Research Article

Long Short-Term Memory Recurrent Neural Network for Predicting the Return of Rate Underframe the Fama-French 5 Factor

Bui Thanh Khoa and Tran Trong Huynh

1Finance, Economics, and Management Research Group, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
2FPT University, Hanoi, Vietnam

Correspondence should be addressed to Bui Thanh Khoa; khoa.bt@ou.edu.vn

Received 22 April 2022; Accepted 19 May 2022; Published 30 May 2022

Abstract

The multifactor approach helps determine the linear connection between a diversified portfolio’s return and risk; however, the efficacy of the model models is still limited in the experiment. Algorithms in machine learning have recently grown in popularity to compensate for some of the shortcomings of theoretical models. This study applied a machine learning technique to compare the performance of the Fama-French 5-factor model (FF5). Two approaches are employed in the Fama-French model: Long Short Term Memory Recurrent Neural Network (LSTM-RNN) and Maximum Likelihood Estimation (MLE). From January 1, 2010, through March 3, 2022, the stock market in Ho Chi Minh City was experimentally researched. The rolling window approach is used in combination with the Root Mean Square Error (RMSE), and the results of the FF5 model with the LSTM-RNN algorithm are more efficient in prediction error than the MLE methodology. This contribution encourages investors and hedge fund managers to use the LSTM-RNN algorithm to boost forecasting efficiency.

1. Introduction

Sharpe [1] introduced the proposed capital asset pricing model (CAPM) based on Markowitz’s portfolio diversification theory. The CAPM model measures the linear relationship between hazardous assets’ return and risk. This concept swiftly became one of the theoretical pillars of modern finance. Because of its simplicity, it is employed by academia, investors, and investment management institutions. CAPM provides a minimum rate of return for risky investment projects for investors to reference. CAPM measures the systematic risk of marketed financial instruments such as stocks and bonds via beta coefficients. However, because CAPM makes too many assumptions that are difficult to meet in practice, its reality has remained a source of contention.

Years later, Banz [2] found the size effect in the US market. Small enterprises, in particular, appear to have larger returns than large firms. This finding shows that the CAPM’s explanation for the scale effect is faulty. Basu [3]’s subsequent work produced similar results to Banz’s. Fama and French [4] established the value-growth impact of equities in 1992. Value equities (those with high B/M ratios) outperform growth companies (those with low B/M ratios). Fama and French proposed a three-factor model by adding two new components while preserving the market factor (later called the 3-factor Fama-French model). Because the 3-factor model explains better than the CAPM patterns previously utilized, Fama and French conducted research using thousands of random stock certificates to test their pattern and discovered that when the ratings and values are combined with the amount of beta, the model can explain 89 percent profit in a varied stock category. With the ability to explain 89 percent of profits compared to the overall market, the investor can construct a portfolio in which they obtain a high-yield rate of relativity in which they have built-in their investment.
Continuing to develop the 3-factor model, Fama and French [5] expanded the three-factor model by including two elements linked to the company’s investment and profit. From 7/1963 through 12/2013, Fama-French tested the 5-factor model in the US market. The model explained roughly 71% to 94% of the volatility in the return series of diverse portfolios. Compared to the 3-factor model, the 5-factor model is more effective in explaining return volatility.

The use of machine learning algorithms to exploit complex correlations between variables is a recent trend. The initial wave of publications used neural networks to forecast derivatives prices [6, 7]. Heaton et al. [8] created a deep learning approach to portfolio selection using neural networks. Shrinkage and selection algorithms were developed to estimate expected returns based on nonlinear connections between variables [9, 10]. Gu et al. [11] offered numerous machine learning algorithms for forecasting market returns, including dimension reduction, boosted regression trees, random forests, and neural networks.

The rapid advancement of the information technology industry, particularly the processing speed of computers, has greatly aided deep learning algorithms. As a result, deep learning algorithms are commonly used to tackle experimental challenges. A recurrent neural network (RNN) predicts future events using time-series data. However, some issues linked to the vanishing gradient problem persist, hurting the prediction model’s effectiveness. LSTM-RNN was created to address this issue to address various difficulties that traditional RNN could not [12–15].

Roondiwala et al. [12] examined the accuracy of stock price projections under the LSTM-RNN when the NIFTY50 share price of the National Stock Exchange of India stock was paired with Open data for the study of stock prices. Consequently, the best results are obtained by combining the four input variables. Furthermore, Zhuge et al. [13] predicted the opening share prices of individual equities. It is concluded that the acquired results appear to be superior to the standard RNN application. When used for time series data processing, it is well known that LSTM-RNN has high efficiency. However, depending on the model-building approach, it might be a means of an effective predictive model. In other words, a good model contains both the underlying theory and an algorithm that fully exploits the latent correlations between variables. LSTM-RNN has also been applied successfully in demand forecasting and financial market forecasting [14–16]. Siami-Namini et al. [17] show that the LSTM-RNN model outperforms the ARIMA model in time series forecasting. LSTM-RNN was utilized in the Forex market rate prediction model by Yildirim et al. [18].

In this work, this study proposed to combine the theoretical framework FFS and the LSTM-RNN algorithm in the model for predicting the series of returns of investment portfolios. The main contribution of the study consists of two parts:

(i) Application of the LSTM-RNN algorithm in the stock return forecasting model.

(ii) Build a pattern that includes financial theory and AI algorithms.

2. Theoretical Foundation and Empirical Evidence

2.1. Fama-French Five-Factor Model. Fama and French [19] proposed a three-factor model frequently employed in academic and experimental research. The CAPM model explains less well than the 3-factor approach (the CAPM model lacks explanations regarding the size premium and the value premium). Some data suggest that the three-factor Fama-French model is insufficient. Novy-Marx [20], for example, shows that profitability is closely connected to average returns. In addition to this issue, Titman et al. [21] and Anderson and Garcia-Feijoo [22] discovered that investment growth is inversely connected with returns. Fama and French [5] presented a 5-factor model that includes both profit and investment factors to address these issues.

The three-factor model augments CAPM with additional factors to capture the size and value premiums. The time series regression equation has the form:

\[
r_{it} - r_{ft} = \alpha_{it} + \beta_1 Mkt_{it} + \beta_2 SMB_{it} + \beta_3 HML_{it} + \epsilon_{it}. \tag{1}
\]

Means:

(i) \(r_{it}\) = Return on asset \(i\) at time \(t\).

(ii) \(r_{ft}\) = The risk-free rate at time \(t\).

(iii) Mkt \(_{it}\) = The excess return of the market portfolio at time \(t\).

(iv) SMB \(_{it}\) = Scale offset (small-scale minus large scale).

(v) HML \(_{it}\) = Value premium (value stocks minus growth stocks).

(vi) \(\beta_{1,2,3}\) = Regression coefficients.

(vii) \(\epsilon_{it}\) = Random error.

The five-factor model adds profitability and investment factors to the three-factor model. The regression equation has the form:

\[
r_{it} - r_{ft} = \alpha_{it} + \beta_1 Mkt_{it} + \beta_2 SMB_{it} + \beta_3 HML_{it} + \beta_4 RMW_{it} + \beta_5 CMA_{it} + \epsilon_{it}. \tag{2}
\]

Means:

(i) \(r_{it}\) = Return on asset \(i\) at time \(t\).

(ii) \(r_{ft}\) = The risk-free rate at time \(t\).

(iii) Mkt \(_{it}\) = The excess return of the market portfolio at time \(t\).

(iv) SMB \(_{it}\) = Scale offset (small-scale minus large scale).

(v) HML \(_{it}\) = Value premium (value stocks minus growth stocks).

(vi) RMW \(_{it}\) = Return factor (high-return portfolio minus low-return portfolio).

(vii) CMA \(_{it}\) = Investment-related factor (small portfolio minus high portfolio).

(viii) \(\beta_{1,2,3,4,5}\) = Regression coefficients.

(ix) \(\epsilon_{it}\) = Random error.
Some tests are based on the standard five-factor Fama-French model. Cakici [23] investigated the stock market in 23 developed nations between 7/1992 and 12/2014. The study’s findings are as follows: the five-factor model is more effective than the 3-factor model in North America, Europe, and International markets, as most of the initial components are present. In other cases, the HML factor is not statistically significant. The two newly added variables of the five-factor model have no statistical significance or have a very low level in the Japanese and Asia Pacific markets. Gruodis [24] investigated the five-factor model in Europe and International markets, as most of the initial components are present. In other cases, the HML factor is not statistically significant. The two newly added variables of the five-factor model are present. In other cases, returns or investment premiums are not statistically significant.

2.2. LSTM-RNN Algorithm. A recurrent neural network (RNN) is a neural network. The output $o_t$ at each node of the RNN depends not only on the input $x_t$ at that node but also on the output $o_{t-1}$ of the previous node in the network. The function can represent in (3)

$$o_t = f(W_{\text{input}}x_t + W_{\text{output}}o_{t-1} + b),$$

where $f$ is the cell’s activation function, $x_t, o_t$ are the input and output of the RNN at time $t$, $W_{\text{input}} W_{\text{output}}$ is the matrix of parameters to be searched for in the model, and $b$ is the bias vector of the model. One of the disadvantages of the RNN model is that it does not solve problems related to long-term memory well. The long short-term memory (LSTM) model introduced by Hochreiter and Schmidhuber [27] is an enhanced/advanced version of RNN that overcomes the inherent weakness of the RNN model.

In a typical architecture of an LSTM, the input of each cell at time $t$ and the input value $x_t$ have the state $C_{t-1}$ and the output value $o_{t-1}$ of the previous step. The cell’s output and the output value $o_t$ also have long-term information carried in the cell state $C_t$. This improves the RNN model and helps LSTM learn more effectively when learning depends on long-term memory.

Mathematically, the model uses the following functions:

$$f_t = \sigma(W_f \cdot [o_{t-1}, x_t] + b_f),$$

$$i_t = \sigma(W_i \cdot [o_{t-1}, x_t] + b_i),$$

$$C_t = \tanh(W_C \cdot [o_{t-1}, x_t] + b_C).$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t,$$

$$o_t = \sigma(W_o \cdot [o_{t-1}, x_t] + b_o).$$

LSTM uses a forget gate to decide which input information should be kept or ignored through the logistic function sigmoid $\sigma$ as shown in (4). The number of information $f_t$ calculated by this function will be used to calculate the output $o_t$ in (7) and (8). In addition, the functions in (5) and (6) say that new information should be combined with the retained data to create the new state and update it to the cell state $C_t$.

Recently, deep learning has been gaining more attention in financial forecasting tasks. Ding and Qin [28] implemented a convolution neural network (CNN) to process events collected from news websites to predict the S&P index. Chen et al. [29] deployed a recurrent neural network to analyze news content posted on social media. Ko and Chang [30] used LSTM-RNN to forecast stock prices; input variables include opening price, closing price, highest price, lowest price, volume, news, and forum.

3. Methodology

Research data include all companies listed on the Ho Chi Minh City Stock Exchange (HoSE) from January 2012 to January 2022. We will exclude companies with a listing period of less than one year and nonstock codes (such as fund certificates and bonds). The collected information includes adjusted closing prices of stocks, VN-Index, and a 1-year bond yield. Price data and VN-Index are collected from the stock exchange, and bond yields are collected from the website of the Ministry of Finance, which can be accessed from the URL: https://vst.mof.gov.vn/webcenter/portal/btc/r/m/trangchu?

We arrange the stocks alphabetically and divide them into ten diversified portfolios. By the end of 2021, on HoSE, there are 404 stocks, which will be classified into ten lists, each of 40 stocks; the 10th list alone will have 44 stocks. The return rate will be calculated with equal weight on stocks. The factors are built and calculated as described in Fama and French [5]. Details are shown in Table 1.

Consider the general linear regression model of the form as

$$y = x^T \beta + \epsilon.$$

In which,

(i) $x$ is the matrix of inputs

(ii) $y$ is the output matrix

(iii) $\beta$ is the matrix of regression coefficients

(iv) $\epsilon \sim N(0, \sigma^2)$ is the random error with the unknown parameter.

Suppose a training set is obtained from a random sample of $k$ inputs $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. The likelihood function is determined by
The MLE method is to find \( \beta_{MLE} \) to maximize the likelihood function. This can be done using gradient ascent or gradient descent for the negative likelihood function. However, we often use log-transformation to minimize the log-likelihood function for the likelihood function.

Using log-likelihood for the normality assumption, we have

\[
\log p(y|x, \beta, \sigma^2) = \log \left( \prod_{i=1}^{n} p(y_i|x_i, \beta, \sigma^2) \right) = \log \left( \prod_{i=1}^{n} N(y_i|x_i^T \beta, \sigma^2) \right).
\]

(10)

The processing process is carried out according to the following steps as shown in Figure 1: (1) collecting and cleaning data, (2) calculating variables and factors, (3) estimating parameters, and (4) computing errors. The study used two estimation methods. The quantities are MLE and LSTM-RNN. We use past data of 5 consecutive years (60 months) to estimate the parameters, as shown in Figure 2. For the LSTM-RNN algorithm, we use batch_size = 20, deep network = 6, and layers = 6.

We chose the root mean square error (RMSE) evaluation criteria to evaluate the error like in previous studies [17, 31]. The root-mean-square error (RMSE) is a measure frequently used for assessing the accuracy of prediction obtained by a model. It measures the differences or residuals between actual and predicted values. The metric compares prediction errors of different models for a particular data and not between data sets. The formula for computing RMSE is as follows:

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}}.
\]

(14)

4. Results

4.1. Descriptive Statistics. The factors are built from diversified portfolios, according to Fama-French (2015). Specifically, in June each year, stocks will be ranked by market capitalization, B/M (Book to Market Ratio), profitability, and investment. Combining sorting by size and B/M ratio creates six similar for-profit and investment portfolios, yielding 18. Then calculate the factors HML, SMB, RMW, and CMA. The Mkt factor will be represented by the market index (VN-index), specifically the difference between the VN-index return and the 1-year government bond yield. Stocks traded on HoSE will be grouped into ten categories in alphabetical order. The descriptive statistics are summarized in Table 2.

There are 147 observations, each with 1 month, from 1/2010 to 3/2022. The rate of return in various volatile portfolios ranges from 0.176%/month to 0.694%/month. The portfolios \( p8 \) and \( p9 \) concentrate most of the codes related to technology and real estate, so they have relatively higher outstanding returns but are characterized by high supply risk and standard deviations are 10,034 and 10,151. In the \( p3 \)
category, the standard deviation is the highest compared to the other portfolio, has a value of 11,919, and has the widest range of values, from −55.68 to 28,591. In this portfolio, most stocks related to imports and export are concentrated.

During the COVID-19 pandemic, most import-export companies faced cross-border production and trade difficulties. Opposite, stocks in the p2 portfolio were quite stable because most were related to banks and financial institutions. As a result, returns and standard deviations are low.

Table 3 shows that the variables have a low correlation (absolute value less than 0.7), in which all variables are statistically significant with Mkt. The market factor is positively correlated with the value factor and negatively correlated with the HML, RMW, and CMA factors. The negative relationship between Mkt and HML shows that investors expect more in growth stocks when the market is upbeat (bullish). Conversely, when the market is down, investors prefer value stocks. The positive relationship between Mkt and SMB shows that, when the market is
growing, investors prefer small-cap stocks, which leads to a corresponding increase in the value of SMB.

4.3. Forecast Results. This study uses the rolling windows method with a time series of 60 months to evaluate the effectiveness of the two forecasting models. As a result, the RMSE calculation is summarized in Table 4.

The results of Table 4 show that the RMSE error in the model using the LSTM-RNN algorithm is superior to that of the regression model using the maximum likelihood method with an average RMSE error of 1,952, 2,591, respectively. The MLE method is most effective in category p7 with an RMSE error of 2,244 and least effective in category p1, RMSE error ranges from 2,244 to 3,024. The prediction model using LSTM-RNN also gives similar results as MLE, most effective in the p7 portfolio and the worst in the p5 and p1 portfolio; RMSE ranges from 1,738 to 2,163.

To show that LSTM-RNN is more effective than MLE, we perform a T-Test for the series of distances between the predicted value and the actual value of the two models with the following hypothesis and hypothesis:

\[ H_0: \text{there is no difference between the two algorithms} \]
\[ H_1: \text{the LSTM-RNN algorithm is more efficient} \]

The test results obtained \( t \)-stat = −7.63, and the corresponding \( p \)-value is less than 0.0001, so reject \( H_0 \). Thus, the LSTM-RNN algorithm is more efficient than the RNN algorithm.

5. Conclusion

The MLE algorithm is considered a more general parameter estimation method than the ordinary least squares (OLS)
method in the case of normally distributed random errors. However, the assumption of a normal distribution is sometimes unrealistic; moreover, the relationship between the variables is not simply linear. Some machine learning algorithms are superior to classical econometric algorithms to exploit latent relationships between variables. For data related to time series, the LSTM-RNN algorithm is considered one of the very effective algorithms in future forecasting.

A signal prediction model requires a combination of two factors: the supporting background theory and an efficient parameter estimation algorithm. For portfolio return forecasting, the FF5 model is one of the most effective explanatory models [23–25, 32]. However, these studies compare with a few other models, such as the CAPM model, 3-factor model, and 4-factor model. Moreover, these studies only stop at the interpretation of statistical significance and \( R^2 \) value without considering the perspective of machine learning, that is, evaluating the prediction error.

The \( T \)-Test results have shown that the MLE or OLS method is not the best method of estimating beta coefficients in the FF5 model in the specific case of the HoSE market. The LSTM-RNN method is more efficient, with the average RMSE error of 10 categories only 1,952. This result is more consistent with some previous studies, such as Zhuge et al. [13], Borovkova and Tsiamas [15], Minami [16], Ko and Chang [30].

Compared with some previous studies, we have overcome some limitations in them. More specifically, the forecasting model that we use is completely based on the theoretical foundation of finance, which has been proven to be effective experimentally. Furthermore, the algorithm we use has proven effective for time series forecasting. The new point of this study is to propose a method to estimate the parameters in the FF5 model to produce a more effective predictive model. Furthermore, our study overcomes some limitations from previous studies; for example, in the study of Ko and Chang [30], the authors exploit the LSTM-RNN algorithm and rely on past information to predict the future. Unfortunately, in the future, the markets are effective [33], all past information fully reflects the stock price. So, it is hard to predict; in other words, a price model is a random \( b \).

The main objective of this study is to apply the LSTM-RNN algorithm in the 5-factor Fama-French model experimentally in the HoSE market. We compare the model using the LSTM-RNN algorithm and the model using the MLE method. The MLE method is considered a more general method than the OLS method. As a result, the model uses LSTM-RNN more efficiently than MLE. From that, we propose to use the estimation method using the LSTM-RNN algorithm in the 5-factor model to increase the accuracy of the forecast. We emphasize that an effective predictive model must combine the underlying theory and a suitable estimation method. Hence, this study proposes some recommendations based on the result.

With the theoretical contributions, first, the five-factor Fama-French model is a good predictive framework for changing the expected returns of diversified portfolios. The model quantifies the linear relationship between risk and expected return. From a Machine Learning perspective, when estimating the optimal input parameters, we can forecast the returns of the portfolios with controlled errors. Therefore, Machine Learning should be considered an alternative to traditional econometric methods. Second, for time series where the characteristic parameters change over time, the rolling window method should be considered instead of other methods such as k-fold cross-validation to increase the model’s reliability. Shape and limit the phenomenon of overfitting. Moreover, the LSTM-RNN algorithm is one of the candidates for estimating the parameters in the predictive model. Deep learning algorithms can “learn” data in-depth, thus having better predictive capabilities than conventional algorithms like MLE or algorithms in economics basis quantity such as OLS regression. Therefore, researchers should consider them in actual forecasting.

In managerial implications for investors and fund managers, the 5-factor model is considered one of the best models to estimate the expected return of the investment portfolio. We can increase accuracy by using algorithms in deep learning, such as LSTM-RNN, to exploit latent relationships between variables.

The scope of research is still narrow, only considering the HoSE market. Furthermore, we have not considered the uncertain events affecting the market, such as the COVID-19 pandemic or crises, special fiscal policies, etc., market distortions that the Fama-French model is difficult to explain. We propose that the next research direction is to combine behavioral finance, multifactor models, and some algorithms in deep learning to build a more effective predictive model.

**Data Availability**

The data are available on request.

**Conflicts of Interest**

The authors declare that they have no conflicts of interest.

**References**


