Proceedings of ICFTBA 2024 Workshop: Finance in the Age of Environmental Risks and Sustainability
DOI: 10.54254/2754-1169/94/2024OX0211

Machine Learning Applications in Stock Price Prediction

Tianyao Shen[1,a,*]

[1]*School of Economics & Management, Beijing Forestry University, Beijing, 100083, China*
a. sty1021@bjfu.edu.cn
**corresponding author*

Abstract: With economic fluctuations, investors focus on the stock market where scientific investment strategies become increasingly crucial. This article utilizes linear regression to predict stock price trends, achieving an R^2 value close to 1, indicating high accuracy in the model's predictions. Additionally, logistic regression and random forest classification models were employed to predict stock price movements, revealing a superior performance with the random forest model achieving an AUC value of 0.64. By employing a combination of regression and classification algorithms, this study offers a diverse perspective on stock price prediction, highlighting the importance of scientific forecasting in financial markets. The findings demonstrate that integrating multiple predictive models enhances the accuracy and reliability of stock price predictions, providing investors with precise market information to facilitate rational investment decisions. Furthermore, this research contributes valuable insights for future studies in stock market prediction, promoting continuous innovation and development in the theoretical and methodological aspects of this field. The results underscore the potential of advanced machine learning techniques in financial forecasting, offering a robust framework for understanding market dynamics and improving investment strategies. This study serves as a reference point for both academic research and practical applications, driving forward the evolution of predictive modeling in finance.

Keywords: Stock Price Prediction, Linear Regression, Machine Learning.

1. Introduction

As technology rapidly evolves, machine learning has gained widespread adoption in finance, particularly in stock price evaluation, demonstrating substantial potential for insightful analysis. Taking a joint-stock limited company as an example, this article delves into the application of machine learning in stock price valuation analysis, focusing on two mainstream methods: linear regression and random forest [1].

Stock price valuation is one of the core issues in financial analysis, closely related to investors' investment decisions and the market value of companies. However, stock prices are influenced by various factors such as the macroeconomic environment, company performance, and market sentiment, making stock price valuation complex and difficult to predict accurately. Therefore, utilizing machine learning techniques to explore the relationship between stock prices and influencing factors has become a subject worthy of research [2].

DOI: 10.54254/2754-1169/94/2024OX0211

In the following sections, this article will elucidate the specific applications of these two methods, and conduct empirical analysis using stock data from Apple Inc. to validate the effectiveness and accuracy of these algorithms.

2. Literature Revie

With the intensification of economic fluctuations, investors are gradually turning to the dynamic stock market, where predicting stock price movements becomes crucial. Quantitative investment has rapidly advanced with the integration of AI and financial derivatives, with machine learning widely applied in studying stock trends. Wang et al. employed Logistic regression to forecast stock price changes, demonstrating significant predictive effectiveness for the stock price trends of Guiyang Bank through confusion matrix and AUC evaluation [3]. Ma et al. focused on the agriculture, forestry, animal husbandry, and fishery industries, using the Random Forest and XGBoost algorithms to predict stock trends with high accuracy [4]. Zhang et al. developed a popular stock analysis and recommendation system based on linear regression [5].

This paper concentrates on the classification and regression methods in machine learning, delving into their application in predicting stock price trends. By model construction and effect evaluation, the aim is to elucidate algorithm principles and provide a scientific basis for investment decisions in the stock market. While not delving into the full evolutionary process of stock valuation in the literature review, the aforementioned studies have adequately showcased the latest advancements in the combination of quantitative investment and machine learning in the field of stock prediction.

3. Data Source and Data Processing

The data used in this study is sourced from the Snowball official website, covering the period from June 2019 to June 2024. In the data preprocessing stage, Apple, NVIDIA, and Microsoft were selected for analysis, offering multiple advantages [6]. Being global tech giants, the market position and financial performance of these three companies have always been closely monitored, resulting in relatively high quality and integrity of the related data [7]. This implies that when analyzing stock price valuation, it can obtain more accurate and comprehensive data, thereby enhancing the reliability of the analysis. The data mainly consist of seven attributes including date, opening price, closing price, high price, low price, adjusted closing price, and trading volume, as shown in Table 1 below.

Table 1: Data description.

Attribute	Data type	Describe
Date	Object	Data
Open	Float64	Opening price
High	Float64	The highest
Close	Float64	Closing price
Low	Float64	Lowest price
Volume	Int64	Turnover

The dataset was enriched by introducing a new column, 'close_next_day', which indicates the closing price for the subsequent day relative to the current entry in the 'close' column. This supplementary attribute, 'close_next_day', can be employed to train novel predictive models or improve existing ones. Table 2 presents the updated data for Apple Inc.

DOI: 10.54254/2754-1169/94/2024OX0211

Table 2: Updated data for Apple Inc.

Date	Close	close_next_day
2024-6-21	207.49	209.68
2024-6-20	209.68	214.29
2024-6-18	214.29	216.67
2024-6-17	216.67	212.49
2024-6-14	212.49	214.24

4. Return to Model Building and Model Evaluation

4.1. Theoretical Methods and Prediction of Regression Models

Linear regression is a fundamental regression analysis method that predicts the target variable by fitting a linear relationship, establishing a linear mathematical model, and evaluating predictions to handle the linear regression relationship between dependent and independent variables. Assuming θi is the weight parameter of linear regression, xi represents the numerical value of sample points, and n represents the number of sample points, establish a regression equation to form a surface in a high-dimensional space to fit all data points [8].

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x. \tag{1}$$

Due to the presence of some error between the actual and predicted values, denoted as ε for error, for each sample:

$$y^{(i)} = \theta^T x^i + \varepsilon^{(i)} \tag{2}$$

In the equation, y(i) represents the true value, and θT xi represents the predicted value. For linear regression algorithms, the error term ε serves as the loss function, where the algorithm aims for a smaller loss value, meaning a smaller ε is preferable. With a sufficient sample size, the errors ε(i) are independent and follow a Gaussian distribution; hence:

$$p\left(\varepsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\left(\varepsilon^{(i)}\right)}{2\sigma^2}\right) \tag{3}$$

By combining Equation (2) and Equation (3), further simplifying by introducing the logarithmic likelihood function, it arrives at the final objective function formula (4).

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (y^i - h_\theta(x^i))^2 \tag{4}$$

In general, gradient descent is typically introduced for solving. For the objective function in equation (4), differentiate with respect to θ, it gets:

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m} \sum_{i=1}^{m} (y^i - h_\theta(x^i)) x_j^i \tag{5}$$

$$\theta_j^{'} = \theta_j + \frac{1}{m} \sum_{i=1}^{m} (y^i - h_\theta(x^i))\, x_j^i \tag{6}$$

The key determinants of stock price movements commonly encompass opening, peak, trough prices, and trading volumes. However, this investigation zeros in on the closing price as the primary observation metric within the temporal framework. Consequently, five explanatory variables are chosen: Date (albeit typically used for indexing), Open Price, High Price, Low Price, and Volume. Initially, data normalization is conducted employing the StandardScaler() method from the sklearn

DOI: 10.54254/2754-1169/94/2024OX0211

library to ensure uniformity. Following this, the dataset is partitioned into training and testing subsets at an 80:20 ratio. The training subset is then subjected to a linear regression analysis utilizing the LinearRegression() module from sklearn, aiming to identify patterns and relationships within the selected variables.

4.2. Model Evaluation

In regression problems, besides caring about the model's accuracy, the proximity between predicted values and actual values is of greater concern. Therefore, this study evaluates the linear regression model using specialized metrics for regression problems such as Mean Square Error (MSE) and R-Squared.

4.2.1. Mean Square Error

Mean Square Error (MSE) represents an advanced evaluation metric employed in assessing the accuracy of model predictions against factual observations. This metric quantifies the average squared difference between the model's forecasts and the corresponding true values, offering a nuanced perspective on the prediction fidelity. A notable reduction in MSE signifies that the model's estimations align more closely with reality, reflecting enhanced predictive prowess. Conversely, an elevated MSE value underscores a model's relatively weaker capability in accurately forecasting outcomes. As such, MSE serves as a valuable benchmark for discriminating between the performance of various models, enabling researchers and practitioners to make informed decisions regarding model selection and refinement for publication-worthy contributions.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{7}$$

In the formulation, 'n' denotes the total number of observations in the sample, with 'yi' representing the individual observed values, while '$\hat{y}_i$' signifies the corresponding predicted values generated by the model. The Mean Squared Error (MSE) metric, though a powerful tool, is typically not employed in isolation but rather within the context of comparative analysis, where it serves as a cornerstone for evaluating the performance of multiple models side by side. This approach facilitates a nuanced understanding of how each model fares in terms of its ability to closely mimic the underlying data patterns, ultimately guiding the selection of the most suitable model for a given task or research endeavor.

4.2.2. R-Squared

The R-squared (R^2), a key metric, quantifies the degree of fit between a model and its corresponding data. Its range spans from 0 to 1, with higher values approaching 1 indicative of a model's superior capability in accounting for the variability present within the data. Conversely, values nearer to 0 denote a model's relatively limited explanatory power over the data, highlighting areas for potential improvement or alternative model consideration.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \overline{\hat{y}i})^2}{\sum_{i=1}^{n}(yi - \overline{y})^2} \tag{8}$$

DOI: 10.54254/2754-1169/94/2024OX0211

Table 3: Prediction based on Random Forest method.

Stock	Test MSE	Test R-squared
AAPL	0.7560161865296143	0.9995634735157912
NVDA	0.1186258197132207	0.9997593256606654
MSFT	2.179096471608457	0.9996162824477595

When utilizing the Random Forest method to predict data for Apple Inc., Nvidia Corporation, and Microsoft Corporation (e.g. Table 3), the results demonstrate extremely high prediction accuracy. Despite differences in Mean Squared Error (MSE) among the companies, with Nvidia Corporation having the lowest MSE (0.1186) indicating the smallest deviation between predicted and actual values, and Microsoft Corporation having a relatively higher MSE (2.1790), all companies exhibit R-squared values close to 1, at 0.9995, 0.9997, and 0.9996 respectively. This implies that the model can explain the majority of the variability in the data.

In conclusion, the Random Forest model has shown outstanding predictive performance on the data of these three companies, displaying a high fit to the data and accurate forecasting of future trends. This proves the robust capability and broad application prospects of the model in complex data analysis.

4.3. Model Prediction Verification

Merging the feature vector X and label y into the DataFrame, predicting the new closing price using the predict() method, combining the new closing price data into the DataFrame to create an updated dataset as shown in Table 4 below. Randomly print five rows and observe that the predicted results closely match the actual values.

Table 4: Comparison of actual and predicted prices.

Actual Close	Predicted Close
310.70	309.81
403.93	404.61
426.28	424.08
429.37	429.11
146.57	141.84

If the model perfectly captures the relationship between the independent and dependent variables, then for each data point in the training set, its predicted value will be exactly the same as or very close to the actual value, as Figure 1, Figure 2, and Figure 3 shown. In this scenario, when the actual closing price is taken as the horizontal axis and the predicted closing price as the vertical axis on a scatter plot, all points will lie on the same straight line, known as the perfect fit line. From the three graphs, it can be observed that the predicted values are nearly equal to the actual values, as they all fall on this line. Among them, MSFT stands out as the most standard.

DOI: 10.54254/2754-1169/94/2024OX0211

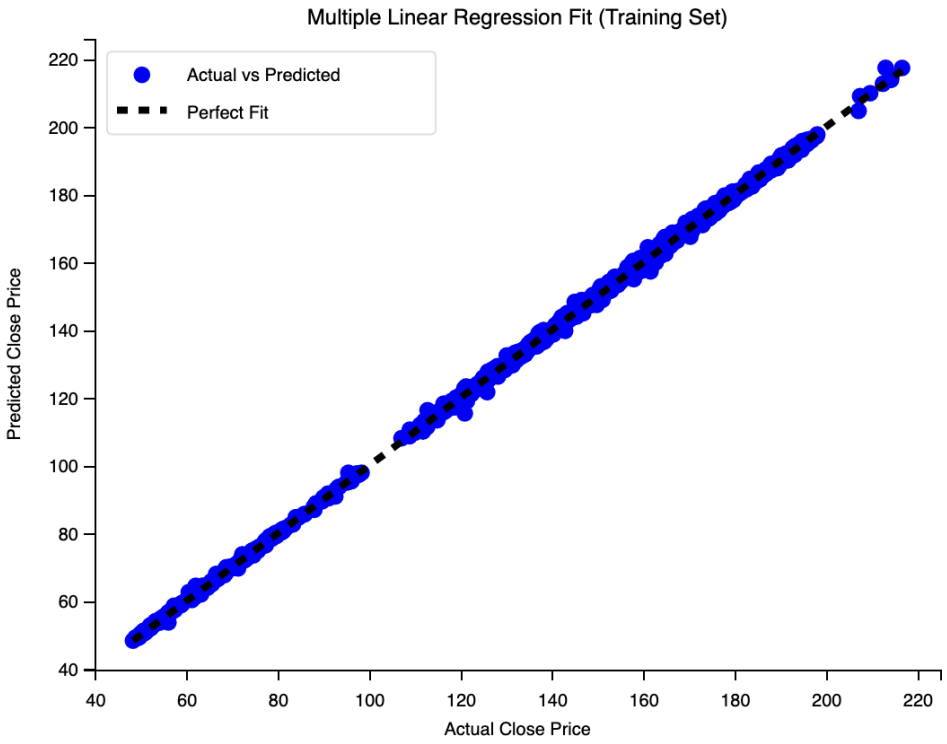

Figure 1: Multiple linear regression fit of AAPL (Photo/Picture credit: Original).

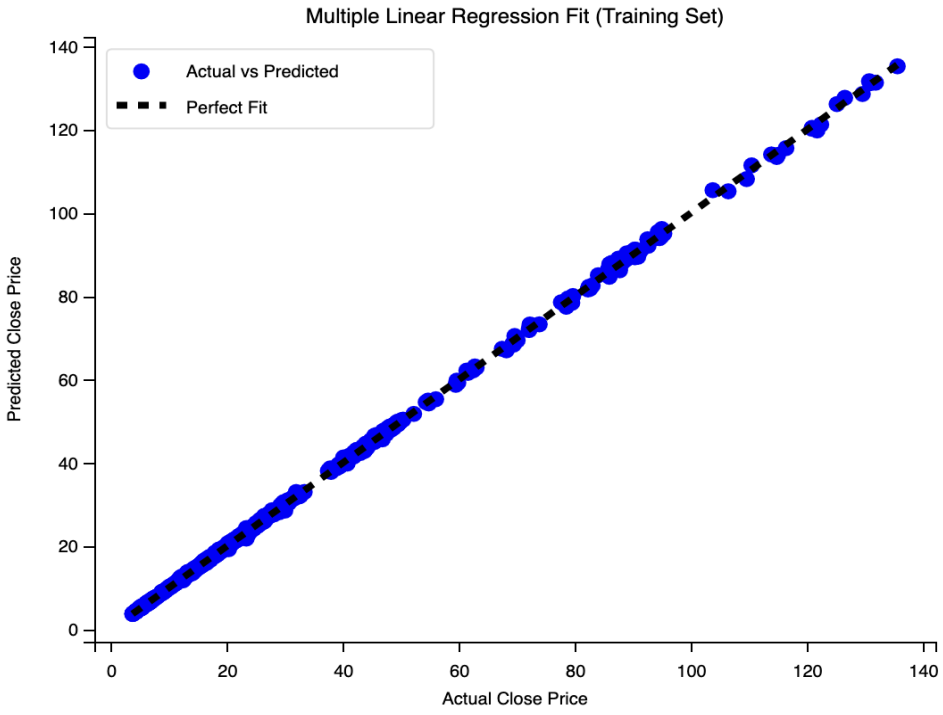

Figure 2: Multiple linear regression fit of MSFT (Photo/Picture credit: Original).

DOI: 10.54254/2754-1169/94/2024OX0211

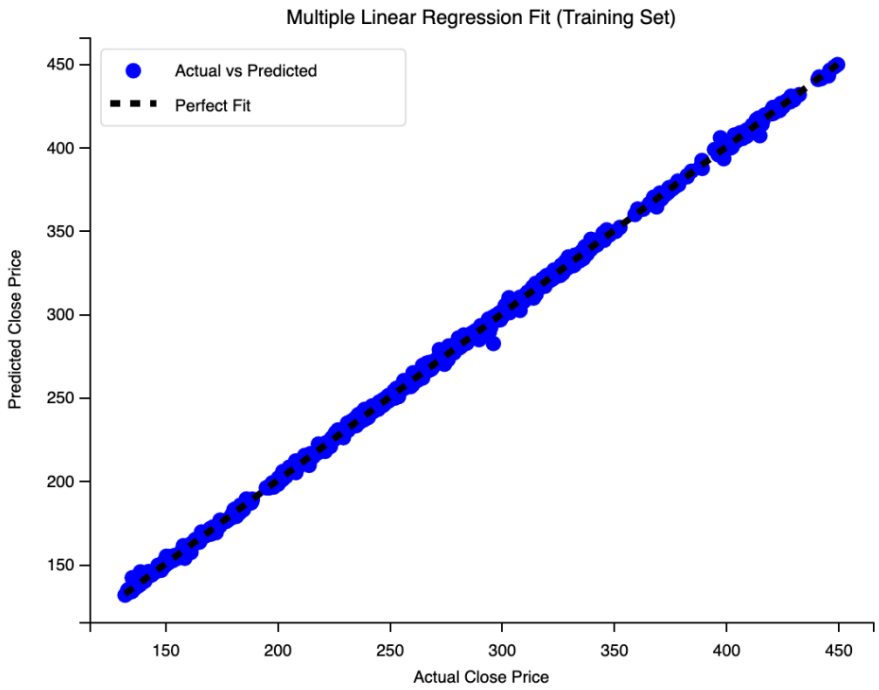

Figure 3: Multiple linear regression fit of NVDA (Photo/Picture credit: Original).

5. Classification Model Construction and Model Evaluation

5.1. Logistic Regression and Random Forest

Logistic regression, a statistical learning technique for handling classification tasks, predicts the likelihood of an input belonging to a specific category. It builds on linear regression, generating predicted values mapped through the sigmoid function to the [0,1] range, converting values into probabilities. The sigmoid function is expressed as:

$$g(z) = \frac{1}{1+e^{-z}} \tag{9}$$

Random Forest is one type of ensemble algorithm that combines multiple decision trees in parallel using two methods: data random sampling and feature random selection, in order to achieve the purpose of integration. Random Forest is widely used for classification problems due to its advantages in handling high-dimensional data and evaluating the importance ranking of features. To study the regression problem as a classification problem, new attributes need to be added to the data. Firstly, calculate the difference between the "Close" field and the "Close" field of the previous data, store the result in a new field called "diff", then determine if the "diff" field is greater than 0. If it is greater than 0, it indicates the closing price is in an "up" phase, otherwise, the stock closing price is considered to be "down" [9]. Finally, use the apply method to generate a new label field based on the condition.

Next, the same 5 attributes from Section 2.1 are chosen as independent variables, specifically the Date, Open, High, Low, and Volume for the regression model. The label attribute is selected as the target variable. The data set is divided into training and testing sets in a 7:3 ratio. Subsequently, logistic regression and random forest models are constructed.

DOI: 10.54254/2754-1169/94/2024OX0211

5.2. Model Evaluation

In classification problems, model evaluation metrics generally include accuracy, recall, F1 score, precision, ROC curve, and AUC value. In this study, accuracy, ROC curve, and AUC value are chosen as evaluation metrics.

5.2.1. Accuracy

Accuracy is defined as follows.

$$ACC = \frac{TP+TN}{TP+TN+FP+FV} \tag{10}$$

In the formula, TP signifies the count of samples where both the predicted and actual industry indices exhibit an upward trend, while TN represents the number of instances where both indices decline concurrently. Conversely, FN captures the scenario where the predicted index decreases despite an actual increase, and FP denotes cases where the predicted index rises, yet the actual outcome declines. The comparative performance of logistic regression and random forest models in terms of accuracy is visually presented in Table 5 below, offering insights into their respective strengths and areas for potential refinement.

Table 5: RFC Accuracy and LGS Accuracy of three stocks.

Stock	RFC Accuracy	LGS Accuracy
AAPL	0.5714285714285714	0.5198412698412699
MSFT	0.503968253968254	0.5674603174603174
NVDA	0.5912698412698413	0.5317460317460317

On the dataset of three companies (AAPL, MSFT, NVDA), Random Forest Classifier (RFC) generally exhibits higher accuracy than Logistic Regression (LGS). Despite LGS showing slightly higher accuracy than RFC on MSFT, RFC demonstrates better accuracy on AAPL and NVDA. This indicates that Random Forest Classifier may have advantages in handling complex data and identifying non-linear relationships, thereby enhancing prediction accuracy.

5.2.2. ROC Curve and AUC Value

To assess model performance, initiate by computing the Receiver Operating Characteristic (ROC) curve utilizing the roc_curve function from the sklearn library. This function yields three key metrics: the False Positive Rate (FPR), which quantifies the proportion of negative instances erroneously labeled as positive; the True Positive Rate (TPR), also known as Recall, reflecting the fraction of actual positive samples accurately predicted as such; and the thresholds employed to plot the ROC curve. Subsequently, employ the auc function to determine the Area Under the Curve (AUC) of the ROC plot. An AUC value approaching 1 signifies superior model performance, indicating a higher degree of separation between positive and negative classes [10].

DOI: 10.54254/2754-1169/94/2024OX0211

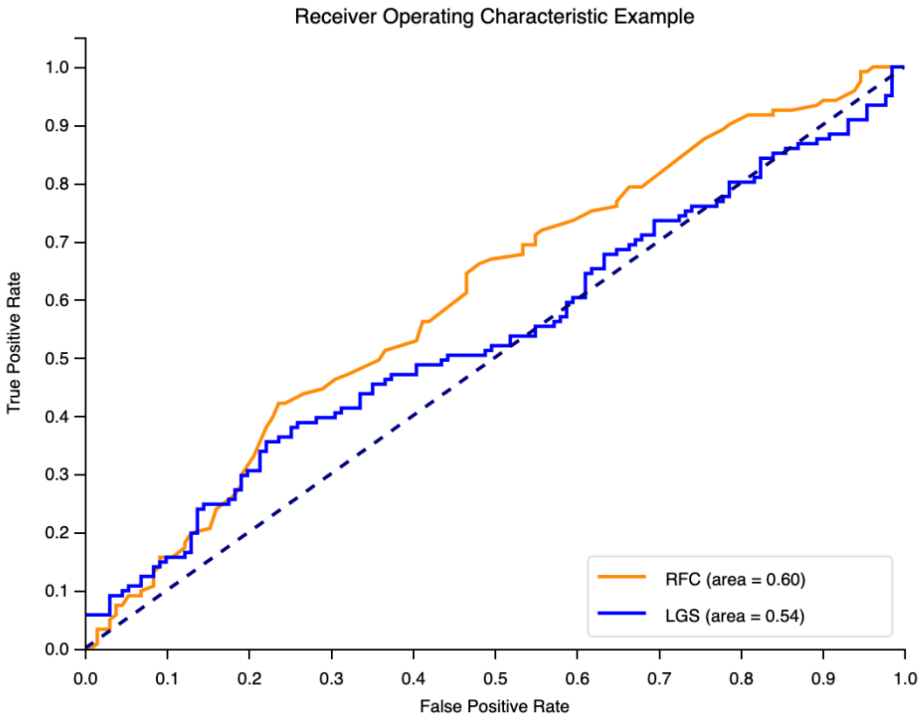

Figure 4: ROC curve of AAPL (Photo/Picture credit: Original).

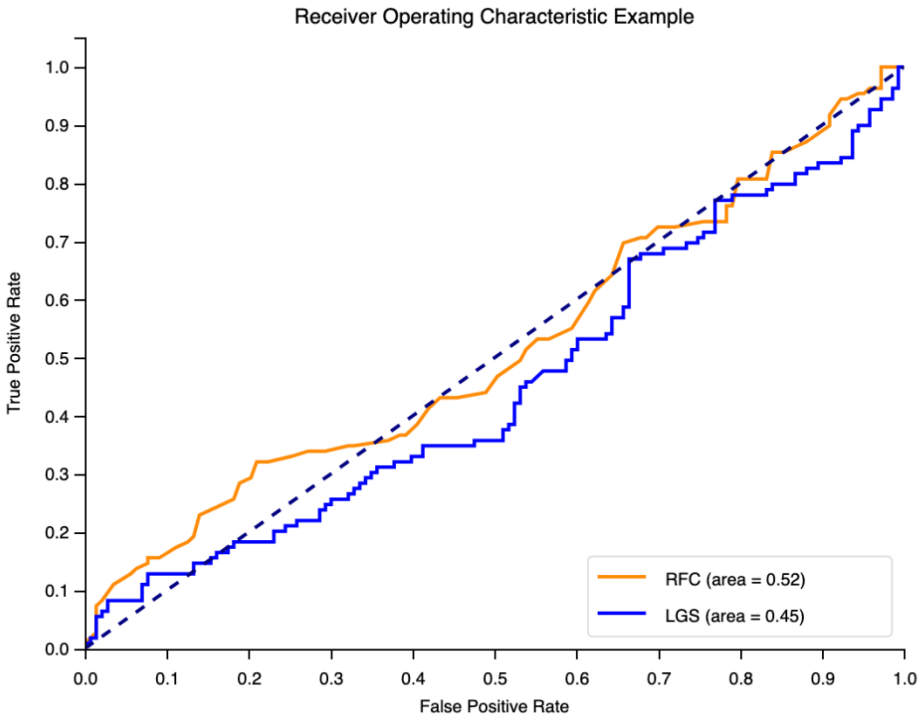

Figure 5: ROC curve of MSFT (Photo/Picture credit: Original).

DOI: 10.54254/2754-1169/94/2024OX0211

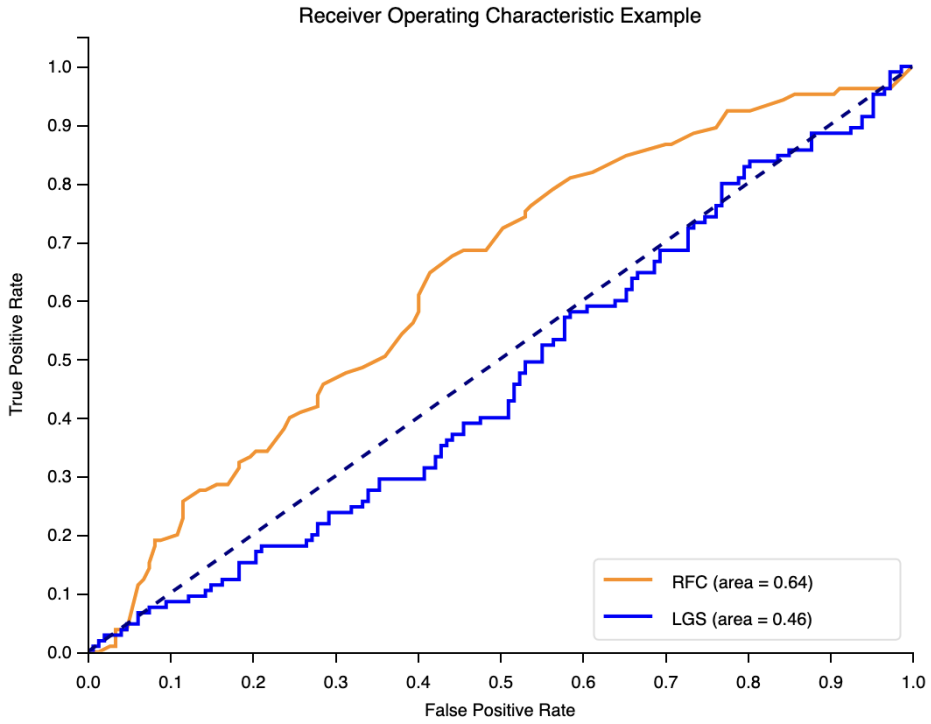

Figure 6: ROC curve of NVDA (Photo/Picture credit: Original).

The experimental results show that on the datasets of three companies (AAPL, MSFT, NVDA), the AUC value of the random forest model is generally higher than that of the logistic regression model, as Figure 4, Figure 5, and Figure 6 shown. Specifically, the random forest AUC values for AAPL and NVDA reached 0.60 and 0.64 respectively, demonstrating a better model discrimination ability, while the corresponding logistic regression AUC values were lower at 0.54 and 0.46. Although for MSFT, the random forest AUC value (0.52) was slightly higher than logistic regression (0.45), both were at a lower level, indicating limited model discrimination ability on the dataset of this company. Overall, the experimental results support that the random forest model outperforms the logistic regression model in predictive tasks. [11].

6. Conclusion

In conclusion, linear regression models exhibit higher accuracy in predicting stock prices with R2 values nearing 1, while random forests excel in predicting stock price movements with AUC values consistently above 0.52, even reaching 0.64. In today's complex stock market landscape, even a slight increase of 0.02 in chances for novice investors could potentially enhance investment profitability to some extent. Stock closing price prediction pertains to a regression problem where the closing price represents a continuous numerical value rather than a discrete category, making linear regression more appropriate for numerical predictions. This study delves into the application of candlestick chart analysis and machine learning algorithms in forecasting stock price trends, offering valuable insights and guidance to investors, financial professionals, and researchers. Despite possible localized fluctuations and short-term volatility in the market, long-term holdings aid in mitigating transient fluctuations and achieving more reliable returns.

DOI: 10.54254/2754-1169/94/2024OX0211

References

[1] *Li, B., & Long, Z. (2023). Research on the predictability of the Chinese stock market: A machine learning perspective. Journal of Management Science in China, 26(10), 138-158.*

[2] *Huang, Y. (2023). Analysis of the factors influencing investor behavior in the stock market. China Management Informationization, 26(23), 131-135.*

[3] *Wang, L., He, Y., & Jiao, D. (2023). Prediction of stock price trends based on logistic regression model: A case study of Guiyang Bank. China Management Informationization, 26(4), 156-158.*

[4] *Ma, G., & Tang, Y. (2023). Stock trend prediction in the agriculture, forestry, animal husbandry, and fishery industries based on machine learning. Investment and Cooperation(6), 47-49.*

[5] *Zhang, X., & Chen, L. (2022). Design and implementation of a hot stock analysis and recommendation system based on linear regression. Modern Information Technology, 6(22), 16-21.*

[6] *Zhao, Y., Tang, Y., & Jiang, Z. (2023). Study on the changes in human step length under load based on linear regression method. Journal of Liaoning Police Academy, 25(4), 57-62.*

[7] *Hsiao Y ,Yang M,Lo Y, et al.A Study on Competitive Trend of Global Top 100 Brands[J].Journal of Economics, Business and Management,2022,10(4):*

[8] *Yeung, W. H., & Lento, C. (2018). Ownership structure, audit quality, board structure, and stock price crash risk: Evidence from China. Global Finance Journal, 37(3), 1-24.*

[9] *Zheng, X. (2023). Comparative analysis of machine learning and deep learning techniques for prediction of the stock market. In Proceedings of the 5th International Conference on Computing and Data Science (part 3). Faculty of Medical and Health Sciences and Bioengineering Institute, University of Auckland, ITM Department, Illinois Institute of Technology, USA.*

[10] *Shi, X., Guo, P., Zheng, Q., et al. (2022). Application of ensemble learning in consumer finance auditing: A case study of random forest detecting credit card fraud. Commercial Accounting(15), 46-51.*

[11] *Zhou, L. (2021). Research on multi-factor investment in stocks based on random forest model. Financial Theory and Practice(07), 97-103.*