F daily for the time period from October 31, 2012 till October 28, 2022. The information on the stock prices for the ten firms is obtained from yahoo finance. The data for the Fama/French 5 Research Factors (2x3) daily derives from the Kenneth R. French data library. The market capitalizations of the ten firms are displayed in the Table 1. The research relies on a sliding window technique for predicting the following day stock price. The window size was set to 60 days which consists of the information overlapping the past 60 days and predictions for the next day. The train data composed of stock prices of companies and the FF5F daily for the period October 31, 2012 to October 29, 2020, validation set is from the period of October 30, 2020 to October 28, 2021 and the test set is from the period of October 29, 2021 to October 28, 2022. There are 11 input variables forming the LSTM model, including Open Price, Highest Price, Lowest Price, Close Price, Volume, Mkt-RF, SMB, HML, RMW, CMA, RF. The general framework of the model is composed of (1) an input LSTM layer containing previous 60 days' data with 11 features, listed by (2) one or two LSTM layers, (3) a hidden layer and (4) an output layer with only one cells that contains the 61st day's predicted stock price.

Table 1. Different Market Capitalization of Companies

Stock code	Market Cap.	Stock code	Market Cap.
ACRX	\$16.98 Million	APPL	\$2.103 Trillion
JAN	\$5.41 Million	AMZN	\$854.80 Billion
MARK	\$0.12 Billion	BRK-B	\$667.85 Billion
SBGI	\$1.05 Billion	GOOG	\$1.137 Trillion
STAR	\$0.66 Billion	MSFT	\$1.775 Trillion

2.2 The Fama-French Three-Factors

The FF3 model is realized by integrating two more variables to the CAPM equation to explain anticipated returns more clearly. It explains that risk premium factors affect the return on the stock market. The formular is:

$$r = R_f + \beta_3(R_m - R_f) + b_s \cdot SMB + b_v \cdot HML + \alpha \quad (1)$$

Among them, r is return ratio of portfolio, R_f is risk-free rate, R_m is the return of the market, and the estimated coefficients correspond to the change of the three factors. To be specific, the influence on the r , where the coefficient beta notion is similar to its in the CAPM model, the company size variable (SMB) indicates the return of the portfolio made up of both small and large market capitalization companies. B/M ratio premium (HML) is the discrepancy between the return of a portfolio of companies with a relatively high book value and companies with a low ratio. In an ideal situation, α , serving as excess rate of return, should be 0 since the investment portfolio's excess return will be explained by all three factors.

2.3 The Fama-French Five-Factors

A FF5 model updated the FF3 with two more factors which is robust minus weak (RMW) and conservative minus aggressively (CMA). The model's effectiveness is unchanged with how its parameters are set. RMW is the profitability beta component, and r_iRMW_t is the profitability premium. The beta factor for investment value is CMA , and the investment premium is c_iCMA_t .

$$R_{it} - R_{ft} = a_i + \beta_i(R_{mt} - R_{ft}) + s_iSMB_t + h_iHML_t + r_iRMW_t + c_iCMA_t + \varepsilon_{it} \quad (2)$$

2.4 LSTM model

A simple c, unlike standard multilayer perceptron neural networks, features a loop to allow data to retain. The chain-like topology in demonstrates that RNN is connected to sequences and lists (Fig. 1A). Traditional RNN, on the other hand, shortage of the capability to deal with long-term reliance [23]. The LSTM block has three different gates including an input gate i_t , a forget gate f_t , and an output gate o_t (Seen from Fig. 1B). Furthermore, c_t is the cell's state, X_t is the input data, and h_{t-1} is the prior state. The model of LSTM may be stated as related to information flow throughout the structure as:

$$\tilde{c}_t = \text{Tanh}(W_c \cdot [h_{t-1}, X_t] + b_c) \quad (3)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (4)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \quad (5)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \quad (6)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \quad (7)$$

$$h_t = o_t * \text{Tanh}(c_t) \quad (8)$$

where $*$ represents the Hadamard product and σ denotes the sigmoid nonlinearity function. Concatenation of the basic LSTM unit can result in more sophisticated networks that have previously been used to handle a range of time serial issues [24].

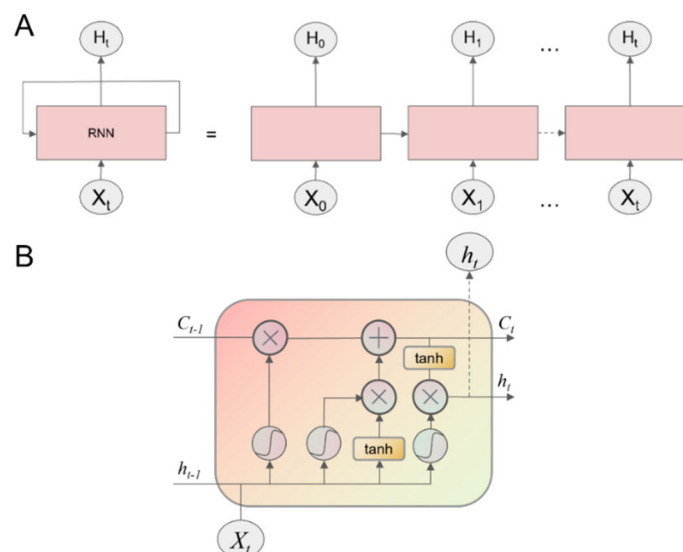


Fig. 1 A sketch of LSTM.

2.5 Evaluate Regression Models Performance

The accuracy of LSTM models integrated with FF5F was examined using *RMSE*, which is preferred in many recursive prediction and optimization strategies. Measures can be expressed as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (9)$$

While the n is the amount of data, the value of predicted and examined are donate as x_i and y_i , at the same time, bar represent as mean of variables.

3. Result & Discussion

3.1 Factor Correlation Interpretation

According to the analysis stock price mostly correlates significantly to explanatory factors, particularly the market component (Mkt). In which the stock correlate relatively weaker than the other portfolios; specifically, the stock's correlation is statistically insignificant to the factors SMB, RMW, and CMA; the relative correlation is weak. In particular, it has no statistical relevance to SMB, RMW, or CMA. Many of the shares in the data sample are traded on multiple stock exchanges. Besides, market moves in these overseas markets, as well as local ones, have a direct impact on share returns. The still-weak connection between HML and SMB is to be expected. Multicollinearity is ruled out by evidence of this modest correlation. If it turned out that the data were strongly correlated, even slight shifts in the data might cause unexpected changes in the coefficient estimates, which would reduce the regression models' capacity to forecast the future.

3.2 Forecasting Results

The findings demonstrated that normalization was quite beneficial for enhancing accuracy. This study generated better results when using equities with the small market capitalization, which implies that changing different type of stocks will alter the forecast's accuracy and that the prediction had been repeated independently 6 times for each stock. Even though the accuracy of the top 5 market capitalization companies was not particularly gratifying, the best results had been obtained by adjusting the number of LSTM layers, activation functions, the number of hidden layers, optimization equations, batch size, units in each layer and dropout rate. The results of the prediction of the small market capitalization indicated that the LSTM with FF5F is the most effective model with a lower

average RMSE about 1.576, while the top five market capitalization stocks are the least effective ones with an RMSE about 14.836 on this model. The accuracy of the models is highly correlated and varies among stocks. As summarized in Table 2, there is a substantial difference in the forecast results of the two different types of stocks, especially for small-cap stocks, where the average RMSE is about one-tenth that the large-cap stocks. It shows that the LSTM model combined with FF5F effect of large market capitalization is poor. The possible reason is that the correlation between the FF5F and large market capitalization companies' stock price is not high; while small market capitalization is effective, indicating that some of FF5F may have better effects on the stock price of small market capitalization companies.

Table 2. Average RMSE Comparison

Stock Code	Avg. RMSE	Stock Code	Avg. RMSE
ACRX	4.103415864	APPL	7.826711349
JAN	0.500132853	AMZN	8.517755936
MARK	0.207123992	BRK-B	38.62411427
SBGI	1.603718757	GOOG	7.121970761
STAR	1.466714268	MSFT	12.08965981
Total Average	1.576221147	Total Average	14.83604243

3.3 Limitation & Future outlooks

This article focuses exclusively on using the LSTM method and FF5F to forecast future stock prices for US stocks with various market capitalizations. As a result, this study is restricted to the analysis of American stocks and solely considers how the FF5F relate to stock prices across a range of market capitalizations using the LSTM approach. However, this study only considers the short-term outlook for stocks. The accuracy of various machine learning techniques, including SVM, neural networks, decision trees, etc., and the LSTM method, may be thoroughly compared with the estimated the stock price of small market capitalization companies in the follow-up study. This procedure should also be validated concurrently in the stock market of different countries.

4. Conclusion

In summary, this study forecasted the prices of diversified stocks using the Fama-French five-factors in building LSTM models. The LSTM model performed well for equities with relatively small market capitalization and resulted in low RMSE, which implies this model can interpret the correlations between the small market capitalization companies stock price and Fama-French five-factors well. It further explains that the improved trial conditions of the model will be different under different conditions. The risk-return connection is quantified by the model. From the standpoint of Machine Learning, this work can estimate stock price with controllable errors by calculating the better input parameters. As a result, Machine Learning should be viewed as a viable alternative to classic econometric methodologies. Further study should be conducted to examine the prediction error components and to employ various machine learning algorithms to select the best forecasting model. Overall, these results offer a guideline for enhancement of short-term stock price prediction through a LSTM model combined with Fama-French Five Factors by applying it on the stocks with different market capitalizations.

References

- [1] Sayavong L, Wu Z, Chalita S. Research on stock price prediction method based on convolutional neural network. 2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS). IEEE, 2019: 173-176.
- [2] Kaur R, Sharma Y K, Bhatt D P. Measuring accuracy of stock price prediction using machine learning based classifiers. IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2021, 1099(1): 012049.

- [3] Sirohi A K, Mahato P K, Attar V. Multiple kernel learning for stock price direction prediction. 2014 International Conference on Advances in Engineering & Technology Research (ICAETR-2014). IEEE, 2014: 1-4.
- [4] Lintner J. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 1965, 47: 12–37.
- [5] Sharpe W. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 1964, 19: 425–442.
- [6] Banz R. The relationship between return and market value of common stocks. *Journal of Financial Economics*, 1981, 9: 3–18.
- [7] Rosenberg B, Reid K, Lanstein R. Persuasive evidence of market inefficiency, *Journal of Portfolio Management*, 1985, 11: 9–17.
- [8] Karp A, van Vuuren G. The Capital Asset Pricing Model And Fama-French Three Factor Model In An Emerging Market Environment. *International Business & Economics Research Journal (IBER)*, 2017, 16(4): 231–256.
- [9] Fama E F, French K R. A Five-factor Asset Pricing Model. *Journal of Financial Economics*, 2015, 116: 1–218.
- [10] Krollner B, Vanstone B J, Finnie G R. Financial time series forecasting with machine learning techniques: a survey. *ESANN*. 2010.
- [11] Agrawal J G, Chourasia V S, Mittra A K. State-of-the-art in stock prediction techniques. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering* 2013, 2.4: 1360-1366.
- [12] Lipton Z C. A Critical Review of Recurrent Neural Networks for Sequence Learning. *arXiv preprint* 2015, arXiv:1506.00019.
- [13] Saad E W, Prokhorov D V, Wunsch D C. Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. *Neural Networks, IEEE Transactions*, 1998, 9(6): 1456-1470.
- [14] Rather A M, Agarwal A, Sastry V N. Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications* 2015, 42(6): 3234-3241.
- [15] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation* 1997, 9(8): 1735-1780.
- [16] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. *Journal of Computational Science* 2011, 2(1): 1-8.
- [17] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation*, 1997, 9(8): 1735-1780.
- [18] Chen K, Zhou Y, Dai F. A LSTM-based method for stock returns prediction: A case study of China stock market. 2015 IEEE International Conference on Big Data (Big Data), 2015, 1: 823-2824.
- [19] Nguyen D H D, Tran L P, Nguyen V. Predicting stock prices using dynamic LSTM models. *International Conference on Applied Informatics*. Springer, Cham, 2019: 199-212.
- [20] Cheng L C, Huang Y H, Wu M E. Applied attention-based LSTM neural networks in stock prediction. 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018: 4716-4718.
- [21] Nikou M, Mansourfar G, Bagherzadeh J. Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms, *Intelligent Systems in Accounting, Finance and Management*, 2019, 26(4): 164174.
- [22] David M, Nelson Q, Adriano C, Pereira M, de Oliveira R A. Stock Markets Price Movement Prediction With LSTM Neural Networks, in *International Joint Conference on Neural Networks (IJCNN)*, 2017:14191426.
- [23] Schumaker R P, Chen H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, 2009, 27(2): 12.
- [24] Saleh K, Hossny M, Nahavandi S. Intent prediction of vulnerable road users from motion trajectories using stacked LSTM network. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2017: 327-332.