Idea Generation for Price Prediction Signal for a Macro Asset or ETF

Likang Wang

Doctor of Philosophy

Department of Computer Science and Engineering

Hong Kong Univeristy of Science and Technology

Hong Kong

lwangcg@connect.ust.hk

Abstract

Accurately predicting the future trends of macro assets and broad-based index ETFs is of paramount importance for the regulation of government policies and the healthy development of capital markets. In this paper, we first conduct a comprehensive analysis of the characteristics of various macro assets and broad-based index ETFs, and on this basis select the S&P 500 as our primary research subject. We thoroughly and deeply summarize the unique characteristics of the S&P 500, and propose a promising price prediction framework. Our preliminary implementation and the corresponding excellent results strongly validate the reliability of this framework. Beyond this, based on the strengths and weaknesses of the current model, we provide beneficial insights on how to further improve the model.

1 Introduction

As per the topic's requirement, we are tasked to select and forecast the price of a macro asset or broad-based index ETF from the myriad of options available. This paper sets its focus on the S&P 500 Index, a premier indicator of the U.S. stock market and a barometer of the global financial market. Our choice to forecast the S&P 500 over other macro assets stems from several reasons:

- **Market Representativeness:** The S&P 500 Index is comprised of the 500 largest companies within the U.S. stock market, establishing it as an exceptional representative for gauging the performance of not only the U.S. but also the global stock markets. Unlike other indices that cover specific industries or smaller companies, the S&P 500 offers a more comprehensive market perspective. For instance, unlike the Dow Jones Industrial Average (DJIA), which is more industrially focused, the S&P 500 provides broader representativeness, including technology-driven companies. In contrast, the NASDAQ concentrates on tech firms without the extensive industry coverage found in the S&P 500. A detailed comparison of the comprehensive characteristics of these three indices is available in Table 1.
- **Economic Impact:** The fluctuations of the S&P 500 Index are considered predictive of the broader economic trajectory. Because it spans multiple industries, its performance is often closely related to national economic health, making it a vital indicator for decision-makers and analysts.
- **Liquidity and Investor Attention:** Due to the large-cap nature of its constituent stocks, the S&P 500 Index typically enjoys high liquidity, drawing significant attention from

Preprint. Under review.

institutional and individual investors. This high liquidity and concentrated investor interest make forecasting the S&P 500 more crucial compared to other macro assets.

- **Availability and Reliability of Data:** Due to its significance, data on the S&P 500 Index is more readily accessible and typically of higher quality. This facilitates deeper analyses of the index, making the S&P 500 an ideal subject for quantitative analysis and machine learning modeling.
- **Development of Forecasting Tools and Methods:** With the advancement of financial technology, numerous forecasting tools and analytical methods have been optimized for major indices like the S&P 500. For such indices, forecasting models are more mature, with more extensive related research and practice.

Table 1: Comparative Analysis of Financial Indices

Index	Coverage	Sector Weightings	Volatility
S&P 500	Broad market exposure	Diverse, including IT, Health Care, Consumer Discretionary	Moderate
DJIA	30 large-cap industrial corporations	Industrials, Financials	Lower
NASDAQ	Over 3000 components, Heavy in tech	Information Technology	Higher

Other macro assets, such as commodities, currency exchanges, or government bonds, are undoubtedly important. However, from an investment analysis and macroeconomic strategy perspective, the S&P 500 is widely considered the cornerstone for portfolio construction, financial forecasting, and understanding market dynamics due to its broad coverage and wealth of information. Therefore, while forecasting other macro assets is valuable, forecasting the S&P 500 is particularly critical because of its unique attributes and influence.

The remainder of this article delves into how to accurately forecast the S&P 500 Index, providing a thorough experimental analysis of the proposed methods and concluding with potential future directions for work.

2 Conceptual Framework

2.1 Task Analysis

In our previous paper, we have successfully forecasted the future trends of the S&P/Case-Shiller U.S. National Home Price Index. Considering that the U.S. National Home Price Index and the S&P 500 datasets are both time-series reflecting the prices of assets, it's imperative to analyze the intrinsic technical similarities and differences of these two tasks. We contend that the main technical differences between forecasting the S&P 500 Index and the S&P/Case-Shiller U.S. National Home Price Index (referred to hereafter as the "Case-Shiller Housing Price Index") predominantly lie in the characteristics of the data, market factors, and choice of models:

- **Data Characteristics and Seasonality Patterns:**
  - The S&P 500 Index reflects the overall performance of the stock market, with data volatility being more susceptible to market sentiment, economic policies, rate changes, and other factors, and the data is updated frequently (usually daily).
  - In contrast, the Case-Shiller Housing Price Index illustrates trends in the U.S. housing market's prices, influenced by the macroeconomic environment, regional supply and demand, as well as seasonal changes (for instance, the housing market exhibits clear seasonal patterns, with prices usually rising in spring and summer due to increased buying activities), and data are updated less often (typically monthly).
- **Market Factors and Economic Indicators:**

  - When forecasting the S&P 500, analysts must consider a wide range of macroeconomic metrics, company earnings reports, and political events on a global scale.
  - In predicting the Case-Shiller Housing Price Index, however, the focus shifts to factors directly related to the real estate market, such as housing inventory levels, new housing construction rates, employment rates, consumer confidence indices, and mortgage loan rates.
- **Model Selection and Analytical Methods:**
  - Due to the volatility and complexity of the S&P 500 data, often-used models for its prediction include Random Forest (Breiman, 2001), deep learning models like LSTM (Hochreiter and Schmidhuber, 1997), and other advanced machine learning models capable of handling non-linear relationships and capturing complex market dynamics.
  - For the Case-Shiller Housing Price Index, time series analysis methods (such as ARIMA (Box et al., 1976) models, seasonal decomposition (Cleveland et al., 1990), etc.) are more commonly employed due to the seasonality and update frequency of the data, which can effectively process and predict data with notable seasonal patterns and long-term trends.

Overall, predicting these two indices requires different methods and models tailored to their respective data characteristics and influencing factors. Moreover, accurate forecasting of both the S&P 500 and the Case-Shiller Housing Price Index necessitates a deep understanding of the related markets and meticulous analysis of extensive historical data.

2.2 Model Selection

Given the capabilities of LSTM in predicting the Case-Shiller Home Price Index and considering that predicting the S&P 500 is not likely to be less challenging, it seems fitting to start our exploration with LSTM. Should it underperform, we'll then contemplate adopting time-series models with greater fitting potential, such as Transformers (Vaswani et al., 2017) and Informer (Zhou et al., 2021).

2.3 Data Selection

Another crucial question is which independent variables should be used as inputs. There are numerous factors influencing stock market performance; we believe the following to be particularly significant:

- **Company Earnings:** The earnings reports of the S&P 500 constituent companies, especially Earnings Per Share (EPS) and revenue, notably impact the index. Both positive or negative results can cause fluctuations.
- **Economic Indicators:** Economic data, such as GDP growth rate, employment figures including unemployment rates, inflation, and manufacturing output, all affect the S&P 500. Economic expansion is generally indicative of rising company profits and stock prices, whereas recessions might trigger declines.
- **Interest Rates:** Policies on interest rates set by the Federal Reserve influence the S&P 500. Typically, low-interest rates foster investment in the stock market, elevating index values, whereas high rates might deter stock investments due to bonds and savings accounts becoming more appealing options.
- **Geopolitical Events and Major News:** Political unrest, trade agreements, and international conflicts could bring about market uncertainties, affecting investor sentiment and the performance of the S&P 500.
- **Global Economic Factors:** Considering the global operations of the companies within the index, global economic health—including foreign exchange rates, economic sanctions, and international trading dynamics—can impact the S&P 500.
- **Commodity Prices:** Fluctuations in the prices of commodities like oil can have a bearing on certain sectors within the S&P 500, and in turn, influence the entire index's trajectory.
- **Sector Performance:** The S&P 500 encompasses various sectors, and the strength or weakness of specific sectors such as technology or healthcare can affect the direction of the index.

- **Market Sentiment:** Driven by various elements such as market trends, analysts' forecasts, and broader economic sentiment, investor perspectives can affect trading behaviors, which in turn affect the S&P 500.

These elements together add to market complexity and highlight the importance of a multifaceted approach to analyze and predict the S&P 500's movements. Regarding how to incorporate the aforementioned influencing factors, our overall concept is:

- For statistical quantities such as company earnings (e.g., EPS and revenue), economic data (like GDP growth rate, employment data including unemployment rate, inflation, and manufacturing output), interest rates, commodity prices, and foreign exchange rates, where direct numerical values can be obtained, we can use their raw data (which might require normalization) as features fed into the neural network.
- For non-quantifiable data like geopolitical events and major news (e.g., political instability, trade agreements, international conflicts), global economic factors (like economic sanctions and international trading dynamics), sector performance (e.g., the emergence of promising new technologies and broad new markets), and market sentiment (such as influential market commentary from major companies or key opinion leaders), we can employ pre-trained large language models or other natural language processing models to infer the impact of these events on the market (such as outputting a positive/neutral/negative classification label, a floating-point number between -1 to 1, or a multi-dimensional embedding). These quantified values can then be incorporated as inputs into the neural network.

We considered various aspects of the market. However, as a starting point for the research, we believe it is essential to first simplify the problem to obtain some preliminary conclusions. Therefore, we will first attempt to use only the raw S&P 500 series with LSTM. If the outcome is unsatisfactory, we will consider gradually integrating more data and larger models.

3 Methodology

In this section, we will detail how we implemented a functional and well-performing S&P 500 forecast model based on the approach mentioned in the previous section.

3.1 Data Acquisition

We obtained the entire historical data of the S&P 500 (from December 30, 1927, to May 10, 2024) from Yahoo Finance and adhered to the mainstream setting of using closing prices for training and validation. Figure 1 displays the variation of the S&P 500 closing prices over time.

We believe using closing prices offers the following six advantages:

- **Market Consensus:** The closing price is the last transaction price at the end of the trading day and is considered to reflect the consensus among market participants. In contrast, other prices (such as the opening, highest, and lowest prices) may be more susceptible to short-term market fluctuations.
- **Stability:** The closing price is more stable than other transaction prices throughout the day and is not affected by short-term fluctuations. For example, the opening price may be influenced by overnight news events, whereas the closing price is the result of considering all market information throughout the day.
- **Technical Analysis:** Many technical analysis methods and trading algorithms are designed based on the closing price. This is because the closing price is often used to calculate various technical indicators, such as moving averages (Murphy, 1999), the Relative Strength Index (RSI) (Wilder, 1978), and Bollinger Bands (Bollinger, 2001).
- **Data Integrity:** For most stocks, the closing price is the most complete and easily accessible data point. The opening, highest, and lowest prices may not be accurately recorded due to market interruptions or other issues.
- **Trading Strategy:** Many traders base their trading strategies for the following day on the closing price, making it a crucial basis for planning trades and estimating market trends.

Figure 1: Changes in the S&P Historical Close Price Over Time.

- **Market Cycles:** The closing price provides clear daily, weekly, or monthly market cyclical information, which helps in identifying trends and patterns.

3.2 Data Partitioning

Regarding the partitioning of our dataset, we adhered to conventional norms, allocating the initial 90% of the sequential data as the training set (encompassing the period from December 30, 1927, to September 15, 2014, equating to roughly 87 years and a total of 31,672 samples) and assigning the subsequent 10% of the data stream for validation purposes (spanning from September 16, 2014, to May 10, 2024, again covering nearly a decade and comprising 3,520 samples in total). The temporal continuity maintained across the dataset ensures the preservation of local correlation and global cyclicity between data points. Amassing a substantial volume of data is conducive to the training of deep learning models and lends statistical credibility to the validation outcomes.

3.3 Data Preprocessing

As stock markets are not open for trading during weekends, public holidays, or special events (such as national mourning), there will be no trade data for these days. For the convenience of training, linear interpolation has been performed for days in the training set that lacked actual values. As can be seen from Figure 2, the effects of interpolation are quite favorable, showcasing a high degree of alignment with the original data.

3.4 Model Configurations

Given our prior success in housing index prediction, we opted to retain the same LSTM (Hochreiter and Schmidhuber, 1997) architecture. That is, setting the number of hidden layers to 2, each layer containing 48 neurons, using the Adam (Kingma and Ba, 2014) optimizer with a learning rate set at 0.01, and the original data will be linearly normalized based on maximum and minimum values before being input into the model and after being output by the model. However, considering the significantly larger amount of data points here, we increased the number of gradient updates from 5000 to 10000 and multiplied the learning rate by a cosine function (Loshchilov and Hutter, 2016) to facilitate convergence.

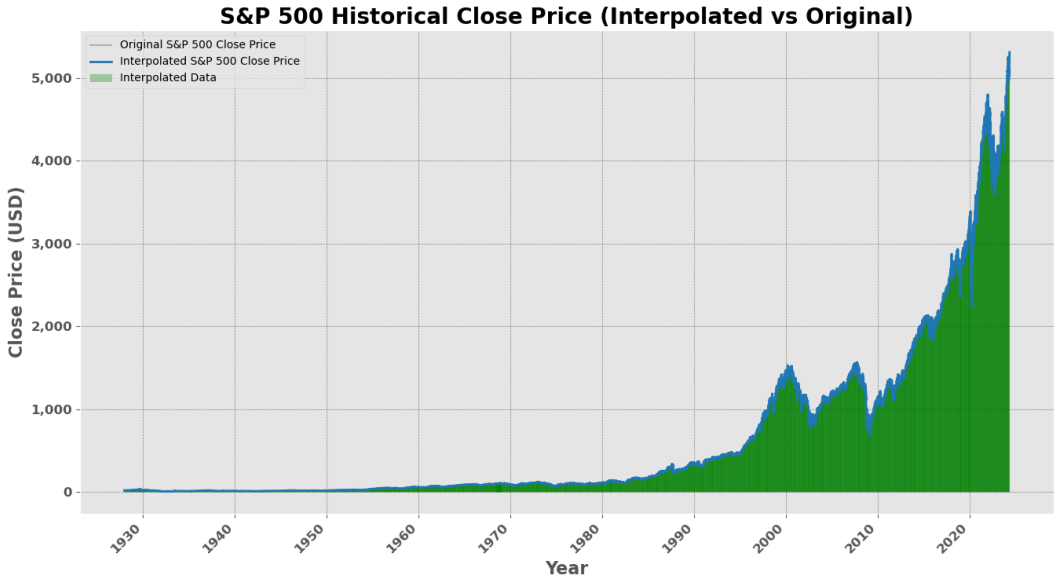

Figure 2: The Impact of Interpolation on the S&P Historical Close Price

3.5 Evaluation Metrics

Since the model's inputs and outputs are continuous floating-point numbers, we believe that a combination of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Relative Error (MRE) can comprehensively evaluate the model's performance. These metrics are calculated as shown in Equations 1, 2, and 3, where n is the total number of time frames, $\hat{y}_i$ is the predicted closing price at time i, and y_i is the actual closing price. Both RMSE and MAE measure absolute error in closing price; the key difference is that the former is more sensitive to larger errors, while the latter does not overlook small errors. MRE is a normalized metric, less affected by the scale of the data. Additionally, as closing prices undergo drastic changes over time, the same absolute error can have different implications at different times. The MRE can account for this issue.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{2}$$

$$\text{MRE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{3}$$

3.6 Result Analysis

As shown in Figure 3, the curve we predicted almost perfectly coincides with the true values. Moreover, dramatic changes in the stock market were accurately reflected, such as the sharp decline between February 7, 2020, and March 27, 2020, and the significant rise from September 22, 2023, to March 28, 2024. Table 2 further quantitatively demonstrates the precision of our S&P 500 predictions. Since the errors are minimal on both the training and test sets, we believe there is no occurrence of overfitting.

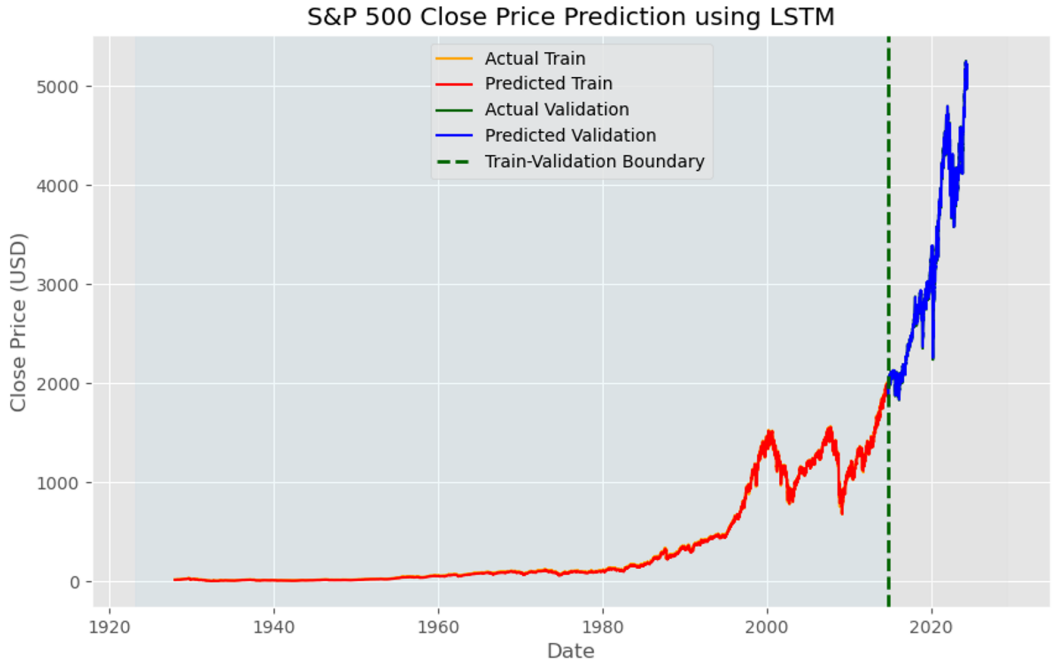

Figure 3: Visualization of Using LSTM to Predict Future Values of the S&P 500 Close Price

Table 2: Numerical Analysis of Using LSTM to Predict the S&P 500 Close Price.

Metric	**Value**
Root Mean Squared Error (Validation)	24.822
Mean Absolute Error (Validation)	14.594
Mean Relative Error (Validation)	0.004
Root Mean Squared Error (Train)	6.755
Mean Absolute Error (Train)	2.769
Mean Relative Error (Train)	0.002

3.7 Deliverability

Considering the high accuracy of the predictions mentioned above, and since the main task is to examine concept generation, as well as preliminary code implementation and evaluation, we believe the current model already meets the project's requirements.

3.8 Future Work

While we have already obtained satisfactory results at this stage, we believe there still exist ample opportunities for further improvement. For future enhancements, we regard the following directions particularly promising:

- Introduce more input variables. Specifically, we can refer to the methods illustrated in Section 2, treating numerical and non-numerical inputs separately. This strategy would effectively raise the performance ceiling of the model.
- Attempt to increase the model's fitting power. This move would allow our model to observe more distant data points, leading to a better recognition of periodic patterns.

- Integrate more domain knowledge pertaining to finance. This strategy can reduce the model's reliance on data.

4 Conclusion

This paper systematically investigates the characteristics of various macro assets and broad-based index ETFs, from which we selected the S&P 500 for future value prediction. We analyzed the characteristics of this data in detail and depth, and proposed a promising overall model framework. Based on the proposed framework, we carried out preliminary implementation and verification, and achieved good results. Beyond this, we also analyzed in depth the advantages and shortcomings of the current method, and provided potential routes for improvement.

References

John Bollinger. *Bollinger on Bollinger Bands*. McGraw Hill, 2001.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 1976.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Robert B Cleveland, William S Cleveland, Jean E McRae, Irma Terpenning, et al. Stl: A seasonal-trend decomposition. *J. Off. Stat*, 6(1):3–73, 1990.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

John J Murphy. Technical analysis of the financial markets. *New York Institute of Finance*, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

J. Welles Wilder. New concepts in technical trading systems. 1978.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11106–11115, 2021.