Dean&Francis
ISSN 2959-409X

MS003532

# Stock Prediction Analysis Based on Logistic Regression and Random Forest

**Yuang Liu**

Shanghai Experimental School, Shanghai, 200126, China

*Corresponding author: 1559522603@qq.com

**Abstract:**

An important factor in a nation's economic growth is the stock market. Accurate stock price prediction is vital for investors, financial analysts, and investment institutions. However, the volatility of stock prices makes prediction challenging. Adopting advanced analytical methods can improve prediction accuracy. This paper investigates the application of machine learning (ML) algorithms, such as logistic regression (LR) and random forest (RF) in predicting stock price trends, exploring their value in stock forecasting. Ten daily features, such as opening price, closing price, and trading volume, were utilized to train models. By exploring the impact of the number of historical days on prediction accuracy, the study found that the LR model achieved its highest accuracy (53.8%) when utilizing data from the previous 4 days. Further comparison among models revealed that the RF model outperformed both LR and BP models, achieving the highest prediction accuracy (58.3%) and AUC (0.576). These findings suggest that RF holds promise as a reliable tool for stock price trend forecasting, providing valuable insights for investors seeking more accurate market prediction tools.

**Keywords:** Logistic Regression; Random Forest; Stock Prediction

## 1. Introduction

The application of different analytical techniques to anticipate future patterns in stock prices is known as the „stock price prediction problem.“, including upward or downward movements. This issue is of great significance in the financial sector, as accurate predictions can yield substantial economic benefits for investors. However, stock price forecasting faces numerous challenges.

The fact that a wide range of intricate aspects impact the stock market is one of the main challenges, such as macroeconomic indicators, market sentiment, political events, and corporate performance. These factors are often difficult to quantify and incorporate into models. Additionally, the stock market exhibits high nonlinearity and dynamic characteristics, with inherent noise and randomness that make it challenging for predictive models to capture all relevant information. Moreover, imbalanced data and the non-stationarity of time series data further complicate the task of building a stable and accurate stock

ISSN 2959-409X

prediction model.

Currently, commonly used stock prediction methods can be categorized into fundamental analysis, time series models, and machine learning techniques [1-3]. Time series prediction typically employs traditional models such as ARIMA and GARCH. Machine learning approaches primarily use methods like logistic regression (LR), random forest (RF), and Artificial Neural Network (ANN) for classification-based predictions.

In Usha's research, a binary logistic regression model achieved a 71.2% accuracy rate when categorizing companies as "good" or "poor," demonstrating the advantages of machine learning (ML) for stock performance prediction [4]. Sheth applied three ML methods to stock price prediction, with the results underscoring the potential of ML in this domain [5]. In Christian's study, the likelihood that equities would outperform the market was predicted using an RF model, focusing on technical indicators [6]. With a Sharpe ratio of 2.49 for equal-weighted portfolios and 1.56 for value-weighted portfolios, the model demonstrated outstanding performance, surpassing benchmarks [6].

This paper applies multiple ML methods, including LR and RF to analyze the stock data of China Merchants Bank from January 2016 to October 2024. The aim is to develop a high-performing prediction model, further analyze the future trends of China Merchants Bank (CMB) stock, and provide more accurate insights for investment decision-making.

## 2. Data Source and Processing

This study focuses on China Merchants Bank (CMB) as a case study to predict stock price trends using logistic regression and random forest models. Historical stock data for CMB was obtained via the stock_zh_a_hist API provided by AKShare. Opening price, closing price, highest price, lowest price, trading volume, transaction amount, amplitude, percentage change, price change, and turnover rate are among the statistics in the dataset.

The training set consisted of stock price data from 2016 to 2023, and the test set consisted of data from 2024. A binary variable was used to describe the closing price trend: if the closing price on a given day was higher than the closing price on the day before, the stock price was labeled as bullish, with y=1. On the other hand, the stock price was labeled as bearish with y=0 if the closing price was less than or equal to that of the day before.

The training set (2016–2023) contains data for 2,144 trading days, while the test set consists of data for 199 trading days up to October 31, 2024. Fig. 1 illustrates the stock price trends of China Merchants Bank since 2016.


**Fig. 1 The stock price trend of China Merchants Bank from January 2016 to October 2024 (Picture credit: Original).**

## 3. Fundamental Principles of Model Construction

### 3.1 Logistic Regression

In statistics and machine learning, LR is a popular predictive analysis technique that is mostly utilized for binary classification issues. It forecasts the likelihood that a specific input instance falls into a given category. LR builds upon linear regression by generating a prediction value and then mapping this value to the range [0,1] using the sigmoid function, thereby transforming it into a probability.

The Maximum Likelihood Estimation (MLE) approach is commonly used to estimate the parameters of the LR model. MLE finds the values of the parameters that maximize the observed data's probability (likelihood function). In LR, the likelihood function is the product of the joint probabilities of all observed data. For binary classification

problems, Finding the parameter values that maximize the log-likelihood function is the aim of parameter estimation. Finding the parameter values that maximize the log-likelihood function is the aim of parameter estimation. Usually, numerical optimization techniques like the gradient ascent method are used to do this.

### 3.2 Random Forest

RF is an ensemble learning approach that builds numerous decision trees to do regression or classification. To increase overall forecast accuracy and resilience, it aggregates the predictions of several models. Because Random Forest lowers the chance of overfitting and can identify intricate patterns in the data, it can produce predictions that are more accurate than those made by a single decision tree. In regression problems, RF computes the average of all the predictions made by the decision trees; in classification issues, it chooses the most frequent class as the final prediction for every sample.

Below are the steps for building a Random Forest model:

1. Sample Extraction: From the initial training dataset, randomly select several subsets, each of which should be the same size as the original dataset. This procedure makes it possible to sample repeatedly.
2. Feature Selection: Choose the optimal splitting point from a collection of characteristics for each decision tree at each split by selecting them at random.
3. Decision Tree Construction: Create decision trees using the feature subsets and sample sets that were retrieved. Without pruning, each tree is given the opportunity to reach its maximum potential.
4. Aggregating Results: RF votes to aggregate all decision tree predictions for classification issues, and the class with the highest votes becomes the final prediction. It determines the average of each decision tree‘s predictions for regression problems.

### 3.3 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models that mimic the structure and function of the human brain, consisting of numerous nodes (or neurons) connected to one another. These nodes are typically arranged in layers, including the input layer, hidden layers, and output layer.

1. Input Layer: Receives input feature data from the external environment.
2. Hidden Layers: Processes the data from the input layer; there can be multiple hidden layers.
3. Output Layer: Generates the network‘s ultimate output, or the outcome of the prediction.

Every neuron uses an activation function to generate an output after receiving input from the layer above. The strength of signal propagation in the network is determined by the weights assigned to the connections between neurons. The ANN reduces the discrepancy between expected and actual output by modifying these weights throughout the learning process. The gradient descent method and the backpropagation algorithm are commonly used to accomplish this operation. The following phase is involved in a neural network model‘s training process:

1. Forward Propagation: The network transmits the input data forward. Every neuron in every layer uses the activation function to produce an output after calculating the total of the weighted inputs.
2. Loss Function Calculation: Mean Squared Error (MSE) or Cross-Entropy Loss is commonly used to compute the difference between the expected and actual outputs.
3. Backpropagation: To update each weight, the gradients determined by the loss function are sent backward through the network.
4. Weight Update: To reduce the loss function‘s value, the weights are updated via the gradient descent technique.

## 4. Model Prediction Performance Evaluation

### 4.1 Evaluation Methods

Common evaluation criteria for classification problems include the ROC curve, accuracy, recall, F1 score, and AUC value. Accuracy, ROC curve, and AUC value are the evaluation measures used in this work.

Confusion Matrix: When dealing with classification issues, this tool is crucial. It is a 2x2 matrix that shows how well a classification model performs. It has the four components listed below:

1. True Positives (TP): The quantity of cases that the model accurately identified as belonging to the positive class.
2. True Negatives (TN): The quantity of cases that the model accurately identified as belonging to the negative class.
3. False Positives (FP): The proportion of cases that are genuinely negative but were mispredicted as the positive class.
4. False Negatives (FN): The proportion of cases that are truly positive but were mispredicted as the negative class.

Accuracy: The percentage of accurately predicted samples among all samples is known as accuracy. It is computed as Accuracy=(TP+TN)/(TP+TN+FP+FN). Accuracy provides a general indicator of the model‘s performance.

The percentage of accurately anticipated positive results among all actual positive samples is known as recall. It

ISSN 2959-409X

is computed as follows: Recall=TP/(TP+FN). Recall is sensitive to the percentage of missed positive samples and focuses on the model's capacity to anticipate positive samples.

A graphical method for visualizing the model's performance at all potential classification thresholds is the Receiver Operating Characteristic Curve or ROC curve. The True Positive Rate (TPR), or recall, is plotted against the False Positive Rate (FPR). The FPR is computed as follows: FPR=FP/(FP+TN) ROC. Better model performance is shown by a greater area under the ROC curve (AUC).

The area under the ROC curve, or AUC, is a single figure that can be used to assess a model's overall performance. AUC values fall between 0.5 (no capacity to discriminate) and 1 (perfect discrimination). Because it is independent of the classification threshold selection, AUC is a significant statistic.

## 4.2 Model Prediction Results

Using the acquired stock data, we built a logistic regression model for this investigation. Ten daily aspects, including opening price, closing price, highest price, lowest price, trading volume, trading amount, price fluctuation, price change, rate of change, and turnover rate, are included in the data. We initially examine the impact of the number of historical days on stock price prediction because past price knowledge influences stock prices. In particular, it generates a dataset of 10*t features by using the 10 features from the previous t days (i.e., opening price, closing price, etc.). This dataset is then utilized to construct the logistic regression model. With a gain represented by 1 and a decrease by 0, this model forecasts whether the stock will rise or fall on day t+1. We set t to 1, 2, 3, 4, and 5, which means that we forecast the price change for the next day using data from the 1, 2, 3, 4, and 5 days prior, respectively. The LR model's prediction results for various values of t are shown in Table 1.

**Table 1. Logistic Regression Prediction Results**

| t | Accuracy | AUC |
|---|---|---|
| 1 | 0.515 | 0.5 |
| 2 | 0.512 | 0.5 |
| 3 | 0.510 | 0.501 |
| 4 | 0.538 | 0.527 |
| 5 | 0.525 | 0.513 |

It is evident from Table 1 that the logistic regression model's overall prediction accuracy is somewhat low. The LR model's prediction accuracy falls between 0.51 and 0.515 whether t=1, 2, or 3. When t=4, the accuracy increases to 0.538, but when t=5, the accuracy drops back to 0.525. The number of days of data significantly affects the prediction results. The model achieves its highest accuracy when t=4.

This study examines how well various machine learning models anticipate changes in stock prices. In addition to the logistic regression model, two commonly used ML classification models, neural networks and random forests, are selected. A multilayer perceptron neural network with two hidden layers of 100 and 16 neurons each was selected.

Based on the previous research, we use the same 10 features from the previous 4 days, combining them into 40 features, to forecast the fifth day's movement in the stock price. The Table 2 displays the predictions made by the three classification models.

**Table 2. Prediction Results of Three Classification Models**

| Model | Accuracy | AUC |
|---|---|---|
| LR | 0.538 | 0.527 |
| BP | 0.528 | 0.516 |
| RF | 0.583 | 0.576 |

The RF model has the highest overall prediction accuracy, reaching 0.583, while the neural network performs the worst, with an accuracy of 1% lower than the LR model.

The three models' ROC curves and matching AUC scores are displayed in the accompanying figure. It is evident from the graphic that the ROC curves of the LR and neural network models are nearly entirely covered by the ROC curve of the random forest model. In comparison to

the other two models, the RF model‘s matching AUC value is the highest, coming in at 0.576 (Fig. 2).

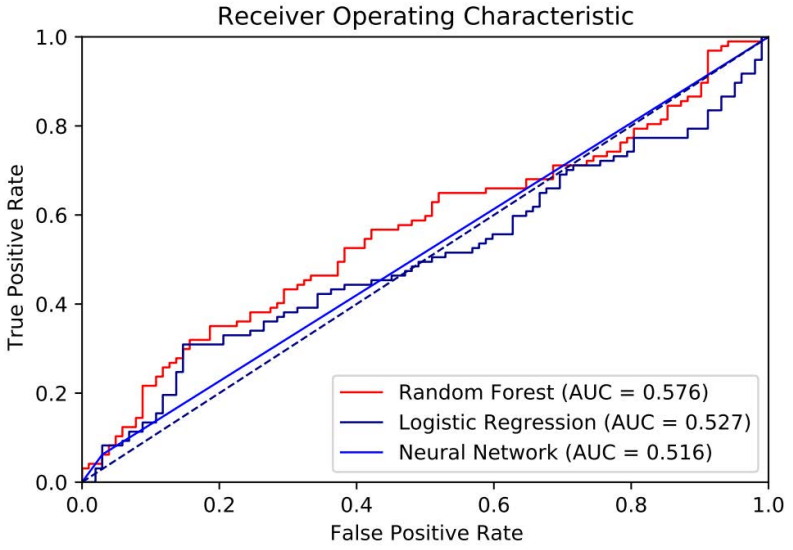


**Fig. 2 ROC Curves of the Three Models (Picture credit: Original).**

## 5. Conclusion

This paper focuses on predicting the stock trends of China Merchants Bank, a bank with strong financial management capabilities, issuance ability, and product development, and conducts an in-depth study of the application of ML algorithms in Forecasting the trend of stock prices. Based on the accuracy of the three ML models—LR, RF, and neural networks—random forest yields the highest accuracy. Therefore, when predicting the stock price of China Merchants Bank, it is recommended that investors use the random forest model for predictions, and combine it with logistic regression for further judgment, providing valuable reference and guidance for investors, financial practitioners, and researchers.

In the future, integrating statistical charts, data analysis, and stock market knowledge may help extract technical indicators that affect stock price trend prediction, thereby raising the precision of stock price trend prediction models.

## References

[1] Wafi A S, Hassan H, Mabrouk A. Fundamental Analysis Models in Financial Markets-Review Study. IISES Economics and Finance Conference; IISES Economics and Finance Conference.2015.

[2] Corizzo R, Rosen J. Stock market prediction with time series data and news headlines: a stacking ensemble approach. Journal of intelligent information systems, 2024, 62(1):27-56.

[3] Zhu Z, Wang W. Stock Type Prediction Based on Multiple Machine Learning Methods. Journal of Intelligent Learning Systems and Applications, 2024, 16(3): 20.

[4] Ananthakumar U, Sarkar R. Application of Logistic Regression in Assessing Stock Performances. IEEE International Conference on Pervasive Intelligence and Computing. 2017.

[5] Sheth D, Shah M. Predicting stock market using machine learning: the best and accurate way to know future stock prices. International journal of systems assurance engineering and management, 2023.

[6] Breitung C. Automated stock picking using random forests. Journal of Empirical Finance, 2023, 72: 532-556.