Can A Fresh Graduate Beat the Market? A Five-Day Horizon Prediction in Taiwan's ETF 0050

1st Ti-Wen Chen
Institute of International Business
National Taiwan University
Taipei, Taiwan
r10724050@ntu.edu.tw

***Abstract*—In this study, I ventured into using an Long Short-Term Memory (LSTM) approach to forecast the upcoming five-day stock prices of Taiwan's ETF, 0050.TW. Recognizing that stock prices don't follow a stationary time series pattern, directly feeding prices into the model for predictions could disastrously fail, especially when it encounters unprecedented price levels. To circumvent this, I opted to predict the log returns and subsequently reintegrate these predicted values to construct the future five-day stock prices. Given the notorious challenge of predicting stock market movements, I concluded the study by crafting a model dedicated to forecasting the five-day stock price trend. Impressively, this model demonstrated commendable performance.**

***Index Terms*—Long Short Term Memory, Stock Price Prediction**

I. Introduction

With the advancement of technology, the dissemination of knowledge has become more convenient, and we can also learn new knowledge through the creative content shared by others on the internet. At the same time, technologies such as machine learning and deep learning have also flourished due to easier communication, with their progress advancing so rapidly that it can even be measured on a monthly basis. With the development of machine learning and deep learning, more and more fields have begun exploring how to integrate these technologies, among which the financial sector is a particularly hot topic of discussion, giving rise to the new term, "fintech." Fintech encompasses a wide range, including fraud detection, mobile payments, recommendation systems, blockchain, and quantitative trading, etc. This particular study will focus on the application of machine learning and deep learning in quantitative trading.

In the realm of traditional financial research, linear econometric models have long been the cornerstone for forecasting stock market excess returns. Notably, the seminal work by Fama and French [1] leveraged such models to unveil significant linear correlations between individual stock returns and various factors, including market excess returns, value, and size factors. Subsequently, Carhart [2] extended this framework by integrating a momentum factor, further enriching the analytical landscape. However, the rapid advancement in computational capabilities has led to the waning dominance of traditional linear models. Their erstwhile advantage of having analytical solutions for swift computations is now overshadowed by their inherent limitations, particularly their inability to capture complex, non-linear relationships that are increasingly relevant in financial markets. In contrast, the burgeoning fields of machine learning and deep learning offer sophisticated tools that excel in identifying and leveraging these non-linear dynamics. Their prowess extends far beyond the capabilities of conventional linear models, offering nuanced insights and enhanced predictive accuracy.

Beyond their ability to capture non-linear relationships, machine learning and deep learning techniques also introduce the possibility of increasing model complexity. This includes, for example, the addition of more features than there are data points, a feat traditional econometric methods cannot achieve. Kelly et al. [3] discuss how, with escalating model complexity, predictive accuracy tends to improve. Expanding on this innovative approach, Kelly et al. [4] ventured into the realm of computer vision, analyzing candlestick charts to predict market movements and successfully generating excess returns.

These advancements illustrate the evolving landscape of financial analysis, where the incorporation of more intricate and nuanced models opens new vistas of understanding. In this study, I will start by analyzing stock information, setting the stage for the application of deep learning algorithm to forecast market trends. A significant focus will be placed on employing LSTM networks, renowned for their efficacy in capturing temporal dependencies and intricate patterns in time-series data.

The structure of the remainder of this paper is as follows. Section 2 commences with an introduction to the data utilized in this study, followed by a detailed discussion on the methodology employed for analyzing the data. Section 3 introduces the LSTM model that used in this study. Section 4 presents the results obtained during the testing period. Finally, Section 5 provides the conclusion of this study.

II. Data

A. Data Sources

The dataset spans from November 2009 to March 2024, providing a substantial historical context for analysis. However, the training period is specifically set from January 2010 to March 2024, which ensures that the data begins at a clean starting point of the year, likely aligning better with complete

fiscal year records and avoiding any partial data from the end of 2009. Within this dataset, I have allocated the first 90% of the data for training purposes and reserved the remaining 10% for testing. This split allows for a significant amount of data to be used for the model to learn and discern patterns, trends, and anomalies within the financial time series.

The data sources for individual stocks and the 0050 ETF are sourced from TEJ, which is a reputable database providing a wealth of financial information and analytics. For options-related data, the information is scraped from the Taiwan Futures Exchange website[1], ensuring up-to-date and relevant market insights. Regarding interest rates, the data is downloaded from the Federal Reserve Economic Data (FRED) website[2], which is a comprehensive source for economic data, offering a wide array of financial indicators and statistics that are crucial for conducting robust financial analyses.

Since 0050.TW is an ETF, I integrated data from the top companies by allocation within this ETF as features to enhance the model. These companies include 2330.TW, 2454.TW, 2317.TW, 2881.TW, 3231.TW, 2382.TW, 2308.TW, 2303.TW, 2891.TW, 2882.TW, and 2884.TW, as well as 3034.TW. For each of these enterprises, the incorporated features are as follows: log return, RS volatility, the ratio of institutional transaction volume to its one-month average, the ratio of margin transaction volume to its one-month average, and the ratio of short selling volume to its one-month average.

Additionally, options market data were included to provide a forward-looking perspective, specifically the ratio of outstanding put and call options in the Taiwanese stock market. Macroeconomic data were also integrated to inform the model of the economy's hidden state, including data like the 10-year interest rate. These enhancements are intended to offer a more nuanced view of the factors influencing the ETF's performance, leveraging both company-specific and broader economic indicators to refine the predictive model.

B. Analyzing 0050.TW Stock Price Data

In the time series analysis conducted within this study, the initial step involves charting the trends, followed by an Autocorrelation Function (ACF) analysis. From figure 1, graph (a), it's apparent that 0050.TW exhibits highly non-stationary behavior, posing significant challenges for traditional econometric prediction methods. To transform this non-stationary time series into a more analyzable form, I compute its log return and depict the trend in graph (b). This transformation reveals a relatively more stationary series compared to graph (a) and hints at potential volatility clustering—a phenomenon I investigate further.

Certainly, analyzing the log return information is a pivotal part of the study. ACF test in graph (a) of figure 2 reveals that log returns exhibit no significant autocorrelation, suggesting that using past returns to predict future returns might not be straightforward if we rely solely on linear relationships.

[1]https://www.taifex.com.tw/cht/3/pcRatio

[2]https://fred.stlouisfed.org/

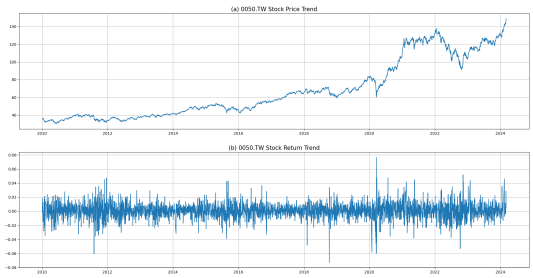

Fig. 1. 0050.TW Price Trend and Daily Return.

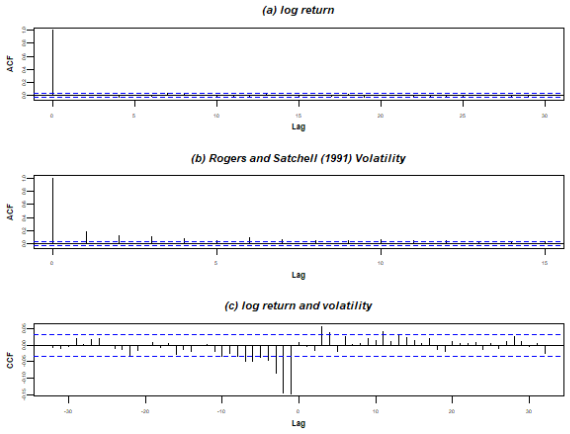

Fig. 2. ACF and CCF of 0050.TW log returns and volatility.

However, the ACF test does not capture nonlinear interactions, which could be crucial in financial data. To delve deeper, figure 3 is introduced to explore potential relationships between returns and their lagged values beyond linear correlations. Figure 3 also does not indicate any significant relationship between log returns and their lagged counterparts, it reinforces the hypothesis that predicting returns using past values might be inherently challenging.

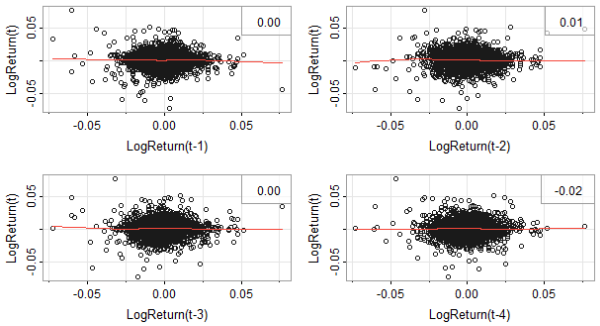

Fig. 3. Log return and lags correlation.

Drawing inspiration from the methodologies of Garman and Klass [5] and Rogers and Satchell [6], which utilize the day's high, low, open, and close prices to calculate volatility, I apply these techniques and conduct ACF and CCF analysis. Figure 4 presents the volatility trends calculated using these two

methods, while graph (b) of figure 2 highlights the significant autocorrelation in the volatility as computed by Rogers and Satchell [6]. The most important observation is that from graph (c) of Figure 2, it is evident that the Cross Correlation Function (CCF) of log return and volatility by Rogers and Satchell [6] is significant, which means log return leads the volatility. However, it also indicates that the volatility has some influence on the log return; therefore, we will consider this feature as a factor for predicting future returns.

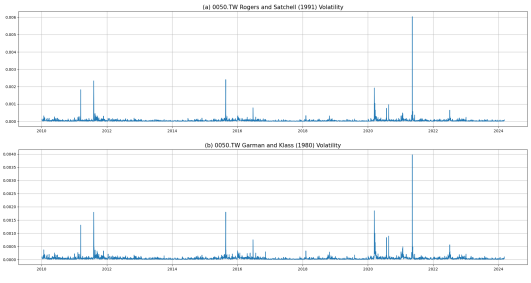

Fig. 4. 0050.TW Volatility Measurements.

III. Methodology

A. LSTM

LSTM networks, a special kind of Recurrent Neural Network (RNN), have emerged as a powerful tool for dealing with sequence prediction problems, capturing long-term dependencies in sequence data where standard RNNs might struggle. Introduced by Hochreiter and Schmidhuber [7], LSTMs were designed to overcome the vanishing gradient problem, enabling the model to retain information over extended periods and making them particularly adept at tasks such as language modeling, speech recognition, and time series forecasting.

LSTM networks distinguish themselves through their unique internal structure, featuring memory cells that store and regulate the flow of information. Each cell is governed by gates: the input gate determines how much new information flows into the cell, the forget gate decides what information is discarded, and the output gate controls the extent to which the value in the cell impacts the network output. This architecture allows LSTMs to make selective decisions about which information to store, modify, or erase, granting them remarkable flexibility and learning capacity for sequences of varying lengths and complexities.

IV. Results

A. Future Prices Prediction

First, for baseline, I posited that stock prices follow a martingale process, leading me to employ today's stock price as the expected price for the subsequent five days. This approach allowed for a nuanced assessment of the improvements offered by advanced modeling techniques over the basic expectation derived from the martingale hypothesis.

Second, I adopted a naive methodological approach, wherein I utilized the historical price data spanning the previous ten days as a predictor for forecasting the stock prices over the subsequent five-day period. Building upon this foundation, I sought to enhance the model's predictive accuracy and sensitivity to temporal patterns inherent in the stock price data. To this end, I incorporated an advanced loss function into my analytical framework, as proposed by Le Guen and Thome [8]. Specifically, I employed the DILATE loss function, a novel metric designed to overcome the limitations associated with conventional MSE loss functions in capturing the intricate temporal dynamics of time series data.

Thirdly, I plan to incorporate additional features to predict the upcoming five-day returns, which will subsequently be utilized to deduce the predicted stock prices. In this context, I will employ three distinct loss functions to optimize the predictive model. The first is the standard MSE loss function, serving as a fundamental benchmark. The second loss function employed is the Lasso loss function, which is instrumental in mitigating the influence of less significant features, thereby facilitating an effective feature selection process. This step is crucial in refining the model by isolating and leveraging only those features that provide substantial predictive value. Lastly, given the well-documented fat-tailed nature of stock returns, I have opted to utilize the Huber loss function.

The hyperparameters are as follows: batch size = 8, learning rate = 10^{-4}, epochs = 200, hidden dimensions = 200, layers = 10.

TABLE I
MSE in testing period for each models

Models	T+1	T+2	T+3	T+4	T+5
Baseline	1.586	3.191	5.128	6.770	8.306
Naive Price LSTM	19.927	22.489	25.129	27.517	30.649
DILATE Price LSTM	20.484	24.266	26.533	32.182	30.814
MSE Return LSTM	2.984	5.302	7.295	9.141	10.296
Lasso Return LSTM	2.209	4.819	6.753	8.167	10.10
Huber Return LSTM	2.793	5.423	7.082	8.961	11.065

From Table 1, it is patently clear that using past stock prices to predict future values yields poor performance. I surmise that this inadequacy stems from the non-stationary nature of stock prices combined with a limited dataset, leading to significant discrepancies when the model encounters stock prices during the testing period that were not present in the training dataset. This hypothesis is visually supported by the observations in Figure 5. In Figure 5, I allocated 80% of the dataset to training data to streamline the process, with the remaining portion designated for testing. The resulting analysis, as illustrated in Figure 5, reveals an interesting pattern: the model predicts relatively accurately when the actual stock prices decline to levels previously observed within the training data. However, when stock prices begin to surge significantly, the model incurs substantial errors.

Furthermore, utilizing additional features to predict the subsequent five-day returns and then extrapolating these predictions back to stock prices has noticeably reduced the error margin. Among the various methods, the Lasso loss function exhibited the most favorable performance. This suggests that

Fig. 5. LSTM price prediction.

within the pool of features, there are elements that minimally impact the prediction of returns, aligning with the initial hypothesis that advocated for the significance of feature selection. However, it is important to note that even the best-performing model, which employs the Lasso loss function, does not outperform the baseline established under the martingale assumption. This observation underscores the inherent challenge of predicting future stock prices, highlighting the complexities and unpredictabilities embedded within financial market dynamics.

B. Future Stock Trend Prediction

Given the challenges associated with predicting precise stock prices, the focus can indeed shift toward predicting the market trend, which might offer a more robust and actionable insight. In this context, I continue to utilize the aforementioned features to forecast the trend five days into the future. The model is designed to categorize its predictions into three distinct outcomes: the future stock price will be higher, unchanged, or lower compared to the current price. The reported metrics values for the four trained models in terms of their classification capabilities are as follows: **f1 score** = 0.548, **recall** = 0.550, **precision** = 0.546, and **accuracy** = 0.550. These values indicate that the model performs better than random guessing, which would have an accuracy of around 50%. This demonstrates that the model has some utility in classifying the trend of stock prices, albeit modestly.

However, it's crucial to emphasize that a slightly above-chance performance does not necessarily translate into profitability. One significant drawback of this model is that it only considers the direction of the price movement for training and does not account for the magnitude of changes. If the correct predictions yield only minor gains and the incorrect ones result in substantial losses, such a strategy is inherently flawed and unsustainable. The model's utility, therefore, is more academic or exploratory, providing insights into the possibility of trend prediction but without offering a reliable mechanism for financial gain. The ability to predict the direction accurately without considering the potential size of the price movement or the associated risks can lead to a misleading evaluation of the model's practical effectiveness in a trading context.

CONCLUSION

In this study, I developed an LSTM model to predict the future five-day stock prices of 0050.TW. The findings suggest that, due to the non-stationary nature of stock prices, directly forecasting the prices results in significant inaccuracies. However, predicting the log returns and then converting these predictions back into stock prices proved to be more feasible. Despite the relative ease of predicting stationary properties like log returns, the results still indicate that my model was unable to outperform the baseline established under the stock price martingale assumption, aligning with the efficient market hypothesis.

As for future work, there is potential for incorporating a broader dataset and additional factors. Furthermore, experimenting with more sophisticated models, such as transformers, could offer improvements in predictive accuracy. These future endeavors could provide deeper insights into stock price movements and possibly refine the forecasting capabilities beyond the current model's performance.

REFERENCES

[1] Fama, E. F., French, K. R. (1992). The cross-section of expected stock returns. the Journal of Finance, 47(2), 427-465.
[2] Carhart, M. M. (1997). On persistence in mutual fund performance. The Journal of finance, 52(1), 57-82.
[3] Kelly, B., Malamud, S., Zhou, K. (2024). The virtue of complexity in return prediction. The Journal of Finance, 79(1), 459-503.
[4] J. Jiang, B. Kelly and D. Xiu. 2023. (Re-) Imag (in) ing price trends. The Journal of Finance, 78(6), 3193-3249.
[5] Garman, M. B., Klass, M. J. (1980). On the estimation of security price volatilities from historical data. Journal of business, 67-78.
[6] Rogers, L. C. G., Satchell, S. E. (1991). Estimating variance from high, low and closing prices. The Annals of Applied Probability, 504-512.
[7] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[8] Le Guen, V., Thome, N. (2022). Deep time series forecasting with shape and temporal criteria. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1), 342-355.