Stock Price Prediction Method Based on XGboost Algorithm

Yifan Zhang(✉)

Lanzhou University of Technology, Qi Lihe, Lan Zhou, China
1091720256@qq.com

Abstract. In this paper, I use the XGboost algorithm model with parameters controlled by the Grid SearchCV search algorithm to forecast stock prices based on the daily time series characteristics of stocks. The results of the model demonstrate that by using a computer algorithm to analyze stock price information with a large amount of data, the model is able to capture the high-frequency time series fluctuation trend of the stock more accurately under control of overfitting and underfitting, and thus can obtain more accurate prediction results about the stock price.

Keywords: XGboost · Stock Price

1 Introduction

Since the 1960s, the introduction of the CAPM model has provided a clear understanding of the relationship between assets and returns, and then various factor models began to be introduced into the practice of asset pricing, the most famous of which is the three-factor model [3]. The core idea of this model is to identify the characteristic factors that can explain the asset returns by looking at the data characteristics. Nowadays, with the booming of financial markets, the characteristic factors that could explain asset returns in the past often no longer work, and because of the proliferation of data, simple financial models can no longer meet the requirements of the current big data era, so we need to find new factors that can explain asset returns, and use new algorithms that meet the requirements of the big data era to predict the price of assets and explain returns. We need to find new methods and explanatory perspectives for asset price prediction and return interpretation using new algorithms that meet the requirements of the Big Data era.

In recent years, the development and maturity of machine learning algorithms have provided new perspectives and methods to analyze and solve financial and economic problems, and the use of computer algorithms to solve financial and economic problems has gradually become a hot topic in academia. In this paper, we will use XGboost algorithm to simulate and predict stock prices, aiming to provide a new method for analyzing and predicting stock prices. eXtreme Gradient Boosting (XGBoost) is an algorithm based on GBDT. The basic idea of XGBoost is the same as GBDT, but with

D. Qiu et al. (Eds.): ICBBEM 2022, AHIS 5, pp. 595–603, 2023.
https://doi.org/10.2991/978-94-6463-030-5_60

some optimizations, such as second-order derivatives to make the loss function more accurate. The regular term avoids tree overfitting; block storage allows parallel computation, etc. XGBoost is efficient, flexible, and lightweight, and is widely used in data mining, recommendation systems, etc. [1] formally proposed the XGboost algorithm. The basic idea of XGBoost is the same as GBDT, but it makes many optimizations. With the continuous development of algorithm technology, the use of computer algorithms to solve economic management problems has now become a mature technical tool. This is exactly why this paper chooses to use the XGboost algorithm for stock price prediction. The sections of the article are organized as follows, Sect. 1 introduces the model, Sect. 2 presents the reasons for the selection of data parameters, Sect. 3 is the model test, and Sect. 4 is the conclusion.

2 Models and Parameters

2.1 Model Introduction

XGboost uses a presentation of the second-order Taylor formula and adds a regular term, which is a control of the tree complexity and prevents overfitting. For handling missing values, the algorithm tries to decide a default direction for handling missing values by enumerating whether it is better for all missing values to go into the left subtree at the current node, or into the right subtree. The most time-consuming part of a single weak learner is the splitting process of the decision tree, which can be selected in parallel using multiple threads for splitting points with different features. XGboost will order the samples in advance according to the size of the features before parallel processing, and by default they are placed in the right subtree, and then recursively take out one sample from small to large and put it into the left subtree, and then calculate the gain based on the then calculate the size of the gain based on the splitting point, and then record and update the largest gain splitting point.

The core algorithmic idea of XGBoost is to.

1. keep adding trees i.e. keep learning a new function f(x) to fit the residuals of the last prediction.
2. Training is completed to get k trees, according to the characteristics of the sample, in each tree will fall to the corresponding a leaf node, each leaf node corresponds to a score.
3. Finally, the scores corresponding to each tree are added up to the predicted value of the sample. Please remember that all the papers must be in English and without orthographic errors.

The XGBoost objective function is defined as:

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + con$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$ which is the complexity of the tree.

The objective function consists of two parts, the first part is used to measure the difference between the predicted score and the true score, and the other part is the regularization term. The regularization term also contains two parts, T denotes the number of leaf nodes and w denotes the fraction of leaf nodes. γ controls the number of leaf nodes and λ controls the fraction of leaf nodes from being too large to prevent overfitting. The newly generated tree is intended to fit the residuals of the last prediction, i.e., when t trees are generated, the predicted fraction can be written as:

$$\begin{aligned}
&\hat{y}_i^{(0)} = 0 \\
&\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
&\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\cdots\cdots \\
&\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}$$

After the parameters are substituted and Taylor expansion is performed, the objective function becomes:

$$Obj^{(t)} = \sum_{i=1}^{n} [g_i w_q(x_i) + \frac{1}{2} h_i \omega_q^2(x_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_{qj}^2$$

Forecasting indicators using the XGboost algorithm has gradually become a mature technical tool, but there is still much room for improvement in using machine learning algorithms to predict financial asset prices. The reason for this situation is that the market used to generally believe that there was a tight economic logic behind financial asset prices and that analytical data alone could not overcome rational markets. It is more important for investors to have theoretical analysis of financial assets from an economic perspective and to judge the laws behind their operation in order to gain returns.

However, in recent years, with the development of behavioral finance, people found that the market is not as rational as the traditional doctrine describes, most investors have irrational investment behavior, using the rational market hypothesis to analyze the behavior of asset prices instead of gaining returns. In this case, using machine learning algorithms to analyze financial big data for prediction can get better results [4]. The reason for this is that the irrational behavior of investors in the market and the lack of professional financial knowledge of most people lead to the fact that most people's investment behavior is to follow the up and down trends rather than to analyze the economic logic. The use of computer algorithms to analyze trends will be more scientific and accurate than the judgment of ordinary people. Therefore, the use of algorithms for price forecasting of financial assets is not simply analyzing data, but has a scientific logic behind it that is supported by behavioral finance that is in vogue today.

The use of time series features to predict stock prices is an improvement on the theoretical analysis taken in the past to predict stock prices. The refined calculation of the algorithm and the processing of big data information allow the time series trend characteristics of stock prices to be further revealed. For daily stock prices with short transition period and unstable volatility, the previous rough analysis can no longer meet the needs of the times if forecasting is to be achieved. Time series features can strongly capture the ups and downs waves of stock prices. The idea of using the XGboost algorithm to predict stocks is to use the time series function of stocks as a feature factor,

to continuously learn the stock's decision on time series fluctuations, and to eventually form a function that can predict stock prices through machine learning and optimization.

2.2 Parameter Selection

2.2.1 Algorithm Parameters

When using the XGboost algorithm, we need to set the parameter seeking parameters, according to the characteristics of the stock price, the parameters set in this paper are shown in Table 1. The meanings of the parameters are: n_estimators: the number of weak learners; max_depth: the depth of the tree; min_child_weight: the weight threshold of the smallest node, less than this value, the node will not split again; gamma: the node splitting brings loss minimum threshold, we use the difference of the objective function to calculate the gain, less than this threshold, the node will not split again; learning_rate: control the weight reduction coefficient of each weak learner; this coefficient will be multiplied by the weight value of the leaf nodes, it is used to weaken the influence of each tree. If the learning rate is small, the number of corresponding weak learners should be increased.

2.2.2 Predictive Index Selection

Before performing the model prediction, we need to perform the selection of data indicators to constitute the characteristic factors of the algorithm so. The data parameters

Table 1. Parameter seek optimization settings (Table credit: Original)

n_estimators	[100, 200, 300, 400]
learning_rate	[0.001, 0.005, 0.01, 0.05]
max_depth	[8, 10, 12, 15]
Gamma	[0.001, 0.005, 0.01, 0.02]
random_state	[42, 43, 44, 45]

Table 2. Predictive index (Table credit: Original)

EMA_9	Stock price index weighted moving average 9-day data
SMA_5	Stock short-term average moving line 5-day data
SMA_15	Short-term average stock price moving line 15-day data
SMA-30	Short-term average stock price moving line 30-day data
RSI	SMA(MAX(Close-LastClose,0),N,1)/SMA(ABS(Close-LastClose),N,1)*100
MACD	EMA12-EMA26
MACD SIGNAL	Moving average of ewm index weights for MACD with a span of 9 days

used in this paper are the daily opening price, closing price, high price, low price and the total number of daily stock indicators of the stock. These data are the key data to form the feature factors, and the composition of the feature factors will be labeled in Table 2.

In this paper, these indicators are selected as the characteristic factors for stock price prediction for the following reasons. Compared to the multi-factor model, which is based on the explanation of asset returns, the model is constructed by selecting all the company fundamentals and other characteristic factors, and the price forecast tends to be more short term, which has higher requirements on the time series trend of the data. The use of time series data results in better approximation of the model when using the XGboost model to predict stock prices [7].

In addition, the relative strength indicator RSI is a technical curve based on the ratio of the sum of up points and down points in a certain period of time. RSI indicator is effective for stock price prediction based on investment strategies [2].

It reflects the level of market sentiment over a certain period of time. It was first developed by Welles Wilder. It was first applied to futures trading by Wells Wilde, but later it was found that among the many technical analysis of charts, the theory and practice of the Strength and Weakness Indicator is extremely suitable for short-term investment in the stock market, so it was used in the measurement and analysis of stock ups and downs. The analysis indicator is designed to reflect the strength of the price trend with three lines. This graph can provide investors with a basis for operation and is ideal for short term spread operations.

3 Model Training Process

Figure 1 illustrates the daily closing prices of a set of stocks and the decomposition of the daily closing of the stocks by taking the decompose function, resulting in a trend, seasonality and residual plot of the stock prices.

We can see that the trend of this group of stock prices first rises and then tends to fall, which is consistent with the rules of time series, and the stock prices have obvious

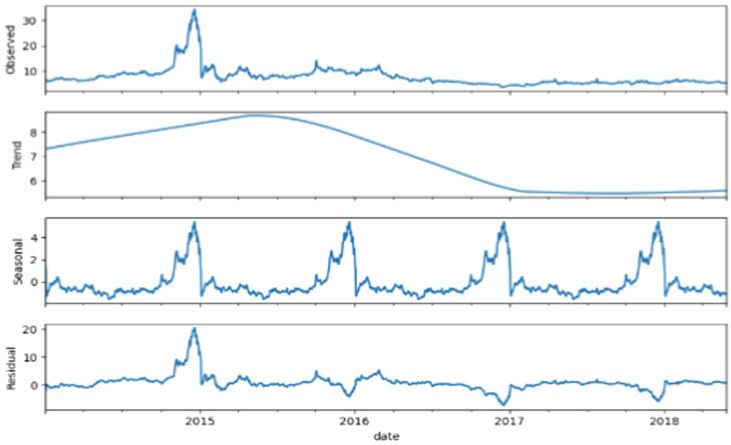

Fig. 1. Price Decompose (photo credit: Original)

seasonality and the residuals tend to be smooth. Therefore, this set of stock data is suitable to be used for price prediction model training.

In the second section on parameters, we give the range of parameters required for the XGboost algorithm decision tree. In order to avoid underfitting or overfitting due to artificially specified parameters, we need to tune the parameters. Grid SearchCV is an efficient way to handle the task of tuning parameters. grid search as a tuning tool is characterized by exhaustive search: it tries every possible parameter among all the candidates, and the best performing parameter is the final result. The principle is like finding the maximum value in an array. The main disadvantage of this method is that it is more time consuming, but it guarantees the accuracy and validity of the results. It will give an evaluation based on the search results of different parameters and finally give an optimal result according to the parameter range that fits the model. Grid Search can effectively improve the accuracy of the model and optimize the model performance [5].

More importantly, the optimization algorithm not only adjusts the parameters of the decision tree required by the XGboost algorithm, but it also analyzes and evaluates the values of the function features needed for XGboost regression, i.e., a series of time series data used to represent stock price features in this paper, and gives the importance of each indicator to the model so that we can have an important role in the data on which features can better help us make stock price predictions.

The results of the parameters of the search algorithm are given in Table 3, while the importance analysis of the metrics in Table 2 is given in Fig. 2, which is all the result of Grid SearchCV.

For the stock price time period given in Fig. 1, the model is set to have the first 70% as the training set, the middle 15% as the validation set, and the second 15% as the test set. The predicted results with importance display plots can be derived based on the merit-seeking parameters, which are shown in Fig. 2 and Fig. 3, respectively.

As shown in Fig. 2 and Fig. 3, the XGboost algorithm with time series features has a high accuracy and significance in predicting stock prices, and the RSI indicator has a high importance in predicting stock prices as in the evaluation of indicators, which justifies the use of the RSI indicator in predicting daily stock prices. Of course significantly visible, the results cannot avoid some overfitting, probably due to the following reasons.

Table 3. Parameter Optimization Results (Table credit: Original)

Best params:	'gamma': 0.01, 'learning_rate': 0.01, 'max_depth': 8, 'n_estimators': 400, 'random_state': 42]
Best validation score	−0.40314337083083407
y_true	[5.89 5.82 5.79 5.69 5.79]
y_pred	[5.132111 5.1830683 5.134022 5.134022 5.079282]
mean_squared_error	0.19183294968417305

Fig. 2. Predicted results (photo credit: Original)

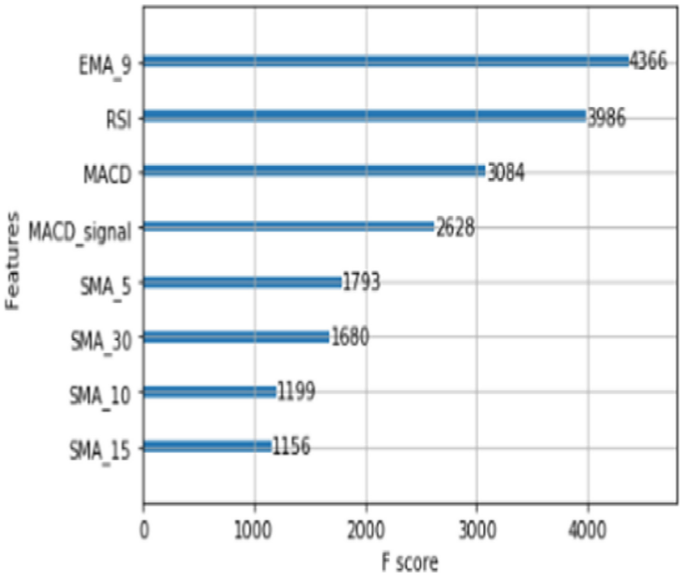

Fig. 3. Feature Importance (photo credit: Original)

1. using time series features to predict stock prices is a trend-focused approach, and doing so cannot avoid economic fluctuations and unexpected financial events behind the stock, such as operational changes in a listed company.
2. The range of parameters delineated is still narrow and a broader range should be considered.
3. The selected stocks are episodic in nature, and more stock data are needed for validation and training so that the model can be more optimized.

It is undeniable that the XGboost algorithm has significant advantages and prospects for predicting stock prices.

4 Conclusions

How to predict the price of financial assets such as stocks has always been an important issue in economic management [6]. With the emergence of constantly developing theories and technologies, this problem has been solved in different ways at different times. In the current era of big data, where the amount of information is exploding, more and more computer algorithms are used to solve problems arising from economic management activities. In this paper, the XGboost algorithm is used to forecast stock prices, and the feature factors are chosen to represent the time series feature data of high-frequency operation trends, while the parameter search method is used to ensure that the XGboost model parameters are more accurate and avoid over-fitting and under-fitting phenomena. The results of the model operations prove that using time series trends to predict daily stock prices with high-frequency characteristics has high accuracy, and using the XGboost algorithm to do forecasting can ensure fine and accurate processing of a large amount of data, with better success in trend capture than previous theoretical analysis.

The model results in this paper further prove that using algorithms to deal with economic management problems is a scientific and effective tool, and that using algorithms to deal with problems will be an important trend in future economic management. We have reasons to believe that in the future, more and more advanced algorithms will be used to predict the prices of financial assets such as stocks and reveal more and more scientific economic features behind these financial assets, so that people can better understand the composition of asset prices, better understand the laws of the market, and can have a better complementary perspective to traditional economic theories or create new doctrinal theories. Also for market investors, more advanced technology can be more effective in ensuring that they can achieve higher returns on their investments.

References

1. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp.785–794.
2. Chong, T. T. L., & Ng, W. K. (2008). Technical analysis and the London stock exchange: testing the MACD and RSI rules using the FT30. Applied Economics Letters, 15(14), 1111–1114.
3. Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of financial economics, 33(1), 3–56.
4. Gu, Shihao, Kelly, Bryan & Xiu, Dacheng, (2020). Empirical Asset Pricing via Machine Learning. The Review of financial studies, 33(5), pp.2223–2273.
5. Huang, Q., Mao, J., & Liu, Y. (2012). An improved grid search algorithm of SVR parameters optimization. In 2012 IEEE 14th International Conference on Communication Technology (pp. 1022–1026). IEEE.
6. Liu, Jianan, Stambaugh, Robert F & Yuan, Yu, (2019). Size and value in China. Journal of financial economics, 134(1), pp.48–69.
7. Wang, Y., & Guo, Y. (2020). Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. China Communications, 17(3), 205–221.