Proceedings of ICEMGD 2024 Workshop: Decoupling Corporate Finance Implications of Firm Climate Action
DOI: 10.54254/2754-1169/110/2024ED0099

Harnessing the Power of Machine Learning to Superchange CAPM Forecasts

Shuangzi Zhang[1,a,*]

[1]Economics, Boston College, 140 Commonwealth Ave, Chestnut Hill, MA, United States
a. zhangarg@bc.edu
**corresponding author*

Abstract: This paper investigates the improvement of the Capital Asset Pricing Model (CAPM) by incorporating machine learning techniques. The objective is to address the model's conventional constraints in accurately predicting stock returns. The conventional Capital Asset Pricing Model (CAPM), which heavily depends on the beta coefficient to explain returns, frequently proves inadequate in emerging markets and neglects to consider market irregularities such as size and value effects. Our methodology enhances the estimation of beta by refining the Capital Asset Pricing Model (CAPM) to exclude the intercept term and incorporating a dynamic rolling regression method. This approach captures the fluctuations of beta more effectively across different periods, resulting in a more precise estimation. This study utilises ten years of weekly closing price data from companies listed on the NSE Nifty 50 index. It employs rolling regression and two-stage regression analysis to improve the prediction of excess returns above the risk-free rate. The revised CAPM model, which omits the intercept, exhibits a superior level of accuracy, as evidenced by higher adjusted R-squared values and improved F-statistics. In addition, the paper analyses the Fama-French models within the Chinese stock market and explores the potential of advanced machine learning models, such as Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN), to improve forecasting methods in unpredictable markets. Integrating LSTM-RNN models presents a promising approach to capturing intricate patterns in financial time series data, demonstrating substantial enhancement in forecasting accuracy compared to conventional methods such as OLS.

Keywords: Capital Asset Pricing Model (CAPM), Machine Learning, Rolling Regression, Asset Returns, Financial Forecasting

1. Introduction

Traditionally, the Capital Asset Pricing Model (CAPM) relies exclusively on beta as a means of explaining returns, disregarding other potentially influential factors. [1] The efficacy of the model fluctuates, frequently exhibiting inferior outcomes in markets that are less developed. In addition, the Capital Asset Pricing Model (CAPM) fails to account for market anomalies such as the size and value effects that are accounted for by other models. CAPM assumes a simplistic linear relationship between expected returns and beta, which may not always be accurate. Moreover, extensive research has disproven the sufficiency of beta as the sole factor accounting for returns.

Nevertheless, a refined approach that integrates machine learning and CAPM can adequately address the initial limitations. The revised model improves upon the traditional CAPM by removing the intercept term in order to enhance the effectiveness of beta. The method utilises a dynamic rolling regression technique to effectively capture the variations in beta values across various time periods. Typically, the improved model calculates beta values for portfolios by excluding the intercept in the regression formula and considering only market risks. Subsequently, it evaluates the effectiveness of the betas in forecasting returns that surpass the risk-free rate, while maintaining the exclusion of the intercept to enhance clarity and accuracy. This enhanced model exhibited superior precision and suitability in comparison to the conventional CAPM. The simplification of the model permits the potential incorporation of supplementary variables that may offer a more comprehensive explanation of returns compared to beta alone.

A recent analysis examined a decade's worth of weekly closing price data (from April 2011 to March 2021) for companies listed on the NSE Nifty 50 index in Todaya Market. From a pool of 50 stocks, a total of 48 were selected based on the presence of comprehensive data for the entire ten-year period being analysed. [2] The Nifty 500 index served as a representative of the market, while the 91-Days Treasury Bill rate was chosen as the rate that carries no risk. The beta of the stocks was calculated using the data from the first two years. Subsequently, yearly data was used to create and analyse five portfolios. This paper incorporates machine learning into the original CAPM model by employing regression models such as rolling regression and two-stage regression analysis. The study utilised rolling regression as a key methodology, which involved dividing the dataset into overlapping segments of 3 years. These segments were then analysed sequentially, with each analysis advancing quarterly.

Another scholarly paper conducts a comprehensive study to assess the efficacy of the Fama-French asset pricing models in the specific context of the Chinese stock market. [3] This market is known for its distinct regulatory and economic conditions. The main objective is to compare the three-factor (FF3) and five-factor (FF5) models with the traditional Capital Asset Pricing Model (CAPM), using stock data from the Shanghai A-share market between 1994 and 2023.

The paper utilises data from January 1994 to December 2023, specifically excluding segments of the market such as the Growth Enterprise Market (GEM) and Key Economic Market (KEM). The study employs a 2x3 portfolio segmentation technique that considers factors such as market capitalization-to-book ratio, market capitalization-to-operating profitability, and market capitalization-to-investment level. This comprehensive segmentation facilitates accurate data analysis and verification.

In addition, the introduction of machine learning methods, specifically deep learning models like Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN), has significantly transformed forecasting approaches. [4] These models offer improved predictive precision and better management of the intricate nature of financial time series data.

In the past, financial forecasting has been based on models that assume linear connections between returns and their predictors, such as the CAPM and later, the multi-factor models proposed by Fama and French. These models have played a crucial role in both academic and practical financial applications, but they often struggle to handle the complexities of real-world data. The adoption of more sophisticated models has been facilitated by recent advancements in computational power and machine learning. Research conducted by Roondiwala et al. (2017) has shown that LSTM-RNN models are highly effective in forecasting stock prices. [5] This indicates that these models have great potential for use in various financial applications.

Another researcher performs experiments to evaluate the effectiveness of a machine learning model. They utilise a dataset that consists of various variables from the Ho Chi Minh City Stock Exchange, spanning from January 2015 to June 2023. These variables include the VN-Index, yield on one-year bonds, and adjusted closing prices of the stock market. [6] The selection of this time period was based on its high level of volatility, which makes it an ideal test scenario for forecasting models. The analysis utilised two primary computational models: Ordinary Least Squares (OLS) and LSTM-RNN.

2. Methodology

2.1. Two-stage regression analysist

The Two-stage regression analysis involves conducting a time series regression to determine the beta (β) value, which quantifies the level of volatility or systematic risk in relation to the market. This serves as the initial stage of the analysis. In the second stage, cross-sectional regression is conducted to examine the correlation between the calculated beta and the excess returns of the portfolios, relative to the risk-free rate.

$$\text{Return of Security} = \alpha + \beta \times \text{Market Return} + \epsilon \tag{1}$$

$$\text{Excess Return} = \beta \times (\text{Market Return} - \text{Risk} - \text{Free Rate}) + \epsilon \tag{2}$$

The intercept in the regression model is denoted by alpha (α), while the beta (β) represents the sensitivity of the stock's returns to market returns. The error term is indicated by the symbol ϵ.

The constrained CAPM Model developed by Bajpai and Sharma (2015) omits the intercept term, which has been shown to provide a more accurate explanation of the fluctuations in stock returns compared to the conventional model. [1] This model enhances the conventional CAPM by removing the intercept term and proposing that the asset's returns are completely and exclusively accounted for by the market's excess returns above the risk-free rate, which are directly proportional to the asset's beta. The t-tests were used to assess the significance levels of the beta coefficients. The results indicated that the constrained model, which did not include the intercept, had a better fit.

By excluding the intercept, this model has the potential to offer a more precise estimation of the correlation between returns and beta, specifically in illustrating how fluctuations in market returns can directly impact asset returns. Furthermore, the constrained model typically demonstrates higher adjusted R-squared values, indicating that it provides a better explanation for the variation in stock returns compared to the traditional model with the intercept. [7] The results showed that this model had a superior fit in terms of F-statistics and the significance levels of beta across different sub-periods analysed using rolling regression techniques.

2.2. CAPM, FF3 and FF5 models

Regression models are constructed to examine the impact of market factors on stock returns. The predictive capabilities of each model, namely CAPM, FF3, and FF5, are evaluated using the same conditions. Ordinary Least Squares (OLS) regression is used to calculate model parameters, which gives a numerical measure of the influence of each factor on stock returns. Visual representations of regression coefficients and confidence intervals can effectively summarise the relationships between variables, providing clear insights into the significance of factors.

The study commences by employing descriptive statistics to offer a comprehensive overview of the characteristics of the data. For each factor, essential measurements such as the mean, standard deviation, and range are computed. An illustrative graph could depict the distribution of returns for each factor during the study period, emphasising periods of high volatility or stability. The efficacy of the models is further examined by conducting thorough regression analysis, which quantifies the impact of each factor on stock returns. Graphs depicting the time-series analysis of factor premiums can aid in visualising patterns and changes in the effects of factors over time.

In order to assess the resilience of the models, the study incorporates sub-period analyses that exclude significant market disruptions such as the 2008 financial crisis. [2] This approach guarantees that the models' performance remains consistent under varying market conditions. A line graph can be used to illustrate the consistency of model predictions over different sub-periods, emphasising any irregularities or consistent trends.

The results are structured to facilitate a direct comparison of the performance of the FF3 and FF5 models with the CAPM. This section would be enhanced by the inclusion of bar graphs that compare the R-squared values of each model. These graphs would visually illustrate the relative explanatory power of the models in relation to each other. The importance of the findings is examined by establishing a connection between the models' performance and the research questions and existing literature. An enhancement to the discussion could involve incorporating a scatter plot that demonstrates the correlation between model predictions and actual market returns, highlighting the areas where the models demonstrate success or failure.

The formula of FF3 model is:

$$E(R\_i) = R\_f + \beta\_mkt \times (R\_m - R\_f) + \beta\_SMB \times SMB + \beta\_HML \times HML \qquad (3)$$

SMB, which stands for Small Minus Big, represents the size premium that measures the additional returns generated by small-cap stocks compared to large-cap stocks. HML, short for High Minus Low, refers to the value premium, which quantifies the additional returns generated by stocks with high book-to-market ratios compared to those with low book-to-market ratios.

The FF5 model further includes two additional factors to the three-factor model:

$$\begin{aligned} E(R\_i) = R\_f + \beta\_mkt \times (R\_m - R\_f) + \beta\_SMB \times SMB + \beta\_HML \times HML + \beta\_RMW \times RMW \\ + \beta\_CMA \times CMA \end{aligned} \qquad (4)$$

RMW (Robust Minus Weak) represents the measure of the profitability premium. The CMA (Conservative Minus Aggressive) factor measures the investment premium, which represents the additional returns generated by companies that adopt a conservative investment strategy compared to those that pursue an aggressive investment approach.

2.3. OLS and LSTM-RNN models

OLS, or ordinary least squares, is a conventional statistical technique employed to estimate the unknown parameters in a linear regression model. The model aims to reduce the total of the squared discrepancies between the observed dependent variable in the dataset and the values predicted by the linear function. The OLS model is chosen as the baseline model in this paper due to its simplicity and wide application in statistical forecasting. The Ordinary Least Squares (OLS) estimator in matrix notation is formally expressed as:

$$\hat{\beta} = (X^\wedge T \times X)^{-1} \times X^\wedge T \times y \qquad (5)$$

X denotes the matrix of input data, which consists of independent variables. Y represents the vector of outputs, which is the dependent variable. β^ refers to the estimated parameters.

Unlike ordinary least squares (OLS), long short-term memory recurrent neural network (LSTM-RNN) is specifically designed to identify patterns in sequential data. This makes it particularly well-suited for analysing time series data, where the data points are arranged in chronological order. The LSTM architecture incorporates feedback connections that enable it to process not only individual data points, as in the case of ordinary least squares (OLS), but also entire sequences of data. A fundamental LSTM unit consists of a cell, an input gate, an output gate, and a forget gate. The cell retains values for indefinite time periods, and the three gates control the movement of information into and out of the cell. LSTM networks excel at mitigating the issue of long-term dependency by utilising the forget gate, which enables the model to eliminate information that is no longer pertinent to the prediction task.

The LSTM-RNN is employed due to its ability to handle sequential data and address the issue of the vanishing gradient, which is prevalent in conventional RNNs. The models were set up to assess their effectiveness, using parameters determined by past performance and unique characteristics of each model. [3] The LSTM model's parameters consisted of the batch size, the number of layers, and the number of nodes per layer. These parameters were optimised through preliminary testing. The following is the formula for the LSTM model:

$$f_t = \sigma_g(W_f \times x_t + U_f \times h_{t\,\_\,1} + b_f) \tag{6}$$

$$i_t = \sigma_g(W_i \times x_t + U_i \times h_{t\,\_\,1} + b_i) \tag{7}$$

$$o_t = \sigma_g(Wo \times x_t + U_o \times h_{t\,\_\,1} + b_o) \tag{8}$$

$$c_{t'} = \sigma_c(Wc \times x_t + U_c \times h_{t\,\_\,1} + b_c) \tag{9}$$

$$c_t = f_t \times c_{t\,\_\,1} + i_t \times c_{t'} \tag{10}$$

$$h_t = o_t \times \sigma_c(c_t) \tag{11}$$

The activation formula for the first forget gate is as follows: ft represents the forget gate, σg denotes the sigmoid function, xt represents the weighted sum of the current input vector, ht−1 represents the previous hidden state, and b represents the bias term. The activation of the second input gate is determined by the input gate it. The activation of the third output gate is represented by the variable ot. The variable ct′ represents the cell state, ht represents the hidden state, σg is the sigmoid function, and σc is the hyperbolic tangent function.

Practically, the process of training an LSTM entails providing it with a substantial quantity of sequential data, from which it acquires the ability to anticipate the subsequent steps in the sequence. For the purpose of financial forecasting, this could involve conducting training using past stock prices in order to anticipate future prices. The LSTM's capacity to retain long-term patterns and discard irrelevant information (via the forget gate) makes it highly suitable for tasks that involve not only complex relationships but also dynamic changes over time. LSTMs offer a strong framework for tackling challenges in financial time series forecasting, outperforming models that do not possess these advanced mechanisms for handling sequential data by utilising these capabilities effectively.

Ultimately, the authors employ Root Mean Square Error (RMSE) as a metric to quantify the precision of the models. The term is mathematically defined as the square root of the mean of the squared errors, which represents the average difference between the estimated values and the actual values being estimated. RMSE is a reliable indicator of the model's ability to accurately predict the response variable. It is the most suitable criterion for assessing the model's fit when the primary objective is prediction. The mathematical expression for calculating the Root Mean Square Error (RMSE) is as follows:

$$\mathrm{RMSE} = \mathrm{sqrt}\ [(\Sigma(Pi - 0i)^2)\ /\ n] \tag{12}$$

where Pi are the predicted values, Oi are the observed values, and n is the number of observations.

The Root Mean Square Error (RMSE) metric was employed to assess and contrast the precision of the OLS and LSTM-RNN models. This metric is especially valuable for quantifying the degree of error in predictions, offering a precise measure of the performance of the model. The LSTM-

RNN model was designed with a deep network architecture to effectively capture the intricate patterns in the financial time series data. The model's performance was consistently assessed using the rolling window method to compare it with the OLS model. This method is crucial for adjusting to dynamic market conditions.

3. Results

3.1. Two-stage regression results

While both models utilised rolling regression, the constrained model's exclusion of the intercept facilitated a more precise examination of the extent to which beta alone could account for returns. Removing the intercept can simplify the model and enhance the clarity of the model's outputs, making it easier to determine the impact of market risk. Although the conventional Capital Asset Pricing Model (CAPM) offers a fundamental comprehension of asset pricing based on the assumption of market efficiency and rationality, it frequently proves inadequate in real-world scenarios, particularly in highly volatile or intricate market conditions. [8] The constrained CAPM model, in contrast, provides a more streamlined and potentially more accurate approach by exclusively considering market risk as the factor influencing asset returns. [9] This is especially valuable in empirical research, where simplicity and precision are highly valued.

To enhance practical applications, portfolio managers and individual investors can utilise the refined model to achieve more precise predictions of stock returns. This, in turn, can result in improved portfolio management and investment strategies. Moreover, the constrained model emphasises the evident correlation between stock returns and market risk, which can be advantageous for enhancing risk management and asset pricing. Nevertheless, this paper stimulates additional contemplation and areas for further investigation. To conduct additional model testing, it is possible to test the constrained CAPM model in various markets and conditions in order to generalise the findings or refine the model further. [10] In addition, future research could incorporate additional explanatory variables, such as size, value, and momentum, in conjunction with beta, to construct multi-factor models that have the potential to enhance the accuracy of stock return predictions. In addition, employing machine learning techniques to optimise the parameters and select variables could further improve the accuracy and flexibility of the asset pricing models.

3.2. CAPM, FF3 and FF5 models

The FF3 and FF5 models utilize the traditional CAPM model as a reference point to comprehend asset returns solely based on market risk. The Fama-French models, unlike the CAPM, include extra factors that consider different aspects of risk and return. These factors are used to measure the Size and Value Premiums (FF3 and FF5 models). Both models suggest that smaller companies and companies with high book-to-market ratios tend to perform better, resulting in premiums that are not accounted for by the CAPM. Additionally, the FF5 model incorporates factors related to profitability and investment premiums, aiming to explain variations in returns that are not captured by the FF3 model.

The results are compared with findings from prior studies to assess the performance of the models in relation to established theories and other empirical evidence. A comparative graph can illustrate the degree of similarity or divergence between the findings of this study and those of similar studies conducted in various markets.

The study concludes by providing a concise overview of the main discoveries, highlighting the exceptional performance of the FF3 model within the specific context of the Chinese market. The study acknowledges its limitations and proposes future research directions, such as investigating supplementary factors or conducting tests in other emerging markets.

3.3. OLS and LSTM-RNN models

The results demonstrate that the LSTM-RNN model outperforms the OLS model across different portfolios. In portfolio 7, the LSTM-RNN model achieved an RMSE of 1.849, while the OLS model had an RMSE of 2.355. This indicates that the LSTM-RNN model has a more accurate and dependable forecasting ability. The graphs clearly demonstrate the improved predictive ability of LSTM-RNN by comparing the RMSE values and portfolio performances over time. These visual aids are essential for comprehending the behaviour of the model across various market phases.

The paper concludes that the superior performance of LSTM-RNN compared to traditional models such as OLS can be attributed to its sophisticated architecture, which enables it to effectively learn and capture long-term dependencies in data. This is especially beneficial in financial markets, as historical data often serves as a reliable indicator of future patterns. The study affirms that LSTM-RNN models are not only appropriate but also surpass traditional forecasting methods in the realm of financial market predictions. Their capacity to adjust to data with non-linear patterns and sustain performance over time provides a notable advantage for fund managers and investors. Subsequent investigations should examine the amalgamation of LSTM-RNN with other machine learning methodologies and the incorporation of supplementary predictive factors to augment the precision of forecasting.

4. Conclusion

This essay has examined the constraints of the Capital Asset Pricing Model (CAPM) and illustrated how the incorporation of machine learning methods can improve its predictive precision. After conducting a thorough examination of three different approaches—Two-stage regression analysis, comparisons with the Fama-French three-factor and five-factor models, and the utilization of Ordinary Least Squares (OLS) and Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN)—it is clear that the Capital Asset Pricing Model (CAPM) by itself does not offer the most precise forecasts.

The Two-stage regression analysis demonstrated a notable enhancement in the accuracy of predictions when machine learning algorithms were utilized in conjunction with the traditional CAPM. Comparatively, when analyzing the FF3 and FF5 models, it became evident that CAPM has limitations in capturing market anomalies, which were effectively addressed by machine learning models. Furthermore, the utilization of Ordinary Least Squares (OLS) and Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) models emphasized the ability of machine learning to adjust to intricate market patterns and improve the forecasting accuracy of the Capital Asset Pricing Model (CAPM).

These findings emphasize that although the Capital Asset Pricing Model (CAPM) offers a fundamental understanding of asset pricing, it is not adequate by itself for making optimal forecasts. By incorporating machine learning techniques, the performance of the Capital Asset Pricing Model (CAPM) is greatly improved. This is achieved by effectively capturing non-linear relationships and being able to adapt to dynamic market conditions. By integrating machine learning with CAPM, we not only address the inherent limitations of CAPM but also enhance its predictive capabilities, resulting in a more robust and dependable tool for financial forecasting. This integrated approach ultimately offers investors and analysts a more robust framework for making well-informed decisions in an increasingly intricate and ever-changing financial landscape.

References

[1] Fama, E. F., & French, K. R. The Cross-Section of Expected Stock Returns.
[2] Roondiwala, M., Patel, H., & Varma, S. Predicting stock prices using LSTM.

[3] Li, H. The performance of Fama-French asset pricing models in the Chinese stock market. Rabha, D., & Singh, R. G. Is CAPM still valid in today's market scenario?

[4] Huynh, T. T., & Khoa, B. T. Utilizing LSTM-RNN algorithm in a multi-factor model for forecasting investment portfolio returns.

[5] Judijanto, L., Mendrofa, Y., Harsono, I., Sebayang, P., & Johari, F. Modern approaches to risk management in investment portfolios: Strategies in market volatility.

[6] Yang, J., Zhang, M., Feng, S., Zhang, X., & Bai, X. A Hierarchical Deep Model Integrating Economic Facts for Stock Movement Prediction. Engineering Applications of Artificial Intelligence. Retrieved from www.elsevier.com/locate/engappai.

[7] Fajarini, N., & Heikal, J. Reassessment of CAPM Relative Accuracy: Comparative Study with Actual Price Movement in Indonesia. DOI: https://doi.org/10.54099/ijmba.v3i1.743.

[8] Muddasir, M., & Kulalı, G. The Validity of CAPM and ICAPM in the Istanbul Stock Exchange. Journal of Research in Economics, Politics & Finance, 9(1), 26-42. DOI: https://doi.org/10.30784/epfad.1383837.

[9] Rabha, D., & Singh, R. G. Is CAPM still valid in today's market scenario? Indian Journal of Finance. Retrieved from http://gnanaganga.inflibnet.ac.in:8080/jspui/handle/123456789/7854

[10] Hazra, A., Kayal, P., & Maiti, M. An examination of the Indian small-cap cycle in relation to the US market. Retrieved from https://doi.org/10.1016/j.iimb.2024.03.008