Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/48/20241621

Predicting the S&P 500 stock market with machine learning models

Boyu Shi[1,4], Chuhua Tan[2], Yue Yu[3]

[1]FedUni Information Engineering Institute, Hebei University of Science and Technology, Shijiazhuang, China
[2]Shanghai Starriver Bilingual School, Shanghai, China
[3]Mathematics, University of Liverpool, Liverpool, UK

[4]bshi@students.federation.edu.au

Abstract. Predicting stock prices through machine learning models is becoming increasingly important in today's world. This essay aims to present a basic analysis of predicting the S&P 500 stock market using machine learning models, that are linear regression and XGBoost. The research utilizes historical data from the S&P 500 stock market in the past few years, dividing it into training, testing, and verification sets with a distribution of 70%, 15% and 15%, respectively. The linear regression model uses MSE (3.051×10^17) and R^2(0.316) in testing group to analyze the accuracy of prediction, while the XGBoost model uses MSE (14816.886) and RMSE (121.725) in testing group to analyze the accuracy. The findings indicate that the accuracy of the XGBoost model surpasses that of the linear regression model. Drawing from the outcomes, one can infer that the XGBoost model is better suited for stock prediction compared to the linear regression model.

Keywords: SP500, Prediction, XGBoost.

1. Introduction and research background

Predicting stock market tendency could bring significant benefits for people which enabled them to either potentially gain their wealth or avoid substantial losses under the result that our models produce. Nowadays exhibiting an extremely wide range of possibilities, machine learning has become one of the hottest topics globally, an increasing number of studies use machine learning models to investigate. So far, machine learning models have been utilized in the various industries due to their ability to process vast datasets and identify intricate patterns that may perform much better than human analysis. For example, there are studies focusing on the application of machine learning including thermal comfort [1], machine fault diagnosis [2], decision support in the insurance sector [3], drug discovery and development [4], solid-state materials science [5] etc. The investigation also takes into consideration foundational aspects, themes, and research clusters pertaining to the focal point of artificial intelligence and machine learning in finance [6]. Additionally, notable attention is given to studies on machine learning techniques employed in financial market prediction [7], as well as decision-making in financial trading utilizing machine learning approaches and portfolio selection [8]. These contributions are recognized for their excellence and significance.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/48/20241621

However, the comparison studies on different machine learning models applied to stock market still lacking. Therefore, we started our study on the comparison of the performances in predicting between two prominent machine learning algorithms: linear regression and XGBoost. The S&P 500, a typical referential index, reflects the overall condition of the US stock market, making it a prime focus for prediction models, so we our comparison was based on the data we collected from the S&P 500.

In this study, we collected the data we needed from the S&P 500, and correspondingly made our prediction with linear regression and XGBoost. As a result, the performance of XGBoost Model is better than Linear Regression Model.

2. Data

In this study, the author collected historical price of the Standard & Poor's 500 (S&P 500) from October 17, 2016, to October 13, 2023, from the Yahoo Finance website. The dataset consists of daily closing prices of the stock market, which serve as the dependent variable for the model. Using Python, the author created an appropriate dataset based on real-life economic performance. Basic information is shown below in Table 1.

Table 1. Descriptive statistics of closing prices

Mean	3356.146
Sample Variance	575971.3
Median	3145.615
Minimum	2085.18
Maximum	4796.56

In Figure 1, we can observe the period from early 2016 to late 2017, during which the S&P 500 index showed a steady upward trend. In early 2018, the index further increased, reaching a historical high, but by the end of that year, the index experienced a decline in prices. Subsequently, from 2019 to early 2020, the S&P 500 index maintained its upward momentum. However, at the beginning of 2020, there was a significant downward plunge, reaching a low point. Over the next year, the index exhibited a rapid upward trend until early 2022, when a period of oscillating decline occurred, reaching a low point in November 2022. Since then, the index has shown a gradual upward trend until the present day.

Figure 1. The s&p500 closing price curve from October 17, 2016, to December 17, 2023

DOI: 10.54254/2755-2721/48/20241621

Given the aim of comparing the prediction accuracy of various machine learning models for S&P 500 price trends, the authors partition the data into three categories: 70% for training, 15% for testing, and 15% for verification. This partitioning and construction of machine learning models aligns with typical and sound methodologies. Such separation allows the authors to train the models based on historical data, demonstrate their performance on unseen data, and make predictions to assess the effectiveness of the models. This partitioning and construction of machine learning models aligns with typical and sound methodologies. Such separation allows the authors to train the models based on historical data, demonstrate their performance on unseen data, and make predictions to assess the effectiveness of the models.

3. Methods

3.1. Linear regression

Linear regression is a foundational machine learning technique employed for predictive modeling purposes. The equation of the model is as follows.

$$Y_i = b_1 * x_1 + b_2 * x_2 + \cdots + b_i * x_i + e_i, i = 1,2, \ldots n \tag{1}$$

This model uses MSE and R^2 to evaluate the accuracy of the model, where,

$$SS_{resid} = (Y - Xb)^T (Y - Xb), \tag{2}$$

$$SS_{total} = (Y - Xb_{total})^T (Y - Xb_{total}), \tag{3}$$

$$MSE = \frac{SS_{resid}}{n - p - 1} \tag{4}$$

$$R^2 = 1 - \frac{SSresid}{SStotal} \tag{5}$$

In the above equations, n is amount of data, p+1 is amount of parameter. MSE represents the fluctuation of error, so the smaller the MSE, the smaller the fluctuation. R^2 represents the fitness of model, so this analysis hopes the R^2 can close to 1, which means the fit is quite good.

To improve model performance, we conducted feature engineering, selecting the most influential economic and financial indicators. The model underwent training on the training set and was subsequently evaluated on both the testing and verification sets to assess its capability in predicting the stock market.

3.2. XGBoost

XGBoost, an ensemble learning algorithm, was employed as the second model in our study. This algorithm is known for its ability to handle complex relationships and non-linearity in the data [9-10]. It builds a decision tree ensemble model to capture interactions between variables.

Much like the linear regression model, feature selection was carried out to identify the most relevant predictors for the XGBoost model. Subsequently, the model underwent training and evaluation using the training, testing, and verification sets. For a dataset containing n samples of m dimensions, the XGBoost model can be represented as:

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in F(i = 1,2, \ldots n) \tag{6}$$

In this equation,

$$F = \{f(x) = w_{q(x)}\}(q: R^m \to \{1,2, \ldots T\}, w \in R^T) \tag{7}$$

DOI: 10.54254/2755-2721/48/20241621

The XGBoost model, characterized by a collection of CART decision tree structures represented by q, w, and T, seeks to find the optimal parameters by minimizing the objective function, which combines the error (L) and model complexity terms (Ω). The objective function can be expressed in the following manner:

$$Obj = L + \Omega \tag{8}$$

$$L = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{9}$$

$$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{10}$$

During the optimization training of the model using the training data, it is imperative to maintain the integrity of the original model while incorporating a novel function f into the model structure.

4. Results

The empirical results reveal that both linear regression and XGBoost models display promising predictive capabilities in forecasting the S&P 500 stock market.

4.1. Linear regression

This study uses prediction graphs (See Figure 2 and Figure 3) and indicators of MSE and R^2 to demonstrate the accuracy of the model (See Table 2). Firstly, the analysis uses prediction graphs to show the accuracy of predictions.

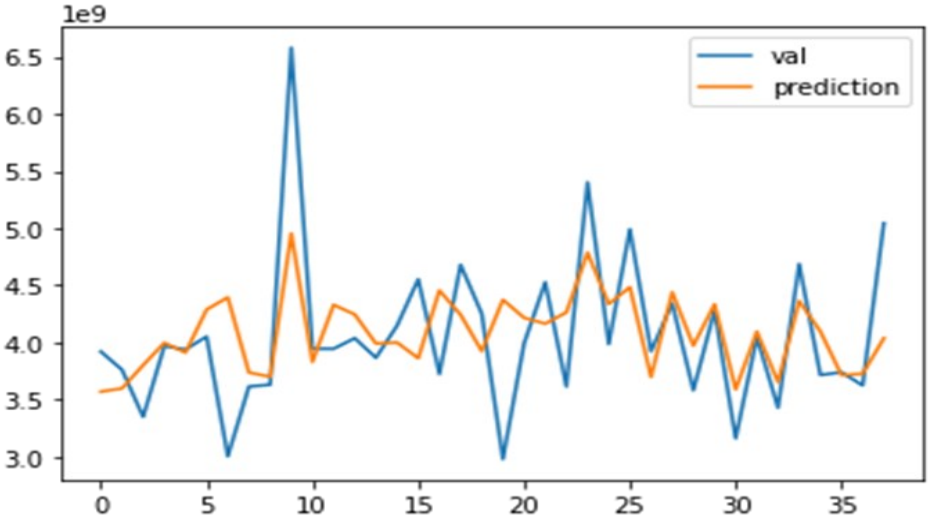

Figure 2. The prediction graph of Test Group

DOI: 10.54254/2755-2721/48/20241621

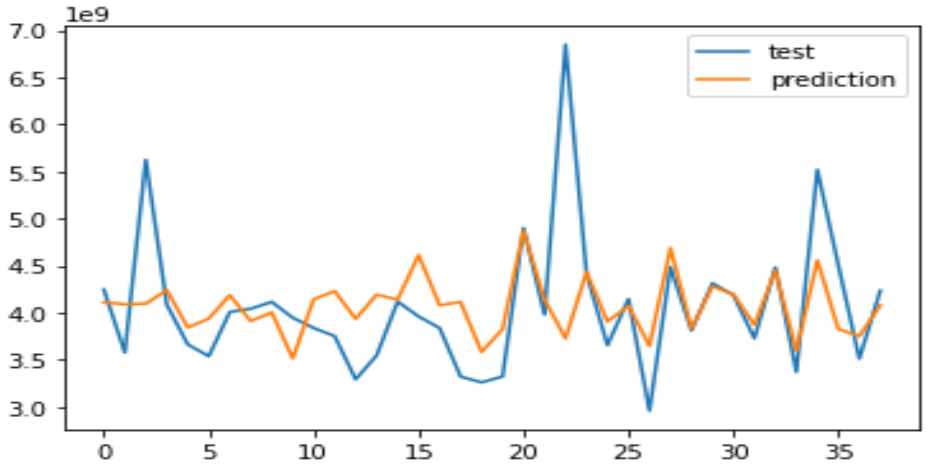

Figure 3. Prediction graph of Predict Group

Table 2. Indicators

	Test Group	Predict Group
MSE	3.051*10^17	4.634*10^17
R^2	0.316	0.106

By calculating the values of MSE and R^2, this analysis finds that the value of MSE is large and the value of R^2 is not close to 0, indicating that the error of modeling this dataset by linear model is large.

The linear regression model demonstrated a moderate ability to predict stock market movements. It captured some of the overall trends and exhibited good performance on the training set. However, its predictive power decreased on the testing and verification sets, suggesting that it might not generalize well to new data.

4.2. XGBoost

In this study, we use XGBoost to predict the price trend of S&P 500 and generate a graph to compare with its historical price and calculate MSE and R^2 (See Figure 4 and Table 3).

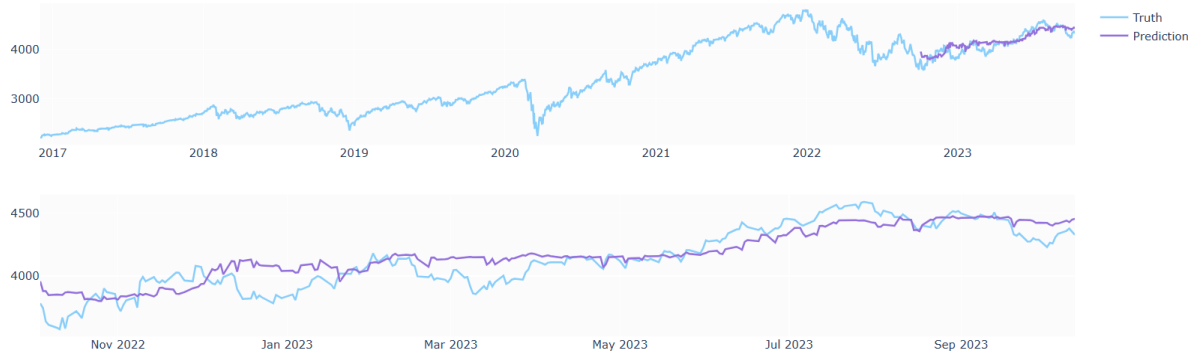

Figure 4. XGBoost model price prediction curve and actual s&p500 price curve

DOI: 10.54254/2755-2721/48/20241621

Table 3. MSE and R^2 of XGBoost model

	Test Group
MSE	14816.886
R^2	0.756

The XGBoost model outperformed linear regression, showing greater accuracy in predicting stock market movements. It exhibited robust generalization capabilities, achieving consistent performance across all three datasets. It is noteworthy that its performance is better than Linear Regression Model partly because of its complexity.

5. Conclusions and recommendations

In conclusion, the purpose of this essay is to provide a fundamental study of the S&P 500 stock market prediction process utilizing XGBoost and linear regression machine learning models. The study employs historical data from the S&P 500 stock market over the past few years, segmented into training, testing, and verification sets with a distribution of 70%, 15%, and 15%, respectively. The XGBoost model utilizes MSE (14816.886) and RMSE (121.725) in the testing group to measure accuracy, while the linear regression model uses MSE (3.051×10^17) and R^2 (0.316) in the testing group. According to the results, the XGBoost model demonstrates superior accuracy compared to the linear regression model.

Author contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Fard, Z. Q., Zomorodian, Z. S., & Korsavi, S. S. (2022). Application of machine learning in thermal comfort studies: A review of methods, performance and challenges. Energy and Buildings, 256, 111771.

[2] Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing, 138, 106587.

[3] Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. International Journal of Information Management Data Insights, 1(2), 100012.

[4] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., et al. (2019). Applications of machine learning in drug discovery and development. Nature reviews Drug discovery, 18(6), 463-477.

[5] Schmidt, J., Marques, M. R., Botti, S., & Marques, M. A. (2019). Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials, 5(1), 83.

[6] Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. Journal of Behavioral and Experimental Finance, 32, 100577.

[7] Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. Expert Systems with Applications, 124, 226-251.

[8] Paiva, F. D., Cardoso, R. T. N., Hanaoka, G. P., & Duarte, W. M. (2019). Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. Expert Systems with Applications, 115, 635-655.

[9] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.

DOI: 10.54254/2755-2721/48/20241621

[10] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).