Creative Economy
2025, VOL. 9, NO. 1, 62-67
DOI: 10.47297/wspceWSP2516-251913.20250901

Research on stock prediction method based on random forest

Yixin Zhong

College of Mathematics and Information Technology, South China Agricultural University, Guangzhou, Guangdong, 510640

ABSTRACT
The complexity and nonlinearity of stock prices have made the study of stock price prediction techniques a significant area of concern in the finance industry. This paper approaches the stock price prediction issue as a categorization challenge, considering the stock's opening, minimum, maximum, trading volume, and daily fluctuation amplitude as key factors. The stock's closing price on the same day is forecasted using logistic regression, support vector machine, decision tree, and random forest algorithms, followed by a comparative analysis of the predictive outcomes from these models. The lattice search is then used to optimize the random forest model. According to the experimental results, the random forest model outperforms the other methods in terms of stock price prediction ability. Following grid search optimization, the model's stability and prediction accuracy are further enhanced.

KEYWORDS
Random forest; Lattice search; Stock price prediction

1 Background

The stock market is among the many traditional businesses that are continuously altering their working methods and procedures in an attempt to advance with the aid of Internet technology as a result of its ongoing development. One of the most crucial components of a nation's financial system is its stock market, which also serves as a gauge of the growth and decline of the national economy. In addition to helping regular investors make more logical choices, stock forecasting can help government regulators make wise selections for policy development, preventing and resolving significant financial hazards[1]. The study of stock forecasting techniques is therefore extremely important from a practical standpoint.

The two primary categories of early stock price prediction techniques are linear and nonlinear prediction models. The predictions of linear prediction models are not optimal because of the features of stock data, including time-series, multifactoriality, and nonlinearity. The traditional linear prediction models in stock prediction are gradually being replaced by machine learning algorithms, such as support vector machines, decision trees, random forests, etc., in financial research due to the quick development of artificial intelligence and big data technologies. Among them, studies on the random forest algorithm have been used in a few marketplaces, and researchers both domestically and internationally have produced some findings.In order to forecast stock prices, HIND DAORI and other researchers selected the support vector machine, random forest, and straight line algorithms. They then confirmed that the random forest approach had the best accuracy[2]. To reduce the danger of stock market investing, Luckyson Khaidem and other academics used the random forest algorithm to forecast stock returns.The model was trained using a stochastic oscillator, relative strength index, and other metrics, which increased the model's accuracy[3]. It causes the model's accuracy to rise. Wang Huiying and colleagues proposed an enhanced random forest model for stock price trend prediction, utilizing the grid search algorithm to optimize the model's parameters, thereby improving its predictive accuracy and generalization capability[4].

In conclusion, the random forest approach performs better in stock prediction than the conventional linear model and a few other machine learning algorithms. Its accuracy and stability can be increased by refining and optimizing it[5]. In this paper, the closing price of the stock for the day is predicted based on the fundamental features of the stock data. The data is trained using the random forest model, logistic regression algorithm model, support vector machine algorithm model, and decision tree model, respectively, and their performance in predicting stock prices is compared. A stock prediction model based on random forest and grid search is then proposed, and the random forest algorithm is optimized based on the results to determine the best parameters and analyze the changes of each evaluation index.

2 Algorithm Introduction

2.1 Logistic regression algorithm

Logistic regression, a statistical learning method for classification, transforms a linear model into probability values, modeling the connection between input and output data to develop a probabilistic model[6]. It is frequently utilized in binary classification issues. The logistic regression algorithm, which learns the model's parameters through Maximum

Likelihood Estimation (MLE), is implemented in this study using Python's LogisticRegression function. Following the normalization of input characteristics in the training phase, linear regression is commonly employed to represent the correlation between these features and the target variable through a logistic function, later transformed into probability figures using the Sigmoid method.

Using the parameters discovered during training, the logistic regression process generates predictions after the model has been trained: For the binary classification task, the logistic regression algorithm maps the probability value to one of the two categories based on a set threshold. Assuming a threshold of 0.5, if the probability is greater than 0.5, the classification is category 1, otherwise category 0;

In the task of categorizing multiple categories, the logistic regression algorithm employs either a one-to-many approach or a Softmax function to determine the likelihood of each category, choosing the one with the greatest probability as the predictive outcome.

2.2 Support Vector Machine Algorithm

In order to find the best solution among all possible classification schemes, Support Vector Machine (SVM), a supervised learning algorithm, calculates the hyperplanes of various categories. It then uses the optimal classification hyperplane to split the training samples so that the individual separated categories have the maximum spacing, minimizing the classification error rate and increasing the separation's degree of confidence[7]. SVMs can be divided into three categories based on the various relationships between the samples: linearly indivisible, linearly separable, and linear SVMs. The support vector machine algorithm is employed in this work to analyze classifications. Due to the non-linear characteristics of stock data, the kernel function facilitates the conversion of this data from its initial space into a more complex dimensional realm. This makes the data capable of linear differentiation in a space of higher dimensions, enabling the identification of a hyperplane that can be differentiated linearly. By optimizing the objective function during training, the Support Vector Machine algorithm makes sure that the classification boundary is accurate while maximizing the distance between the two classes of samples. The SVM uses the derived decision boundary to classify the new samples at the end of training.

2.3 Decision Tree Algorithm

One nonparametric supervised learning approach that can be used for both regression and classification problems is the decision tree. With internal nodes representing the dataset's features, edges representing the features' potential values, and leaf nodes representing the categories, a decision tree, which resembles a tree structure in a flowchart, builds a tree with nodes, edges, leaf nodes, etc. to describe the relationship between one or more features and their subsets. Starting from the root node, a decision tree technique is employed in this study to predict stock data[8]. Decisions are made by evaluating the value of a feature's attribute and dividing the dataset so that, to the greatest extent feasible, the samples in the split sub-dataset fall into the same category. Data confusion steadily diminishes and exhibits regularity as the decision tree level rises, allowing for the construction of classification and prediction models.

2.4 Random Forest Algorithm

Random Forest, employing a voting system for classification and regression forecasting, is a unified learning framework that addresses classification and regression issues through the amalgamation of various decision trees, each trained on a distinct subset. In every decision tree, the count of divided nodes is selected at random, dependent on the quantity of sample attributes[9].The random forest algorithm is used in this study to predict stock prices. The algorithm takes samples from the training dataset and divides them into multiple sample subsets. It then trains a decision tree for each sample subset, chooses k features from all the features at each split node, chooses the best split features from them, and uses a voting mechanism to choose the category with the most occurrences as the final classification result. Due to the integration of many decision trees, the random forest algorithm typically performs better in terms of prediction accuracy than the separate decision tree model. It also has a stronger capacity for generalization and is less susceptible to the overfitting issue.

3 Model Building

In this study, the historical data of American Airlines from February 8, 2013 to February 7, 2018, such as opening price, closing price, high price, low price, and volume, were selected as the dataset, and the data were split into the training set and test set according to the ratio of 3.5:1, and the problem was transformed into a classification task, and the opening

price, the highest price, the lowest price, volume, and the magnitude of the daily price fluctuation were selected as the features, and the The closing price as the target variable is divided into three intervals according to the quartiles, and then the logistic regression algorithm, the support vector machine algorithm, the decision tree algorithm and the random forest algorithm are used to train the data respectively, and compare their prediction results, and then optimize the random forest model to further improve the model, the specific steps are as follows:

(1) gathering information on stock history and pre-processing it, which includes handling outliers and looking for missing values; (2) To extract feature variables and target variables for the training and test sets, respectively, use feature engineering and target variable specification; (3) Use grid search and cross-validation to define the random forest, logistic regression, support vector machine, decision tree, and optimized random forest models; (4) Use each of these models to train the training set and then predict the test set; (5) To assess the models' efficacy, compute the performance metrics—such as accuracy, recall, etc.—derived from each model independently for comparison;

Figure 1 Flow chart of the experiment

4 Model evaluation

All the models in this study transform the problem into a classification task to be processed, and in classification problems, the metrics for model evaluation are generally accuracy, F1 Score value, recall, AUC value, etc., and in this paper, we choose accuracy, F1Score value, precision, and recall as the evaluation metrics.

(1) Accuracy rate

Indicates the proportion of correctly predicted samples to the total samples of the model, which is a finger reflecting the overall correctness, and the formula is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

In this case, TP (True Positive) indicates how many samples possessed the positive class correctly predicted to be

positive, TN (True Negative) shows how many samples had the negative category correctly predicted to be negative, FP (False Positive) demonstrates how many samples had the negative class mistakenly anticipated to be positive, and FN (False Negative) indicates how many samples had the positive class erroneously predicted to be the negative class.

(2) Precision rate

Precision rate reflects the ratio of samples forecasted to be in the positive category to those genuinely in the positive category, showcasing the accuracy of the positive category prediction, using the subsequent formula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

(3) Recall

The recall rate reflects the ratio of genuinely positive samples to accurately forecasted positive classes, showcasing the model's capacity to identify most positive class samples using a specific formula:

$$\text{Recall} = \frac{TP}{TP + FN}$$

(4) F1 Score

F1 Score is the reconciled average of precision and recall, summarizing the effect of the trade-off between these two metrics in a single value, reflecting the balance between precision and recall, and focusing on the model's ability to make comprehensive predictions for positive class samples, with the following formula:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The evaluation metrics for each model obtained after training are shown in Table 1.

Table 1 Evaluation indicators for each model.

	Accuracy rate	Precision rate	Recall	F1 Score
Logistic regression	0.9495	0.9639	0.9495	0.9529
Support vector machine	0.9639	0.9698	0.9639	0.9654
Decision Tree	0.9675	0.9680	0.9675	0.9677
Random forest	0.9711	0.9720	0.9711	0.9715

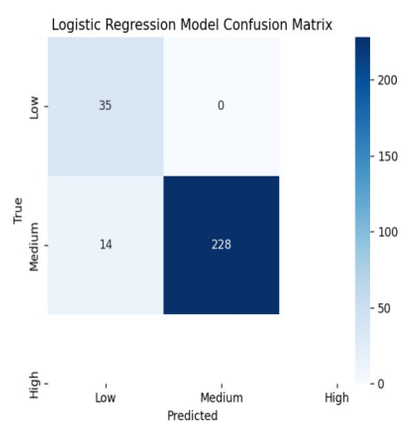

Figure 2 Logistic regression model confusion matrix

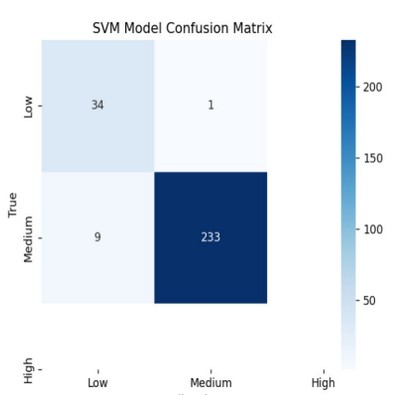

Figure 3 SVM model confusion matrix

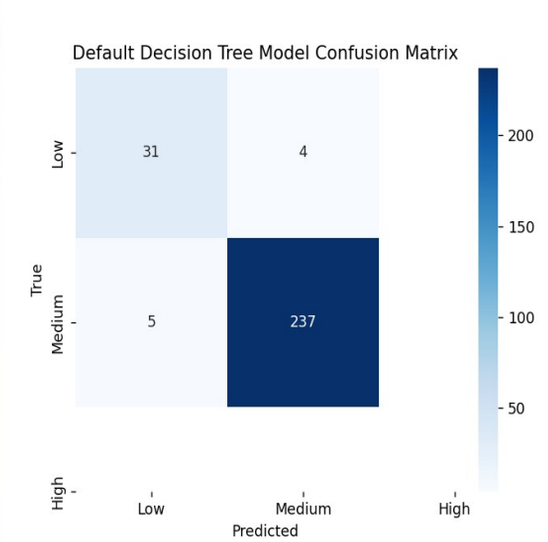

Figure 4 Default decision tree model confusion matrix

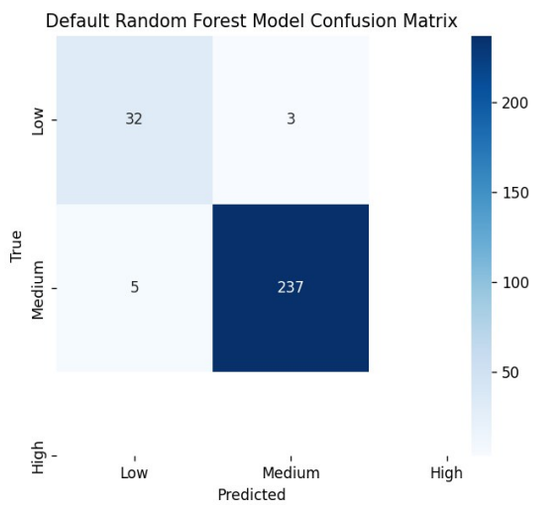

Figure 5 Default random forest model confusion matrix

According to Table 1, comprehensively comparing each evaluation index of the above four models, the logistic regression model performs relatively weakly, and as a linear model has limited performance in complex data scenarios or nonlinear relationships, in contrast, the support vector machine model and the decision tree model perform better, while the random forest model performs optimally in terms of accuracy, precision, recall, and F1 Score, and by integrating multiple decision trees, reducing the risk of overfitting in a single model, and demonstrating strong comprehensive performance.

The evaluation index is set to be the average of the scores of the 3-fold cross-testing set, and the optimal parameters chosen are used to finally construct the optimized random forest model. The basic random forest model uses the grid search to optimize five parameters: the number of decision trees, the maximum depth of decision trees, the minimum number of samples required for splitting nodes, the minimum number of samples required for leaf nodes, and whether to use bootstrap sampling.

The optimal parameter combinations derived from the grid search are: n_estimators=100, max_depth=5, min_samples_split=20, min_samples_leaf=1, and bootstrap=True. After determining the optimal parameters, keep the other parameters the same as those of the default Random Forest model, and compare the The results of each evaluation index of the two models are shown in Table 2.

Table 2 Evaluation indicators for two random forest models.

	Accuracy rate	Precision rate	Recall	F1 Score
Optimizing the Random Forest	0.9783	0.9783	0.9783	0.9783
Random Forest	0.9711	0.9720	0.9711	0.9715

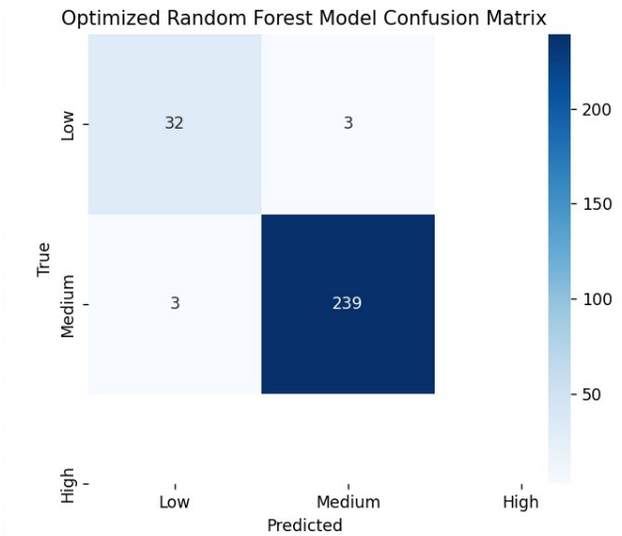

Figure 6 Optimized random forest model confusion matrix

The parameter-optimized random forest model has improved across all evaluation indexes, indicating that the grid search's optimal parameters make the random forest model better suited to the stock data's characteristics. By setting the optimal values of each parameter in a reasonable manner, the model's computational efficiency is optimized, unnecessary branches are reduced, the model's performance and computation costs are effectively balanced, and the risk of overfitting is also decreased, resulting in a higher generalization performance on the test data.

5 Conclusion

The research investigates the effectiveness of the random forest model in forecasting stock prices, in contrast to logistic regression, support vector machines, and decision trees. Subsequently, it proposes a refined random forest model for predicting stock prices, utilizing the lattice search algorithm to boost its predictive precision.

According to the experimental data, Random Forest outperforms other models in predicting stock prices. The model's total performance is further enhanced by the optimized grid search algorithm, which gives it greater stability and accuracy as well as a stronger capacity for generalization.The experiment offers a more dependable model option for the stock price prediction problem in addition to demonstrating the value of random forests in the realm of financial forecasting. The random forest model can be applied to more complicated situations in the future by combining it with additional models or algorithms to better improve it in terms of feature construction, indicator selection, etc.

References

[1] Gao Ziyi. (2021). Research on Stock Price Trend Prediction Based on Random Forest. Master's Thesis of China University of Political Science and Law.

[2] Daori H, Alharthi M, Alanazi A, et al. (2022). Predicting Stock Prices Using the Random Forest Classifier. https://www. researchgate. net/publication/365386418_Predicting_Stock_Prices_Using_the_Random_Forest_Classifier

[3] Khaidem L, Saha S, Dey S R. (2016). Predicting the direction of stock market prices using random forest. https://arxiv.org/abs/1605.00003

[4] Wang H, Hao Y. (2021) Stock Price Trend Prediction Algorithm Based on Technical Index and Random Forest. Modern Computer, 27(27):43-47,52.

[5] Zheng J, Xin D, Cheng Q, et al. (2024). The Random Forest Model for Analyzing and Forecasting the US Stock Market in the Context of Smart Finance. https://arxiv.org/abs/2402.17194

[6] Zhang H. (2022) Scenario Analysis for Linear and Logistic Regression. Automation & Instrumentation, 10:1-4,8.

[7] Zhang L. (2020). Stock Market Trend Analysis and Forecasting Based on Support Vector Machine. Master's Thesis of Nanjing University of Posts and Telecommunications.

[8] Liu X, Cui W. (2023) A decision tree-based classifier for building quantitative strategies. Journal fur die Reine and Angewandte Mathematik, 39 (3): 339-349.

[9] AI Yu. (2020). Research on CSI 300 Index Trend Prediction Based on Random Forest Optimization. Master's Thesis of Shandong University.