Dean&Francis
ISSN 2959-6157

Statistical and Machine Learning Methods for Stock Price Prediction

Chun Hei Feng

Diablo Valley College, California, 94523, United States
Corresponding Author: cfeng846@insite.4cd.edu

Abstract:

Financial markets are very dynamic and thus difficult to predict the price of stocks. There are financial theories like the efficient market hypothesis, which indicate that the future prices of stocks are hard to predict. Nonetheless, empirical research also suggests that there can be some patterns and inefficiencies in financial markets. In this paper, the author reviews and analyzes the typical methods that researchers apply to stock price predictions, such as traditional statistical methods, machine learning methods, and deep learning methods. The paper is aimed at comparing the various methods used to analyze financial time series data, including ARIMA and GARCH models, and machine learning algorithms, including the support vector machine, random forest, and neural network. Besides, the latest trends regarding the implementation of deep learning and the utilization of alternative data are explained to give a more comprehensive outlook on the contemporary prediction frameworks. According to the existing literature, machine learning and deep learning models are likely to be more successful in attaining the nonlinear relationship and the temporal dependence in stock prices than traditional statistical models. Nevertheless, their performance is very dependent on the quality of data, selecting features, and the market conditions.

Keywords: Stock price prediction; Machine learning; Deep learning; Financial markets.

1. Introduction

There is a lot of uncertainty and unpredictability in the financial markets. Thus, it has been a proven difficult task in the past to accurately predict the movement of stock prices. Numerous investors, analysts, and researchers are trying to come up with models that can enable them to consistently make investment decisions. Nevertheless, due to the numerous factors affecting financial markets, which include macroeconomic indicators, company-specific data, investor sentiment, and world events, it has become very challenging to create an effective model.

Financial theory indicates that the Efficient Market

Hypothesis postulates that the prices of stocks reflect all information existing about businesses. The consequences of this theory are that the stock prices are randomly moving or randomly walking. The implication is that any sort of analysis cannot always make meaningful predictions. Despite the theoretical stand on efficient markets, studies have shown that the markets might have certain patterns and inefficiencies. These inefficiencies give a chance to those who desire to use statistical models to analyze and forecast stock price changes. This is the reason that the researchers have used a wide range of methodologies, including simple statistical models, complex machine learning, and deep learning methodologies, to help them in their comprehension and prediction of the stock price movement. A number of questions have been developed by implementing stock price forecasting using different types of statistical models and previously utilized machine learning frameworks.

This paper offers several contributions to the previously published literature. Firstly, this paper provides an organized and comparative analysis of each of the statistical modeling, machine learning techniques, and deep learning strategies that have been used for predicting stock prices, while identifying advantages and disadvantages for each type of modeling technique or methodology. Secondly, this paper focuses on data quality, feature selection, and market environment as critical factors influencing the performance of each of the models presented here, providing a more practical perspective in addition to theoretical analysis. Lastly, this paper examines the increasing importance of using alternative data and hybrid modeling approaches when predicting financial markets, as well as offering potential directions for future studies in financial prediction.

The result of the paper is organized as follows. Sec. 2 will present the models and methods, Sec. 3 will show the Financial Economics Perspective, and Sec. 4 will display Alternative Data. The last section is the conclusion.

2. Models and Methods

2.1 Statistical Models

In financial prediction, traditional statistical techniques constitute one of the simplest aspects of finance to forecast stock prices, and were one of the earliest instruments to be used to forecast financial events. The prevalence of statistical techniques is explained by the possibility to clearly delineate a mathematical framework of the underlying phenomena of interest; furthermore, statistical techniques are easier to comprehend than most other predictive models. Financial time series data are also normally applied to statistical methodologies to predict the future behavior of those variables and potential patterns or interrelationships among the variables. Although statistical models are a little older in their development, they still remain as a point of comparison in the evaluation of more recent prediction methods.

Prior to the recent rapid development of deep learning, traditional statistical models were used by the majority of stock price forecasting research papers, such as regression models, time-series models, and classification algorithms. There was general use of various statistical models, which were easy to interpret and analyze financial time series. They usually used the autoregressive integrated moving average (ARIMA) and generalized autoregressive conditional heteroskedasticity (GARCH) models. Nevertheless, such statistical models typically make the assumption that the relationship is linear and the process is stationary, which is not often the case with most financial markets.

2.2 Machine Learning Methods

The increasing popularity of machine learning models in predicting stock prices is due to their ability to formulate the complex and nonlinear relationships that cannot be established through statistical methods. Machine learning algorithms are more flexible than statistical methods, which are based on assumptions of stationarity and linearity. This enables them to be applied to high-dimensional data sets when variables interact with one another in complex ways. Machine learning is a trendy methodology because it enhances the precision of predictions in the financial markets.

Due to the above-mentioned assumptions, scientists began to apply machine learning algorithms to determine the nonlinear patterns that occur in financial data. As an example, Patel et al. examined a variety of machine learning methods [1], such as artificial neural networks, support vector machines, and random forests, to predict stock index changes. Their conclusion was that when people add both the technical indicators of a stock and machine learning algorithms, people will, to a large extent, enhance people's chances of accurately predicting the stock prices as compared to when people make use of the traditional statistical techniques. Along with the research of Patel et al., Ballings et al. experimented on a wide range of classification algorithms when it comes to stock direction prediction [1,2]. It was established that ensemble models, e.g., random forests and boosting methods, would tend to work quite well as compared to single classifiers. These two papers proved that machine learning methods can detect and quantify more complicated relationships in financial data. The success of machine learning solutions in arriving at

ISSN 2959-6157

correct predictions continues to depend on the quality of features chosen and the market conditions.

More recent studies have involved empirical evidence on a large scale to substantiate this argument. According to the research of Gu, Kelly, and Xiu [3], machine learning models could predict cross-sectional stock returns to a larger extent than the conventional econometric models. According to them, non-linear interaction among variables had valuable predictive information that could not be portrayed through linear models. Based on their findings, Cakici et al. showed that machine learning models can accurately forecast stock returns in international markets [4], which implies that machine learning models are resilient when applied in other economic settings. On the same note, Hanauer and Kalsbach found that machine learning models were more effective in emerging markets than in other forms of markets since inefficiency and informational asymmetry are more common in such markets [5]. These works imply that machine learning can be valuable not only in terms of enhancing predictive accuracy but also in helping researchers uncover new patterns or structures that can be present in financial data. Nevertheless, like in other fields of finance, machine learning is not guaranteed to bring good results in all scenarios because it relies on quality data, effective selection of features, and good market conditions.

2.3 Deep Learning Approaches

In recent years, researchers have shown greater interest in deep learning approaches because they are a more advanced form of machine learning compared to traditional machine learning approaches. The primary distinction between deep learning and conventional machine learning is that deep learning is capable of discovering numerous levels of abstraction using larger volumes of data. Since financial time series are characterized by numerous non-linear relationships among each other, time relationships, and noise, deep learning is highly applicable in the analysis of such sequences. With both growth in computing power and the availability of data, it seems that deep learning can be among the most suitable solutions to creating predictive stock-price models.

A second research field that has changed in the last decade is the use of deep learning techniques to forecast financial data. The financial forecasting aspect of deep learning has gained growing popularity because of the ability of deep learning to automatically learn hierarchical features in large volumes of data and nonlinear temporal dependencies. Fischer and Krauss carried out one of the initial investigations of the use of long short-term memory (LSTM) networks to predict the financial market [6]. This paper demonstrated that LSTM models can also make more correct predictions than conventional machine learning models when predicting stock returns in large equity markets.

Similar to the study of Fischer and Krauss [6], Bao, Yue, and Rao [7] developed a new method of forecasting using stacked autoencoders combined with LSTM networks to model financial time series. The authors were in a position to develop a process that possessed greater predictive power than conventional neural networks and autoregressive models. There have been subsequent studies that have expanded on the concept of this study since the initial study of Bao, Yue and Rao [7], by employing several deep learning architectures. A recent study that was published is the study of Kanwal et al. [8], who created a hybrid predictive model that employed a blend of bidirectional LSTM and convolutional neural networks to detect time dynamics as well as local feature patterns in stock data.

For instance, the design of Kanwal et al. can be defined as a hybrid one, as the authors have employed Bidirectional LSTMs [8], together with CNNs. By using the method, the model is capable of processing not only long-term temporal dependencies but also short-term localized patterns of a specific set of historical stock prices. Similarly, Rezaei, Faaljou, and Mansourfar also adopted an integration approach between frequency domain decomposition and deep learning approaches [9]. They have shown that their work has better performance in the case of non-stationarity of financial time series. These more recent works suggest that the models in the field of deep learning are still being developed, starting with simple architectures and moving to more complex and flexible architectures, capable of capturing complex dynamics in the financial markets.

The other recent paper was by Rezaei, Faaljou, and Mansourfar [9], who also created a deep learning model in which a mixture of frequency decomposition and neural networks is used to enhance the analysis of non-stationary financial signals. Collectively, these works demonstrate how deep learning models have resulted in significant advances in modeling intricate financial processes, particularly in scenarios where extensive data and computing power are accessible.

There are a number of limitations to deep learning models. As an example, these models should be trained using huge volumes of quality data. Unluckily, such data is not always available or exceptionally difficult to get in financial markets. The other disadvantage of deep learning models is that they are commonly known as black box systems. That is, it is highly challenging to know the reason behind a specific forecast. In finance, it is equally important to know the rationale behind a forecast as well as the forecast itself. Due to the complexity of interpreting

outputs of deep learning models, numerous researchers do not suggest replacing other types of machine learning (traditional) and statistical models with deep learning, but rather co-exist.

1. Financial Economics Perspective

Much of the present-day research has ceased to focus on simply applying machine learning techniques to enhance the accuracy of the prediction, to understand how machine learning techniques can assist us in learning more about the nature of financial markets. Researchers would rather know whether the models will produce insight into such things as asset pricing, predictability of returns, and market behavior rather than merely testing the hypothesis that a model generates good forecasts. The fact that more focus is placed on the acquisition of insight into the consequences of applying machine learning techniques to the financial domain constitutes a more significant paradigm shift in financial economics. In more recent times, predictive models are not regarded as the means of producing forecasts alone but also as the way of testing theoretical hypotheses and of achieving a better insight into the underlying arrangements of the financial markets.

Besides purely predictive modeling, there is another literature on the bigger connection between machine learning and financial economics. Rather than assessing the consistency of price forecasts based on machine learning, this literature assesses the ability of machine learning to improve the knowledge of asset pricing and predictability of returns. In a general empirical investigation to compare the traditional econometric models and machine learning methods to predict cross-sectional stock returns, Gu, Kelly, and Xiu studied the issue [3]. The findings of this paper show that machine learning algorithms are capable of observing nonlinear correlations between economic factors and make better forecasts than the old linear factor models. This study suggests that machine learning is capable of enriching traditional financial theories and not replacing them. Building on the concept, Cakici et al. conducted a study to identify the performance of machine learning models in forecasting the international stock returns across different markets [4]. The findings of the current paper are that machine learning models work best when applied to predict the stock returns with large cross-sectional data that include a large number of firm-specific features. Besides this, Hanauer and Kalsbach investigated the machine learning practices in the emerging equity markets and found that nonlinear models tended to perform better than the traditional model in the high-volume and inefficient information markets [5].

Cakici et al. [4], for example, investigate the effectiveness of machine learning algorithms in a variety of international markets. They conclude that they have the best applications when they use large cross-sectional data sets containing lots of firm-specific information. Hanauer and Kalsbach present results of the effectiveness of machine learning methods in new markets where conventional models are less applicable due to a higher degree of volatility and lower information effectiveness [5]. Because of this, the research by the authors indicates that the usefulness of machine learning can be determined by the nature of the algorithm that is being used and the nature of the market place that the algorithm is applied.

Additionally, Murray, Xia, and Xiao dispute the traditional idea of the lack of empirical evidence of technical analysis by applying machine learning methodologies to assess the different technical indicators [10]. The authors prove that not only can machine learning methods predict some technical indicators. Besides that, Ait-Sahalia et al. also examine the high-frequency stock returns and suggest that microstructural impacts in the market [11], such as order flow and trade friction, are the key determinants of predictability in the short term. Both studies show that the problem of predicting stocks is not merely a statistical/computational problem, but rather a financial issue per se, and directly relates to the essence of market structure and trader behavior.

The studies above prove that machine learning can have a much greater value in the world of finance than merely forecasting price movement in the future. The machine learning methods are able to examine the dynamics of returns, the relationship between variables, and the market structure of the markets.

2. Alternative Data

Besides applying more sophisticated methods in predicting future stock prices with past prices, other types of data are being integrated into the most sophisticated predictive models. The concept here is to employ alternative sources of information other than just previous values and various ratios on the basis of financial statements. High-frequency trading data, textual data in the form of news, social media, and other electronic forms of communication, and technical patterns are some examples of the kind of alternative data that can be used to predict the price of the stock. Alternative data are the possible additional signals that can be discovered and modelled outside the traditional models. The market sentiment projected by the different forms of electronic communication may also directly influence the decision of an investor on his/her investments, which will also tend to influence the price of the stock of a company.

The classical stock forecasting models can typically be restricted to past prices and financial ratios. While there have been many studies recently regarding the incorporation of alternative data types, for example, text-based,

ISSN 2959-6157

high-frequency trading data, and technical patterns to improve forecasting models, researchers have found that the use of additional data types can improve the forecasting performance. Sirignano and Cont demonstrated that deep neural networks could extract universal features of price formation in financial markets using high-frequency limit order book data [12]. The researchers' study indicated that the universality of price dynamics in financial markets may be due to specific structural characteristics of those markets.

Similarly, a growing body of literature has shown how the use of alternative data may increase model performance. A good example is analyses related to sentiment, which demonstrate that textual data, in addition to other quantitative factors from financial news and social media, may create additional predictive power for investors [12, 13]. The results found in this study support the notion that markets' behavior is affected as much or more than just quantitative indicators, but also by investors' sentiment and expectations.

Conversely, Murray, Xia, and Xiao also used machine learning to establish the predictive capability of typical chart patterns [10]. The researchers demonstrated that there are certain common patterns of charts that have traditionally been employed in the support of technical analysis that indeed do have predictive ability on the basis of modern computational methods. Trying to develop this point of view further, Ait-Sahalia et al. carried out research that measured the predictability of high-frequency stock returns [11]. They found out that predictability in short-term returns seems to be due to microstructure effects and trading frictions, and not necessarily explained by fundamental information.

As much as there are benefits of using alternative data, it also has its challenges. Alternative high-frequency data is commonly noisy and extremely hard to process. Sentiment data is both biased and incomplete [14]. Moreover, the inclusion of extra data may not always provide extra predictive power to a model. It can happen on numerous occasions that models may become less precise following the existence of large volumes of irrelevant or redundant information.

3. Conclusion

Altogether, the literature on stock price prediction seems to be an emerging field. Statistical models do offer a theoretical construct to work with in terms of making predictions, but they are not applicable to the nonlinear and dynamic behavior of financial markets. In like manner, machine learning techniques can be used to enhance prediction power by identifying the vast number of potential relationships between predictor and response variables. Similarly, the ability of deep learning algorithms to detect both time-based relations and hierarchy is another chance to create more efficient forecasting models.

Nevertheless, researchers in this field have been increasingly appreciating that although precise forecasts are desirable, it is also needed in the overall behavior of a model. That is, a model can be an ideal fit to the data in the past, but fail in the real world because of a variety of influences, such as transaction costs incurred by conducting trades, evolving market behavior, or overfitting to the sample on which it was trained. Thus, there is now interest among researchers in developing hybrid models that apply a statistical theory and machine learning methodologies. The major objective of such models is to balance three important qualities of good models, namely, interpretability, robustness, and forecasting the relevant outcomes accurately.

With the increasing complexity of financial markets due to the creation of more and more sophisticated data sets, future research undertakings will likely put much focus on: integrating multiple streams of data into their modeling activities, offering a better understanding of the mechanisms of the models, and quantifying the economic importance of the variations in the forecasted values compared to merely comparing the statistical performance of models.

References

[1] Patel J, et al. Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 2015, 42(4): 2162-2172.

[2] Ballings M, et al. Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 2015, 42(20): 7046-7056.

[3] Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning. Review of Financial Studies, 2020, 33(5): 2223-2273.

[4] Cakici N, et al. Machine learning goes global: Cross-sectional return predictability in international stock markets. Journal of Economic Dynamics and Control, 2023, 155: 104725.

[5] Hanauer M X, Kalsbach T. Machine learning and the cross-section of emerging market stock returns. Emerging Markets Review, 2023, 55: 101022.

[6] Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 2018, 270(2): 654-669.

[7] Bao W, Yue J, Rao Y. A deep learning framework for financial time series using stacked autoencoders and long short-term memory. PLoS ONE, 2017, 12(7): e0180944.

[8] Kanwal A, et al. BiCuDNNLSTM-1dCNN—A hybrid deep learning-based predictive model for stock price prediction.

Expert Systems with Applications, 2022, 202: 117123.
[9] Rezaei H, et al. Stock price prediction using deep learning and frequency decomposition. Expert Systems with Applications, 2021, 169: 114332.
[10] Murray S, Xia Y, Xiao Y. Charting by machines. Journal of Financial Economics, 2024, 153: 103791.
[11] Aït-Sahalia Y, et al. How and when are high-frequency stock returns predictable? Cambridge: National Bureau of Economic Research, 2022. (Working Paper No. 30366)
[12] Sirignano J, Cont R. Universal features of price formation in financial markets: Perspectives from deep learning. Quantitative Finance, 2019, 19(9): 1449-1459.
[13] Shah J, et al. A comprehensive review on hybrid deep learning approaches for stock prediction. Intelligent Systems with Applications, 2022, 16: 200111.
[14] Hu Z, Zhao Y, Khushi M. A survey of forex and stock price prediction using deep learning. Applied System Innovation, 2021, 4(1): 9.