# Literature Review: Machine Learning in Stock Predictions

Neal Li *

Toronto French School, Toronto, Canada

* Corresponding Author Email: nealnianli@gmail.com

**Abstract.** Machine learning has revolutionized the field of stock prediction by offering a wide range of models capable of handling complex patterns and making accurate forecasts. Machine learning models vary widely in their application, uses, and effectiveness, and stocks vary as well in terms of volatility within the stock and also between stocks of different industries and at different market conditions. As such, the selection of the proper algorithmic tool to aid an investor is often difficult. This literature review paper provides an overview of ten popular machine learning models over two problem types (prediction and classification), namely Linear Regression, XGBoost, LSTM, ARIMA, GARCH, Random Forest, Logistic Regression, Adaboost, GRU, and CNN. By providing an exploration of these ten machine learning models, this literature review offers valuable insights into their underlying principles, applications and uses, results strengths, and limitations. This paper equally, by consequent, facilitates informed decision-making and encourages further research in the field of machine learning.

**Keywords:** Machine Learning; Stock Prediction; Prediction; Classification; Neural Networks; Ensemble Learning; Boosting.

## 1. Introduction

The most fundamental question in the application of deep learning toward the prediction of stocks lies in its success rate. In order to be viable, deep learning techniques and models must not only be more successful than random chance but also be able to understand and forecast, in some way, the volatility and changes within the stock market. This can't dismissed as an easy task, as stocks are inherently irregular and volatile - subject to changes in public opinion and events in the world - while machines only have access to past data. This is equally by no means unimportant, as machines can, unlike humans, handle large data with high speed and efficiency, and assist with several factors such as analysis, reduced bias, and the spotting of patterns. In general, when looking at stock prediction problems, one can classify them into two major types: Prediction and Classification. Prediction problems aim to predict a certain value, in this case, the exact stock price that the stock will be at in, for example, and a week. Classification problems, on the other hand, aim to predict a binary state, such as whether the stock will go up or down. In order to evaluate the performance of machine learning models across both Prediction and Classification problems, this literature review paper will go through a spectrum of papers on previously researched models as well as their performance in stock predictions. The remainder of the paper will be composed of an overview of the various models, a discussion of the model performances, and a conclusion to the paper.

## 2. Organization of the Text

### 2.1. Prediction Problems

#### 2.1.1. Introduction

As previously indicated, Prediction problems in stocks aim to predict the exact stock price that a stock will rise or fall to within a certain time period. The prediction models that will be looked at are Linear Regression, XGBoost, LSTM, ARIMA, and GARCH, followed by an overall conclusion of their relative performances and suitability toward stock predictions.

### 2.1.2. Linear regression

A typical machine learning approach for forecasting a desired output variable based on input or independent factors is Linear Regression. It depicts the relationship between the variables used as input and output as a linear equation or a straight line. [1]. Due to how Linear Regression aims to create a line that best models the relationship between independent and dependent variables, which isn't always possible, proper model evaluation and validation are usually necessary, as well as statistical analysis on the accuracy of the results [2]. Nevertheless, although Linear Regression may be simplistic, it can still be applied in a number of domains, such as in that of stock predictions by capturing underlying patterns driving variables. For example, through past input variables like stock price, economic indicators, market indices, and trading volume, one can predict values such as trading volume with relative accuracy (Multiple R of approximately 0.6, R Squared of 3.58, and Adjusted R Squared of 3.47) [2]. As such, Linear Regression is viable in stock predictions due to strengths such as its simplicity, efficiency, and speed, as well as being a well-established technique and method. On the other hand, it also has several jarring limitations, such as its sensitivity to outliers, a lack of nonlinear transformations, and an assumption of a linear relationship, as well as a potential for underfitting, all of which can affect performance [1]. However, these problems aren't as present or impactful in terms of stock predictions.

### 2.1.3. XGBoost

Extreme Gradient Boosting (XGBoost), is a well-known ensemble learning machine learning technique. It is designed to improve the performance of predictive models and has increased accuracy and versatility when compared to other Gradient Boosting algorithms in various domains, such as that of stock predictions [3]. This is due in part to the fact that ensemble learning involves combining the predictions of several separate models to get a final prediction that is stronger and more precise. Furthermore, XGBoost is a model specifically designed to address several of the main limitations of standard gradient boosting. As such, it often sees applications in stock predictions, and produces desirable results, being capable of fitting the rise and fall of stocks and approximating and generalizing to an appropriate standard [4]. XGBoosts therefore can be used in the prediction of stocks due to advantages such as consistent accuracy, the ability to handle complex relationships and missing data, and built-in cross-validation. However, XGBoost is also complex and its parameters require time-consuming tuning. It may also not scale well to large datasets and encounter problems handling imbalanced datasets as well as struggle with non-linear patterns [5][3]. Regardless of its drawbacks, XGBoost is still a powerful tool in deep learning.

### 2.1.4. LSTM

Long Short-Term Memory (LSTM) is an advanced type of recurrent neural network (RNN) model. RNNs are designed to process data sequences, where each data point is associated with a specific time step. RNNs have connections that allow information to be passed between time steps, making the capture of temporal dependencies and patterns possible [6] [7]. LSTMs equally have gating mechanisms, allowing the disposal of unnecessary information as well as weighting and picking out of more important information contributing to correct predictions. Thus, LSTM is particularly suited to the prediction of stock prices. This means that it can frequently outperform the baseline in the forecast of prices [6] with some results returning RMSEs of less than 0.01 [7]. These results are possible because of the model's capability at capturing relationships between distant data points. LSTMs are also, by consequent, capable of recognizing more complicated patterns and trends than most other RNNs and maintaining a more stable gradient. However, LSTMs also have problems. Vanishing Gradient remains a problem in complex sequences and the model is computationally intensive to train with large datasets. In addition, it is vulnerable to overfitting and is heavily dependent on objective data in sufficient quantities [6]. Nevertheless, LSTMs continue to be of capital importance in sectors like stock predictions.

### 2.1.5. ARIMA

ARIMA (Autoregressive Integrated Moving Average) is a time-series forecasting model. Its main usage lies in the prediction of future values from past observations of a time-dependent variable. The model is designed to capture patterns and trends in time-series data, making it useful for tasks like stock price prediction [8] [9]. ARIMA can be applied by leveraging its ability to capture autocorrelation and trends. This has been clearly showcased in a number of studies, some of which found that the accuracy of ARIMA, when applied, exceeds 85% [8], with other more conservative results demonstrating ARIMA's potential as well as its existing ability to compete against other models [9]. This is, in part, accounted for by ARIMA's unique set of strengths and weaknesses. This is because ARIMA is rather simple to implement while being effective for stationary data and especially for short-term forecasting. ARIMA equally doesn't necessitate any external factors for predictions, as some models do. However, ARIMA still remains unreliable to the more complex non-linear patterns present in stocks as well as volatile and non-stationary data, while equally being sensitive to outliers [9]. Although these drawbacks are far from having no impact, ARIMA remains a viable tool for stock predictions.

### 2.1.6. GARCH

A statistical model known as GARCH (Generalized Autoregressive Conditional Heteroskedasticity) is used to assess and forecast volatility in time series data, particularly financial data like stock prices. It was developed specifically to account for the phenomenon of heteroskedasticity, which is the changing variance of a time series over time [10]. GARCH addresses limitations faced by traditional linear models by taking into account the fluctuation of volatility and modeling the conditional variance in the function of past observations and forecast errors. The model has already seen applications in stock predictions, which have yielded promising results [10]. It is equally frequently used in combination with ARIMA as a hybrid to enhance its performance, and is capable of returning more accurate predictions as a result [11]. GARCH models are useful tools, as they are able to capture time-varying volatility, an important feature for financial markets that can shift between periods of high and low volatility. Furthermore, it has an inbuilt flexibility and adaptability to other stocks and models. On the other hand, GARCH operates on a number of assumptions such as a continued historic volatility trend and an unwarranted focus on volatility while failing to account for potential discontinuities in data [10][11]. Despite these limitations, the GARCH models remain a valuable tool for forecasting.

### 2.1.7. Summary

The five prediction problem deep-learning models discussed in this section were Linear Regression, XGBoost, LSTM, ARIMA, and GARCH, all of which exhibit varying levels of performance. Linear Regression offers simplicity but struggles with capturing complex market behaviors. XGBoost, on the other hand, excels in capturing non-linear patterns and can adapt to changing market conditions, often yielding accurate results. LSTM is appropriate for use with stock data since it can capture long-term relationships and sequential patterns. ARIMA performs well in capturing time series patterns, but its effectiveness is limited when dealing with non-stationary data and more complex relationships. GARCH is advantageous in modeling volatility changes over time, enhancing risk assessment. All five of the listed models and algorithms are viable for use in stock predictions.

## 2.2. Classification Problems

### 2.2.1. Introduction

As previously indicated, Classification Problems in stocks aim to predict a binary state, such as whether the stock will go up or down The prediction models that will be looked at are Random Forest, Logistic Regression, AdaBoost, GRU, and CNN, followed by an overall conclusion of their relative performances and suitability towards stock predictions.

### 2.2.2. Random Forest

Random Forest is equally a type of ensemble learning method like XGBoost and Adaboost, which integrates the forecasts of several separate models to get a more precise and reliable final forecast. However, unlike the two models, Random Forest, uses decision trees as individual models. By combining the predictions of multiple decision trees and the purposeful introduction of random chance, Random Forest reduces overfitting and increases the overall accuracy and reliability of predictions as well as its ability to be generalized [12][13]. As such, Random Forest is viable for stock predictions and returns promising results. For example, previous applications have yielded prediction results between 80 and 85% in the long term [12], as well as results indicating that it can predict with reasonable accuracy but that the accuracy may vary depending on the quality and amount of input data [13]. This can be attributed to a number of advantages of the model. It is, notably, efficient and resistant to overfitting, and very well-suited for binary classification. On the other hand, it can be limited when being applied in non-linear interactions and its limited ability to capture complex relationships can impact predictive accuracy. Furthermore, Logistic Regression can struggle with imbalanced data. It is equally very sensitive to outliers which can disproportionately influence the model's coefficients and predictions [12]. Logistic Regression remains, however, a powerful tool in stock forecasts.

### 2.2.3. Logistic regression

Logistic Regression is a statistical model and a type of regression analysis used for binary classification tasks. Logistic Regression focuses on estimating the probability that an input belongs to one of two possible classes and is best suited for instances when the dependent variable has two possible outcomes and is binary. The algorithm uses a logistic function to represent the connection between the input and output variables, which in this case is the likelihood of belonging to a particular class. [14]. this means that Logistic Regression is viable for use in the prediction of stocks and has yielded positive results. For example, past studies and implementations of the algorithm have shown that Logistic Regression is reliable and capable of capturing patterns and relationships that contribute to accurate predictions [14][15]. Logistic Regression has a number of advantages such as its efficiency and low computational cost. It is equally resistant to overfitting, and very well-suited for binary classification. However, it can be limited when being applied to non-linear interactions and its limited ability to capture complex relationships can impact predictive accuracy. In addition, Logistic Regression can struggle with imbalanced data. It is equally very sensitive to outliers which can disproportionately influence the model's coefficients and predictions [15]. Despite these shortcomings, Logistic Regression is still very capable and often used in stock forecasts.

### 2.2.4. AdaBoost

A well-known machine learning ensemble approach for classification problems is called AdaBoost (Adaptive Boosting). By integrating the outputs of weak learning algorithms to produce a powerful overall predictor, it is intended to enhance their performance. AdaBoost is particularly effective in situations where individual models struggle to provide accurate predictions [16]. AdaBoost's strength in boosting the performance of weak learners is advantageous in capturing patterns within financial data, making it a strong option when it comes to algorithms for stock predictions. This is seen in a number of applications that demonstrate that AdaBoost can offer reasonable results, with strong performances in forecasting [16] and results above the baseline [17]. However, the effectiveness can vary by a large margin due to AdaBoost’s inherent traits. The algorithm is innately more accurate than traditional ensemble algorithms, and is capable of handling complex relationships within data as well as non-linearity. Furthermore, it’s reliable and able to be generalized without being prone to overfitting. On the other hand, it’s sensitive to outliers and noisy data while being limited by the performance of its weak performers. Adaboost is also susceptible to imbalanced data as well as overfitting as it runs too many iterations. Nevertheless, AdaBoost remains a powerful model for stock forecasting.

### 2.2.5. GRU

RNN architectures with gated recurrent units (GRUs) are utilized for sequential data processing. It is designed to capture long-range dependencies and patterns in sequences, making it suitable for tasks like stock forecasting. GRU was specifically designed as a way to address some of the limitations of traditional RNNs, such as the vanishing gradient problem [18] [19]. GRU has a more simple architecture than LSTM, and is equally as viable when it comes to being used for stock predictions. This is due to the fact that it is capable of yielding strong results [18] as well as outperforming other high-achieving models [19]. This is, in part, due to its fast training speed and efficiency at capturing short-term dependencies within sequences. They're also not prone to overfitting. However, they can struggle to capture complex patterns that involve long-range dependencies and they may not always be the most efficient or reliable depending on dataset characteristics [18]. Although these drawbacks are far from having no impact, GRU remains a viable tool for stock predictions

### 2.2.6. CNN

CNN (Convolutional Neural Network) is a deep learning model used to analyze and predict stock price movements based on historical data. Although CNNs are typically used for processing images and videos, they may also be utilized for sequential data, such as time series, which are frequently employed in stock market analysis. When applied to stock prediction, the model learns to automatically extract relevant patterns and features from the input data. CNNs are thus able to identify recurring sequences or behaviors that might be indicative of future stock price movements [20] [21]. Previous applications of CNN have demonstrated mixed, but overall rather positive results [20]. There are, however, challenges in using only CNN, hence the widespread prevalence of hybrids such as CNN and LSTM [21]. This is because CNN tends to excel at capturing local patterns such as trends or recurring sequences, and an automatic ability to learn relevant features and an ability to focus on the features without being distracted by outliers. At the same time, CNN is only able to focus on said local patterns, meaning that it can miss broader context. CNN may also fail to capture more intricate relationships and relies on the presence of a large amount of high-quality data [20]. Although CNN faces many challenges, it has potential and is still viable for the forecast of stocks.

### 2.2.7. Summary

The five classification problem deep-learning models discussed in this section were Random Forest, Logistic Regression, Adaboost, GRU, and CNN, all of which exhibit varying levels of performance. Random Forest excels in handling complex relationships, producing accurate results through ensemble learning, and handling outliers well. Logistic Regression, being efficient and resistant to overfitting, provides valuable insights into binary outcomes. Adaboost's ability to combine multiple weak models makes it proficient in classification tasks, offering improved accuracy. GRU, a variant of LSTM, is suitable for capturing temporal dependencies and is particularly effective when dealing with sequential stock data. CNN, commonly used in image analysis, can also be applied to time series data with its ability to identify patterns. All five of the listed models and algorithms are viable for use in stock forecasting.

## 3. Conclusion

In this literature review, ten machine learning models (Random Forest, Logistic Regression, Adaboost, GRU, CNN, Linear Regression, XGBoost, LSTM, ARIMA, and GARCH) have been explored. These algorithms span from traditional statistical approaches to sophisticated deep learning architectures, and each offers varying capabilities, opening the possibility to tackle a wide range of challenges. For example, as previously discussed in the paper, ensemble learning methods like XGBoost, Adaboost, and Random Forest can vary widely in terms of their applications. Although the mentioned ensemble learning methods can all be classified within the category, the models appear in different sectors and excel in varying applications (high performance, minimizing errors, and

robustness and versatility respectively). The listed models all have high potential for growth in the future, as they continue to develop and increase their capabilities in terms of pattern recognition, data analysis, and optimization. Machine learning algorithms are thus valuable assets and tools for investors that are only likely to grow in value. However, one should also not be too reliant on these models, as they hold disadvantages as well. For example, models are frequently incapable of accounting for volatility or sudden recessions from real-world events. In addition, the presence of a "lag" between new financial data and an updated prediction is inevitable, as well as occasional false positives which can lead to financial losses if not properly managed. As the field of machine learning continues to advance, it is crucial to stay abreast of new developments and leverage the strengths of different models for improved performance and generalization. Through the use of models and algorithms, one can make more accurate predictions and thereby achieve better results in the stock markets.

## References

[1] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, et al. Introduction to Linear Regression Analysis, 2021, 6: 0-561

[2] Farhad Soleimanian Gharehchopogh, Tahmineh Haddadi Bonab, Seyyed Reza Khaze, et al. A Linear Regression Approach To Prediction of Stock Market Trading Volume: A Case Study, IJMVSC, 2013, 4: 1-7

[3] Tianqi Chen, Carlos Guestrin, et al. XGBoost: A Scalable Tree Boosting System, 2016, 1-10

[4] Tian Liwei, Feng Li, Sun Yu, Guo Yuankai, et al. Forecast of LSTM-XGBoost in Stock Price Based on Bayesian Optimization, Intelligent Automation & Soft Computing, 2021, 1-14

[5] Kyung Keun Yun, Sang Won Yoon, Daehan Won, et al. Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process, Expert Systems with Applications, 186: 1-19

[6] David M. Q. Nelson, Adriano C. M. Pereira, Renato A. de Oliveira, et al. Stock market's price movement prediction with LSTM neural networks, 2017, 1-8

[7] Murtaza Roondiwala, Harshal Patel, Shraddha Varma, et al. Predicting Stock Prices Using LSTM, International Journal of Science and Research (IJSR), 2017, 1-3

[8] Prapanna Mondal, Labani Shit, Saptarsi Goswami, et al. Study of Effectiveness of Time Series Modeling (ARIMA) in Forecasting Stock Prices, IJCSEA, 4: 1-17

[9] Ayodele A. Adebiyi., Aderemi O. Adewumi, Charles K. Ayo, et al. Stock Price Prediction Using the ARIMA Model, 2014, 1-7

[10] Basel M. A. Awartani, Valentina Corradi, et al. Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries, International Journal of Forecasting, 2005, 21, 1-17

[11] Narendra Babu Chindanur, Eswara Reddy B, et al. Selected Indian stock predictions using a hybrid ARIMA-GARCH model, 2014, 1-6.

[12] Luckyson Khaidem, Snehanshu Saha, Sudeepa Roy Dey, et al. Predicting the direction of stock market prices using random forest, Applied Mathematical Finance, 2016, 1-20

[13] Isaac Kofi Nti, Adebayo Felix Adekoya, Benjamin Asubam Weyori, et al. Random Forest Based Feature Selection of Macroeconomic Variables for Stock Market Prediction, American Journal of Applied Sciences, 2019, 1-13.

[14] Syed Shahan Ali, Muhammad Mubeen, Adnan Hussain, et al. Prediction of stock performance by using logistic regression model: evidence from Pakistan Stock Exchange (PSX), 2018, 15-31.

[15] Jibing Gong, Shengtao Sun, et al. A New Approach of Stock Price Trend Prediction Based on Logistic Regression Mode, 2009, 1-6.

[16] Victor Chang, Taiyu Li, Zhiyang Zeng, et al. Towards an improved Adaboost Algorithmic Method for Computational Financial Analysis, Journal of Parallel and Distributed Computing, 2019, 134: 1-28.

[17] Fayaz Tunio, Ding Yi, Agha Amad Nabi, Agha Kinza, et al. Financial Distress Prediction Using Adaboost and Bagging in Pakistan Stock Exchange, Journal of Asian Finance, Economics and Busines 2020, 8: 1-9

[18] Chengyu Li, Guoqi Qian, et al. Stock Price Prediction Using a Frequency Decomposition Based GRU Transformer Neural Network, Methods and Applications of Data Mining in Business Domains, 2022, 1-18

[19] Ming-Che Lee, et al. Research on the Feasibility of Applying GRU and Attention Mechanism Combined with Technical Indicators in Stock Trading Strategies, Advances in Artificial Intelligence: Machine Learning, Data Mining and Data Sciences, 2022, 1-19

[20] Ehsan Hoseinzade, Saman Haratizadeh, et al. CNNpred: CNN-based stock market prediction using a diverse set of variables, Expert Systems with Applications, 2019, 129: 1-37

[21] Sidra Mehtab, Jaydip Sen, et al. Stock Price Prediction Using CNN and LSTM-Based Deep Learning Models, 2020, 1-7