Comparative Analysis of XGBoost Algorithm and Linear Regression in Predicting the Trend of Investor Overreaction

Milad Shojaei Nasir Abadi[1], Habib piri[2,*] and Reza Sotudeh[3]

[1] PhD Student, Department of Accounting, Faculty of Humanities, Zahedan Branch, Islamic Azad University, Zahedan, Iran;
[2] Assistant Professor, Department of Accounting, Faculty of Humanities, Zahedan Branch, Islamic Azad University, Zahedan, Iran;
[3] Assistant Professor, Department of Financial and Accounting, Faculty of Humanities, Meybod University, Meybod, Iran;

* Correspondence: hhpiri1354@gmail.com

Citation: Shojaei Nasir Abadi, M., Piri, H., & Sotudeh, R. (2025). Comparative Analysis of XGBoost Algorithm and Linear Regression in Predicting the Trend of Investor Overreaction. *Business, Marketing, and Finance Open*, 2(2), 125-137.

Received: 19 January 2025
Revised: 06 February 2025
Accepted: 25 February 2025
Published: 01 March 2025

Abstract: Overreaction is one of the observable anomalies in financial markets that can lead to market inefficiency. This phenomenon is particularly prevalent in emerging and less developed markets. Evidence suggests that investors tend to overreact to financial events, which introduces bias into their decision-making processes. Consequently, the market deviates from its optimal efficiency. Detecting and predicting such reactions can assist investors in making more rational decisions regarding the purchase and sale of stocks and other securities. For this purpose, methods such as linear regression and the XGBoost algorithm are employed. Due to its high capability in modeling complex relationships, the XGBoost algorithm can play a significant role in analyzing investor behavior. The objective of this study is to compare the performance of these two methods in predicting the trend of investor overreaction and to contribute to the improvement of investment strategies and risk management. This study is descriptive-causal in nature and is conducted based on an experimental design with a post-event approach. To test the hypotheses, a multivariate linear regression method based on panel data and a combination of time series was used. The required information was collected through library research, and financial data from companies within the statistical population were examined. The statistical population includes all companies listed on the Tehran Stock Exchange during the period from 2011 to 2021. Using a systematic elimination sampling method, 110 companies were selected as the sample. In data analysis, the relationships between variables were examined using regression methods, and the findings were compared with the results obtained from the XGBoost algorithm. The findings indicate the superiority of the XGBoost algorithm over linear regression in terms of the coefficient of determination and the mean squared error (MSE) index. Specifically, the highest coefficient of determination in the test data for the XGBoost algorithm was found to be 0.5713, whereas for the linear regression model, it was 0.4938. Additionally, the MSE index for the XGBoost algorithm in the test data was reported as 0.002288, while for the linear regression model, it was 0.0042. These results demonstrate that the XGBoost algorithm outperforms linear regression in terms of reducing error and increasing predictive accuracy. The XGBoost algorithm, with its ability to detect complex and nonlinear patterns, offers higher accuracy in predictions. By reducing error and increasing the coefficient of determination, this model enables more precise and reliable forecasting of the persistence of investor overreaction trends. Therefore, utilizing the XGBoost algorithm can be considered an efficient method in financial data analysis and investment decision-making improvement.

Keywords: Overreaction, investment decisions, XGBoost algorithm.

1. Introduction

Financial markets, influenced by investors' irrational behaviors, encounter a phenomenon known as overreaction, which can lead to market inefficiency. This anomaly is considered one of the major challenges in capital markets, resulting in consequences such as stock price deviations from their intrinsic values. Studies conducted in major stock exchanges worldwide have indicated that investor overreaction is a key issue in this domain [1]. Emerging and less developed markets are particularly prone to this phenomenon, as evidence suggests that investors often overreact to financial events, leading to unrealistically inflated or deflated stock prices. Subsequently, these deviations are corrected, and prices return to their true levels—a process known as price reversal. Overreaction, rooted in psychological factors, causes behavioral biases in investors' decision-making, especially under uncertainty, ultimately reducing market efficiency [2]. Research, including the studies by Zarowin (1990) and Sangari et al. (2024), has shown that firm size influences the extent of investor overreaction. According to their findings, smaller firms, due to their lower market share, are more susceptible to this phenomenon, whereas an increase in firm size decreases the likelihood of overreaction [3, 4].

Behavioral theories argue that sudden fluctuations in stock markets stem from investors' adverse reactions to new information. Among the extensive research in this field, investor overconfidence has been identified as a key factor in explaining momentum. This line of research began with the study by Daniel et al. (1998), where they introduced a theoretical model called the "Investor Confidence" model, incorporating the concept of misattributed confirmation bias to explain how overreaction and underreaction form in stock markets. Subsequent studies have provided various empirical evidences supporting this model [5]. Moreover, Byun et al. (2016) emphasized that Daniel et al.'s (1998) model links the development of investor overconfidence to the persistence of overreaction behavior. According to this model, if the persistence of overreaction, as proposed by Daniel et al. (1998), contributes to momentum, a more direct measure of this persistence could offer a more accurate prediction of future returns compared to past returns. In this regard, Byun et al. (2016) proposed a method for directly measuring overreaction persistence using volume-weighted signed measures to validate the DHS model in explaining momentum profits [6]. They also introduced a trading strategy in which stocks with high positive overreaction are purchased, while stocks with significant negative overreaction persistence are short-sold. Their findings indicate that this trading strategy can generate substantial abnormal returns [7].

Security prices play a crucial role in shaping investor overreaction. Research has shown that these two variables exhibit an inverse relationship, meaning that as security prices decline, the likelihood of overreaction increases, and vice versa. Furthermore, capital market characteristics significantly influence the extent of overreaction. Factors such as price fluctuation limits, base volume for determining closing prices, trading halts, the proportion of free-floating shares, and the speed of information reflection in prices all impact this phenomenon. Reinganum (1981) argued that the likelihood of overreaction is higher in smaller firms than in larger firms, attributing this to the lack of reliable and credible information about smaller companies [8]. This information deficiency causes investors to rely excessively on rumors and past price trends, making investment decisions based on incorrect or incomplete data [9]. Consequently, stock price fluctuations are generally linked to systematic changes in firms' fundamental values, and irrational investor behavior does not directly affect returns. However, a positive correlation has been observed between investor sentiment trends and the returns of stocks that are subject to higher subjective evaluation [10].

Some investors do not use scientific methods for stock valuation and instead rely on personal judgment, mental perceptions, non-scientific information, and the prevailing psychological and emotional conditions in the market. These conditions are primarily shaped by how companies disseminate information [11]. Continuous shareholder reactions to company news and changes may be based on new analyses or received information. Such behavior reflects the market's ability to absorb information and make autonomous decisions, which can alter investors' perceptions of a company or the market. However, these reactions sometimes lead to extreme fluctuations and emotional behaviors, resulting in sudden increases or decreases in stock values. Additionally, susceptibility to rumors or incomplete information can intensify these fluctuations. Therefore, predicting investor overreaction trends in financial markets is of great importance, as these reactions significantly influence stock prices and investor decisions [12]. Firm size also plays a crucial role in determining stock risk and returns. Larger firms typically have broader financial resources, greater access to capital, and demonstrate more resilience during financial crises compared to smaller firms. In this context, accurate and reliable information can assist investors in making better decisions. Companies with more transparent informational environments are generally more attractive to investors. Additionally, stock returns reflect a firm's financial performance in the market, encompassing profitability assessments, stock valuation, and relative performance analysis compared to competitors. Trading volume serves as an indicator of the number and value of transactions conducted on a company's stock, representing the firm's market role, activity level, and the extent of available market information [1, 13-15].

The literature on investor overreaction in financial markets highlights various factors influencing investment decisions and stock price fluctuations. Taheri Nia et al. (2024) examined the impact of analytical paralysis on investor decision-making in the Tehran Stock Exchange, finding a negative and significant relationship between analytical paralysis and short-term investment decisions, with political factors having the greatest influence [14]. Khoramabadi et al. (2024) studied the effects of the COVID-19 pandemic on Iran's capital market and reported abnormal stock returns and trading volume fluctuations following the pandemic's announcement [16]. Shojaei Nasir Abadi et al. (2024) compared artificial neural networks and linear regression in predicting investor overreaction, demonstrating the superiority of neural networks in terms of R-squared and MSE [17]. Taheri Nia (2023) identified information salience effects in stock markets, associating overreaction with biases such as granularity, misjudgment, and incorrect forecasts [18], while Zia Qasemi et al. (2023) linked IPO overreaction in the Tehran Stock Exchange to market sentiment, valuation support, and firm size [15]. Hajian Nejad et al. (2022) found that investor sentiment significantly affects reactions to earnings announcements, with stronger responses to positive news in high-sentiment firms [2]. Abdoli and Heidari (2021) confirmed the existence of overreaction, underreaction, and herd behavior in Tehran's stock market [19], while Kamiyabi and Javadinia (2021) found a positive relationship between investor sentiment and accounting conservatism, though managerial ability did not moderate this relationship [20], aligning with findings by Hasanzadeh-Diva and Bozorg-Asl (2021) on behavioral biases and financial reporting quality [21]. Faghfor-Maghrebi et al. (2020) explored intermediary effects of perception and processing fluency on investor judgment, finding that linguistic sentiment and disclosure readability influence stock return expectations [22]. Davood Abadi (2020) showed that overreaction persistence enhances momentum profitability in markets with price restrictions [23], while Gol Arzi and Danaei (2019) observed overreaction among large-cap but not small-cap firms in Tehran's stock market [9]. Mohanty and Misha (2024) studied stock market overreaction in India during COVID-19, noting initial excessive reaction followed by correction [24], while Galvao et al. (2024) found that investor overreaction in Hungary and Slovakia led to predictable price trends rather than randomness [25]. Zakamulin et al. (2024) identified asymmetric overreaction patterns in bull and bear markets, with faster corrections in bear markets [26]. Quy Dong and Bertrand (2023) found momentum effects in Vietnam's stock market, where stocks with extreme upward reactions showed higher average returns, and momentum strategy returns remained strong even after adjusting for overreaction effects [27]. Truong et al. (2023) confirmed overreaction persistence in the Ho Chi Minh Stock Exchange, with loser portfolios outperforming winners in subsequent months, supporting weak-form market inefficiency [28]. Handayati et al. (2023) found that firms with higher intangible assets exhibited stronger overreaction in Indonesia's stock market [29], while Riadi et al. (2023) showed that Indonesian LQ45 stocks exhibited overreaction within three months of the COVID-19 pandemic declaration [30]. Loang (2022) observed investor sentiment-driven overreaction in European portfolios during COVID-19 but found no pre-pandemic evidence in the U.S. market [31].

Therefore, identifying and predicting overreaction responses can help investors make more rational decisions regarding stock and securities transactions. These reactions have a direct impact on market behavior and are usually associated with severe and discontinuous stock price fluctuations, which can create both opportunities and new risks for investors. Consequently, recognizing and forecasting the persistence of overreaction trends serves as a valuable tool for investors, financial analysts, and investment managers. Utilizing this tool allows them to make decisions based on more precise and informed analyses, maximizing available opportunities while minimizing financial transaction risks.

Thus, this study examines the efficiency of linear regression and the XGBoost algorithm in achieving accurate predictions of stock overreaction trends. Accordingly, the primary research question is whether the persistence of stock overreaction trends can be predicted through a comparative analysis of the XGBoost algorithm and linear regression.

2. Methodology

This study is descriptive-causal in nature and is designed as an experimental study with a post-event approach. To test the hypotheses, multivariate linear regression based on panel data, which is a combination of time series, has been used. These methods analyze the relationships among research variables by utilizing statistical and econometric techniques. The required information was collected through a library research method, and the data for hypothesis testing was extracted from the financial statements of the companies in the statistical population. The software EViews was used for data analysis.

The statistical population of this study includes all companies listed on the Tehran Stock Exchange during the period from 2011 to 2021. A systematic elimination sampling method was used to select the sample, resulting in 110 companies based on specific conditions. The criteria for selecting sample companies are as follows:

- Investment companies, banks, financial intermediaries, holdings, leasing companies, and insurance companies.
- Companies that changed their fiscal year during the study period.
- Companies that had trading suspensions for more than six months during the study period.
- Companies that were delisted from the stock exchange during the study period.
- Companies whose fiscal year does not end on March 19.
- Companies whose financial information was not available during the period from 2012 to 2020.

Additionally, financial data and information of the companies were collected from the statistical archives of the Securities and Exchange Organization and the Central Bank. Excel software was used for data classification and variable calculation, while SPSS software and Python programming language were utilized for hypothesis testing.

Independent Variables

X1: Investor Confidence - Defined as the percentage change in shares held by shareholders (Batska et al., 2018).

X2: Decision-Making Based on Private Information

The price effect model was first introduced by Roll (1992) and later expanded by Mark and Young Vieu (2000). According to the theory of Durnow, Mark, and Young (2004), if a company's stock returns exhibit strong correlation with market and industry returns, the likelihood of confidential information being embedded in the stock is low. Conversely, if confidential information exists, stock returns will have a low correlation with market and industry returns. In general, stock return variations can be divided into three components: market-related variations, industry-related variations, and firm-specific variations.

The first two components measure systematic variations, while the last component determines firm-specific variation or price nonsynchronicity. This value can be estimated using the formula 1 - R squared, where R squared is derived from the following regression equation:

$$r_{(i,j,t)} = \beta_{(i,0)} + \beta_{(i,m)} * r_{(m,t)} + \beta_{(i,j)} * r_{(j,t)} + \varepsilon_{(i,t)}$$

where $r_{(i,j,t)}$ represents the return of company i in industry j at time t, and $r_{(m,t)}$ represents the market return at time t.

The daily market return is calculated using the stock price index as follows:

$$R_{mt} = (TEDPLX_{(t+1)} - TEDPLX_t) / TEDPLX_t$$

where $TEDPLX_t$ is the stock price index (Kordestani & Teimouri, 2018).

X3: Firm Size - Defined as the natural logarithm of total assets.

X4: Stock Price - Defined as the logarithm of the market price at the end of the year.

X5: Price Limit Range - Defined as the change in price limit:

$$\Delta PriceLimit_i = (Avg(pricelimit_i)_{post}) / (Avg(pricelimit_i)_{pre})$$

where $Avg(pricelimit_i)_{post}$ is the average price limit of stock i one year after the change, and $Avg(pricelimit_i)_{pre}$ is the average price limit of stock i one year before the change.

X6: Base Volume - Calculated as follows:

$BaseVolume_{(i,t)} = 0.0008 * Capital_{(i,t)}$ if $Capital_{(i,t)} < 1000$ billion rials

$BaseVolume_{(i,t)} = 0.0005 * Capital_{(i,t)}$ if $1000 < Capital_{(i,t)} < 3000$ billion rials

$BaseVolume_{(i,t)} = 0.0004 * Capital_{(i,t)}$ if $Capital_{(i,t)} > 3000$ billion rials

where $Capital_{(i,t)}$ is the capital of company i and $BaseVolume_{(i,t)}$ is the base volume of company i. The change in this variable is calculated as follows:

$$\Delta BaseVolume_i = (Avg(BaseVolume_i)_{post}) / (Avg(BaseVolume_i)_{pre})$$

where $Avg(BaseVolume_i)_{pre}$ is the average base volume of company i in the year before the change, and $Avg(BaseVolume_i)_{post}$ is the average base volume of company i in the year after the change.

X7: Trading Halts - In this study, trading suspension is measured by the number of months without trading activity (Esfahani, 2019).

X8: Free-Float Shares - Free-float shares of industries are calculated on a quarterly basis using the weighted average method as follows and are used for testing the second hypothesis:

$$PlF = (\sum(TCS * PCF)) / TIS$$

where PlF represents the percentage of free-float shares in the industry, TCS is the total number of company shares, PCF is the percentage of free-float shares of the company, and TIS is the total number of shares in the industry (Madan-Haghighi, 2013).

X9: Timely Information Reflection in Stock Prices - This variable is measured using the equation proposed by Beeks and Brown (2006) and Andrew (2011) in their research:

$M\ squared = (\sum_{t=-364}^{t=0} |\ln(P_0) - \ln(P_t)|) / 365$

where M squared represents the timely reflection of information in stock prices, P_0 is the stock price on day zero (the beginning of year t), and P_t is the daily stock price in the market from day -364 to day zero. For simplicity, a 365-day calendar period is used.

Dependent Variable

Overreaction Trend

Based on the research by Yang et al. (2019) and Byun et al. (2016), the average of three overreaction trend indicators is considered as a measure of the overreaction trend. These three indicators include volume-based overreaction, change-in-volume-based overreaction, and idiosyncratic volatility-based overreaction. The measurement methods for these indicators are as follows:

1. **Volume-Based Overreaction (COv)** (Zhang, 2019; Byun et al., 2016):

$COv_{(i,t)} = (\sum_{J=1}^{12} (w_J * SV(i,t-J))) / ((\sum_{J=1}^{12} (Vol(i,t-J))) / 12)$

where $SV_{(i,t-J)}$ is the trading volume in month t-J.

2. **Change-in-Volume-Based Overreaction (COdv)** (Byun et al., 2016):

$COdv_{(i,t)} = (\sum_{J=1}^{12} (w_J * SdV(i,t-J))) / ((\sum_{J=1}^{12} (Vol(i,t-J))) / 12)$

where $Vol_{(i,t)}$ is the monthly change in trading volume in month t.

3. **Idiosyncratic Volatility-Based Overreaction (COIVOL)** (Byun et al., 2016):

$COIVOL_{(i,t)} = (\sum_{J=1}^{12} (w_J * SIVOL(i,t-J))) / ((\sum_{J=1}^{12} (Vol(i,t-J))) / 12)$

where $Vol_{(i,t)}$ is the monthly idiosyncratic volatility in month t.

Idiosyncratic volatility is defined as the standard deviation (sigma) of residuals from the Fama-French (1993) asset pricing model on a monthly basis:

$R_i - R_f = \alpha_1 + \beta_i (R_m - R_f) + S_i (SMB) + h_i (HML) + e_i$

where R_m represents the monthly market return, calculated as:

Market return = (Index value at the beginning of the month - Index value at the end of the month) / Index value at the beginning of the month

where R_i is the monthly return of each stock, R_f is the risk-free return, and β, S, and h are regression coefficients.

3. Findings and Results

In this section, the results of the linear regression model are first presented. Then, after identifying the variables with a significant relationship with the overreaction trend, these variables are selected as input for the data mining model. Subsequently, the results of the XGBoost algorithm are presented. Finally, the results of both approaches are compared.

To examine the general characteristics of the variables and analyze them in detail, familiarity with descriptive statistics (Table 1) related to the variables is necessary. In this section, nine influential parameters on the persistence of overreaction trends, including investor confidence (X1), decision-making based on private information (X2), firm size (X3), stock price (X4), price fluctuation range (X5), base volume (X6), trading halts (X7), free-float shares (X8), and timely information reflection in stock prices (X9), are selected as independent variables for the linear regression model.

Table 1. Descriptive Statistics of Research Variables

Symbol	Observations	Mean	Median	Standard Deviation	Minimum	Maximum	Skewness	Kurtosis
y	1210	0.118	0.102	0.074	0.000	0.733	1.722	9.709
X1	1210	0.233	0.000	0.552	0.000	3.697	3.065	13.066
X2	1210	0.465	0.460	0.191	0.150	0.800	0.051	1.753
X3	1210	14.228	14.018	1.595	10.104	20.183	0.786	3.797
X4	1210	8.449	8.259	1.038	6.011	11.653	0.444	2.574
X5	1210	0.658	0.116	1.823	-0.873	24.234	4.700	39.670
X6	1210	0.969	0.787	1.459	0.010	28.779	13.575	232.630
X7	1210	2.004	1.000	1.730	0.000	8.000	1.600	4.926
X8	1210	0.240	0.210	0.145	0.023	0.860	0.982	3.831
X9	1210	0.0292	0.0298	0.013	0.000	0.169	0.692	10.725

To examine the normality of the research variables, the Shapiro-Wilk test was used. In these tests, if the significance level is less than 5 percent (Sig < 0.05), the null hypothesis is rejected at a 95 percent confidence level.

Table 2. Shapiro-Wilk Test

Symbol	Observations	Test Statistic	Significance Level	Result
y	1210	10.978	0.000	Not Normal
X1	1210	12.350	0.000	Not Normal
X2	1210	8.622	0.000	Not Normal
X3	1210	8.277	0.000	Not Normal
X4	1210	7.103	0.000	Not Normal
X5	1210	14.397	0.000	Not Normal
X6	1210	15.678	0.000	Not Normal
X7	1210	11.402	0.000	Not Normal
X8	1210	9.730	0.000	Not Normal
X9	1210	9.295	0.000	Not Normal

Based on the results in Table 2, the significance level in the Jarque-Bera normality test for the variables is less than 5 percent, indicating that the data does not follow a normal distribution. According to the central limit theorem, since the number of observations exceeds 30, the normality assumption is not required.

Table 3. Harris Stationarity Test for All Research Variables

Variable	Test Statistic	Significance Level	Result
y	-32.294	0.000	Stationary
X1	-35.056	0.000	Stationary
X2	-28.777	0.000	Stationary
X3	-30.213	0.000	Stationary
X4	-29.744	0.000	Stationary
X5	-34.220	0.000	Stationary
X6	-34.188	0.000	Stationary
X7	-33.136	0.000	Stationary
X8	-30.688	0.000	Stationary
X9	-29.553	0.000	Stationary

As seen in Table 3, the significance level of the stationarity test for all variables is less than 5 percent, indicating that all variables are stationary.

Table 4. Chow Test Results

Model Name	Test Statistic	Significance Level	Result
Model	1.40	0.005	Panel Data

As observed, the significance level in the Hausman test is greater than 5 percent, indicating that the model follows random effects.

Table 5. Heteroskedasticity Test Results

Test Model	Test Statistic	Significance Level	Result
Hypothesis (Model)	1720.59	0.000	Heteroskedasticity Exists

The results in Table 5 indicate that the significance level of the ARCH test in the research model is less than 5 percent, confirming the presence of heteroskedasticity in the error terms. This issue was resolved in the final model estimation using the GLS method.

Table 6. Hypothesis Test Results

Variable	Symbol	Coefficient	Standard Error	t-Statistic	Significance Level	Collinearity
Investor Confidence	X1	0.003	0.004	0.96	0.337	1.43
Decision-Making Based on Private Information	X2	0.093	0.015	5.99	0.000	2.83
Firm Size	X3	0.001	0.002	0.81	0.419	2.96
Stock Price	X4	-0.004	0.002	-1.79	0.079	2.99
Price Fluctuation Range	X5	-0.001	0.001	-0.89	0.375	4.37
Base Volume	X6	0.009	0.002	4.78	0.000	2.54
Trading Halts	X7	0.001	0.001	1.25	0.211	2.07
Free-Float Shares	X8	0.092	0.019	4.69	0.000	2.83
Timely Information Reflection in Stock Prices	X9	0.710	0.188	3.77	0.000	2.61
Intercept	-	0.035	0.027	1.26	0.209	-

Wald Statistic: 682.06; Wald Significance Level: 0.000; Mean Squared Error (MSE): 0.0042; Adjusted R-Squared: 0.4938; Durbin-Watson: 2.351

The results in Table 6 show that the variables decision-making based on private information, base volume, free-float shares, and timely information reflection in stock prices have a significance level of less than 5 percent, and their coefficients are positive. Therefore, they are accepted at a 95 percent confidence level. However, the variables investor confidence, firm size, stock price, price fluctuation range, and trading halts have a significance level greater than 5 percent and are therefore not confirmed at a 95 percent confidence level.

In this section, all rows of the dataset were examined for missing data. The dataset used in this study had no missing data. The final step involved normalizing the column values. To achieve this, the column values were normalized between 0 and 1. Equation 1 illustrates the normalization process. For each column, the maximum and minimum values were calculated. Preprocessing the data makes interpretation and utilization easier. This process also allows categorical data to be used in the model training process. Additionally, data preprocessing ensures that missing values due to human or system errors do not exist. The advantage of using linear normalization in this study is that it preserves the relationships between the original data values.

The XGBoost model was then trained using the dataset for predicting the persistence of investor overreaction trends. Since XGBoost training involves a set of parameters, the possible values for these parameters are presented in Table 7.

Table 7. Tunable Parameter Values in the XGBoost Algorithm

Parameter	Possible Values	Optimal Value
max_depth	[3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]	10
gamma	[0.001, 0.005, 0.01, 0.02]	0.02

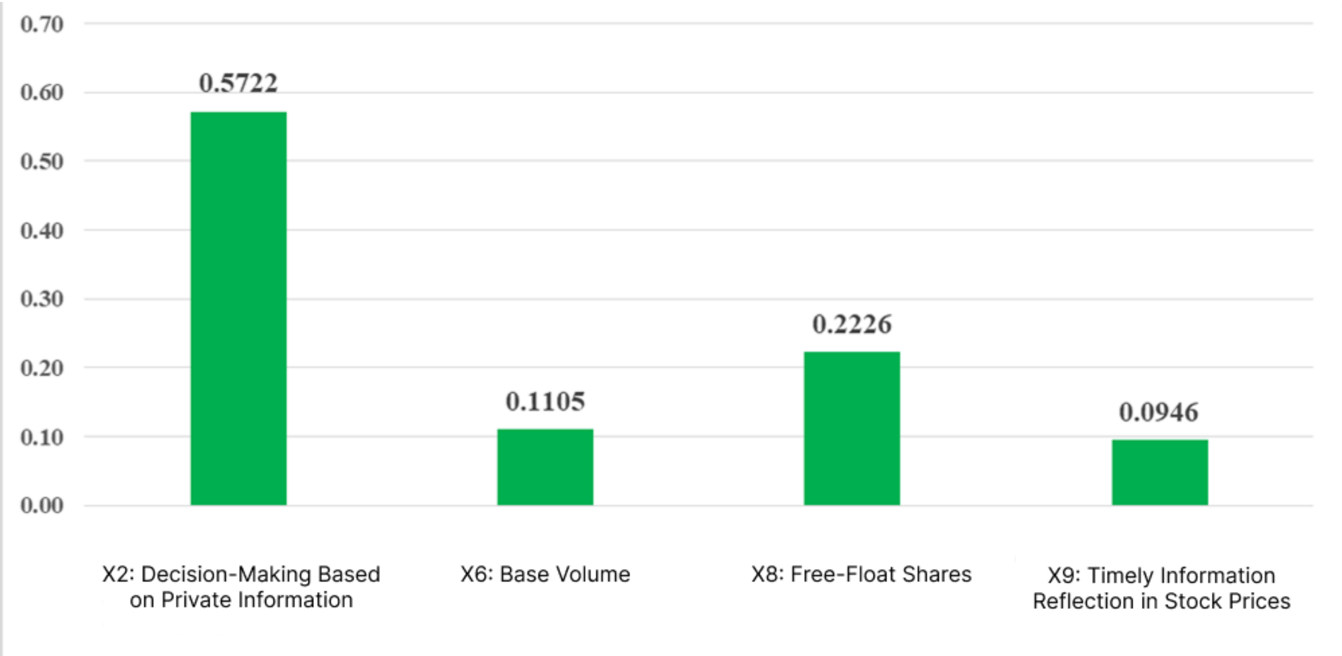
learning_rate	[0.001, 0.005, 0.01, 0.05]	0.05
n_estimators	[100, 200, 300, 400, 500, 600]	400

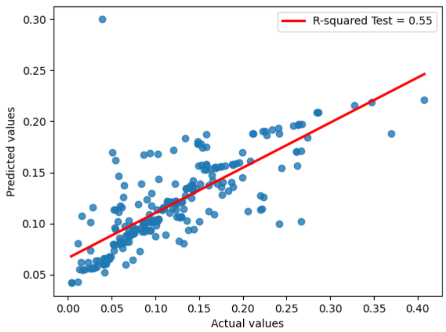
For training the XGBoost model, the Python programming language was used, specifically employing the popular xgboost and Hyperopt libraries. The dataset was divided into 80 percent for training and 20 percent for testing. A grid search (using Hyperopt) was applied to the specified parameter values (Table 7) for determining optimal values. Hyperopt is a powerful Python library for hyperparameter optimization developed by James Bergstra. It utilizes a form of Bayesian optimization to tune parameters, enabling the selection of the best parameters for a given model. This library helps define optimal parameters from predefined hyperparameter values. At the end of the process, the best parameters (Table 7) were selected from the listed hyperparameters.

Table 8. Comparison of XGBoost Algorithm R-Squared and MSE Results

Test MSE	Training MSE	Test R-Squared	Training R-Squared
0.002403	0.002113	0.5496	0.6248

As observed in Table 8, the best results (for testing the second hypothesis) correspond to the XGBoost algorithm with max_depth = 10, gamma = 0.02, learning_rate = 0.05, and n_estimators = 400. In this case, the R-squared value is at its highest, while the MSE is at its lowest compared to other configurations.





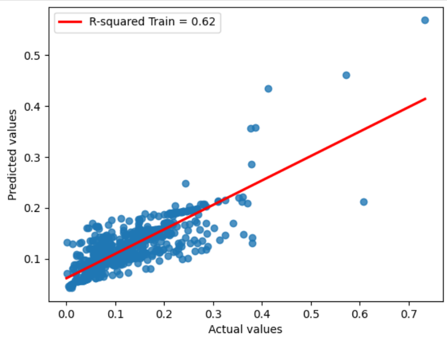


Figure 1. Importance of Input Parameters and R-Squared of the XGBoost Model

The results in the figure indicate that the MSE values for training and test data are 0.002113 and 0.002403, respectively. The R-squared values for the training and test data are 0.6248 and 0.5496, respectively. The importance of input parameters in the XGBoost model is also displayed in the figure. The calculated score for each feature indicates how valuable or useful it was in constructing the boosted decision trees.

The comparison of the presented results (for testing the third hypothesis) in Table 9 demonstrates the superiority of the XGBoost algorithm in terms of R-squared and MSE. Specifically, the highest R-squared value for the XGBoost algorithm for test data is 0.5496, whereas for the linear regression model, it is 0.349. Additionally, the results in Table 9 indicate that the MSE for the XGBoost algorithm on test data is 0.002403, while for the linear regression model, it is 0.004. Thus, similar to R-squared, the MSE index is also better in the XGBoost model.

Table 9. Comparison of XGBoost Algorithm and Linear Regression Model R-Squared and MSE Results

Test MSE	Training MSE	Test R-Squared	Training R-Squared	Model
0.002403	0.002113	0.5496	0.6248	XGBoost Algorithm
0.004	0.4938	0.349	-	Linear Regression

4. Discussion and Conclusion

This study was conducted with the aim of examining and modeling the prediction of the persistence of investor overreaction using the XGBoost algorithm. Investor overreaction is considered one of the key factors influencing stock returns, and understanding and predicting this phenomenon can help investors and economic decision-makers make optimal decisions. In this regard, the primary objective of this research is to provide an efficient model for predicting the persistence of this behavior. The XGBoost algorithm, which is based on decision trees and a gradient boosting framework, was chosen as the main predictive method due to its high capability in identifying complex and nonlinear patterns. This model, by utilizing advanced structures, enables more accurate and reliable predictions of stock return trends.

The significance of this research stems from the fact that the capital market, due to its complex and nonlinear nature, requires more advanced and precise analytical methods to predict investor behavior and price fluctuations. The findings of this study can provide more effective analytical tools for investors and economic policymakers and improve the accuracy of capital market-related predictions. The results of hypothesis testing related to the first hypothesis are presented as follows:

There is no significant relationship between investor confidence (X1) and the persistence of overreaction (Y). In other words, investor confidence alone is not capable of predicting stock behavior in the market. Although it may be assumed that high confidence has a significant impact on market fluctuations, research indicates that this relationship is weak or even negligible. Some studies have also confirmed the hypothesis that there is no significant relationship between these two variables. This issue is partly due to the complexity of financial market behavior, as multiple factors such as economic growth, debt rates, inflation, economic news, and global events influence stock behavior. As a result, the persistence of investor overreaction is affected by a set of factors beyond investor confidence. Examining this issue requires more detailed analyses and comprehensive studies across different time periods and market conditions.

There is a significant positive relationship between decision-making based on private information (X2) and the persistence of overreaction (Y). This means that the use of private information in investment decision-making can significantly impact stock behavior and improve market trend prediction. This can lead to greater price fluctuations and, in some cases, increased accuracy in market behavior predictions. However, reliance on private information may result in informational asymmetry, market unfairness, and even instability. The impact of this factor largely depends on regulatory frameworks and the level of financial market development, as strict regulations can prevent insider trading abuses and maintain market balance.

There is no significant relationship between firm size (X3) and the persistence of overreaction (Y). Although larger firms have more financial resources to meet their ongoing needs, and investors take this into account, the market's reaction to disclosed information is not necessarily influenced by firm size. For example, large firms may focus on future investments due to better access to financial resources, but this alone does not predict stock behavior in the market. Studies [32-34] have confirmed that firm size does not significantly affect investor sentiment. However, findings from other researchers, such as Ananzeh et al. (2013), Choi et al. (2012), and Lini (2012), contradict these results and suggest a significant relationship between firm size and investor sentiment.

There is no significant relationship between stock price (X4) and the persistence of overreaction (Y), meaning that stock price changes alone cannot predict overreaction behavior in the market. Despite the possibility that extreme price fluctuations may influence investor reactions, this effect is not always definitive and may vary under different market conditions. In other words, an increase or decrease in stock prices does not necessarily strengthen or weaken the overreaction trend. The results of this study contradict the prior findings of [32-35] which believed that stock returns have a positive impact on investor behavior.

There is no significant relationship between the price fluctuation range (X5) and the persistence of overreaction (Y), meaning that changes in the fluctuation range cannot predict overreaction behavior in the market. In contrast, base volume (X6) has a significant positive effect on the persistence of overreaction, meaning that an increase or decrease in base volume can cause notable changes in this trend. Additionally, trading halts (X7) do not have a significant relationship with the persistence of overreaction and cannot be used as a predictive factor. On the other hand, free-float shares (X8) have a significant positive relationship with the persistence of overreaction, meaning that changes in the level of free-float shares can affect the stability of this trend. Furthermore, timely information reflection in stock prices (X9) also has a significant positive effect on the persistence of overreaction, indicating that the speed of information dissemination and market reaction can cause changes in this trend.

The best results for confirming the second hypothesis of the study correspond to the XGBoost algorithm with max_depth = 10, gamma = 0.02, learning_rate = 0.05, and n_estimators = 400. In this model, the R-squared value for test data was 0.5713, while for the linear regression model, it was 0.4938, demonstrating the higher accuracy of XGBoost. Additionally, the MSE value for XGBoost was 0.002288, compared to 0.0042 for the linear regression model, further confirming the superiority of XGBoost in reducing model error. Therefore, the results indicate that the XGBoost algorithm performs better in predicting the persistence of overreaction compared to traditional models.

Investors should move away from making decisions based on emotions and overconfidence, instead prioritizing financial analysis, company performance evaluation, and risk assessment. Relying on objective financial indicators rather than sentiment-driven reactions can help mitigate the effects of overreaction in stock markets. Market participants, including institutional and retail investors, should adopt systematic investment strategies that incorporate thorough risk analysis, industry trends, and macroeconomic factors. Additionally, regulatory bodies can enhance market stability by promoting transparency in financial disclosures and encouraging investors to rely on data-driven decision-making processes. By reducing reliance on psychological biases, investors can minimize irrational market fluctuations and make more informed investment choices.

The integration of machine learning techniques, particularly ensemble learning models, can improve stock market predictions and provide more reliable insights into investor behavior. These advanced models help uncover hidden relationships in market data, allowing investors and analysts to develop more precise trading strategies. Furthermore, examining the broader implications of overreaction on corporate performance and financial markets through economic modeling can assist investors in refining their approaches to portfolio management. Implementing strategies that account for the rapid dissemination of financial information and its effect on stock prices can also help reduce excessive market volatility. Investors and policymakers should focus on creating adaptive strategies that not only enhance investment returns but also prevent destabilizing speculative behaviors.

Authors' Contributions

Authors equally contributed to this article.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all participants who participate in this study.

Conflict of Interest

The authors report no conflict of interest.

Funding/Financial Support

According to the authors, this article has no financial support.

References

[1] J. W. Goodell, S. Kumar, P. Rao, and S. Verma, "Emotions and stock market anomalies: a systematic review," *Journal of Behavioral and Experimental Finance,* vol. 37, p. 100722, 2023, doi: 10.1016/j.jbef.2022.100722.

[2] A. Hajian Nejad, H. Amiri, and A. Khorram Kouhi, "Analyzing the impact of investor sentiment on reactions to earnings news of companies listed on the Tehran Stock Exchange," *Journal of Financial Accounting Research,* vol. 14, no. 1, pp. 1-20, 2022, doi: 10.22108/far.2022.132213.1852.

[3] P. Zarowin, "Size, seasonality and stock market overreaction," *Journal of Financial and Quantitative Analysis,* vol. 25, no. 1, pp. 113-125, 1990, doi: 10.2307/2330891.

[4] M. Sangari, M. A. Aghaei, F. Avaz Zadeh Fath, and A. Pirzad, "Investigating the effect of underreaction and overreaction on the risk premium in the Iranian capital market," *Investment Knowledge,* vol. 13, no. 52, pp. 711-726, 2024.

[5] K. Daniel, D. Hirshleifer, and A. Subrahmanyam, "Investor psychology and security market under-and overreactions," *The Journal of Finance,* vol. 53, no. 6, pp. 1839-1885, 1998, doi: 10.1111/0022-1082.00077.

[6] S. J. Byun, S. S. Lim, and S. H. Yun, "Continuing overreaction and stock return predictability," *Journal of Financial and Quantitative Analysis,* vol. 51, no. 6, pp. 2015-2046, 2016, doi: 10.1017/S0022109016000594.

[7] N. T. Yang, H. H. Chu, K. C. Ko, and S. W. Lee, "Continuing overreaction and momentum in a market with price limits," *Pacific-Basin Finance Journal,* vol. 48, pp. 56-71, 2018, doi: 10.1016/j.pacfin.2018.01.005.

[8] M. R. Reinganum, "Misspecifications of capital asset pricing: Empirical anomalies based on earnings yields and market values," *Journal of Financial Economics,* vol. 9, pp. 19-4, 1981, doi: 10.1016/0304-405X(81)90019-2.

[9] G. Gol Arzi and K. Danai, "Investigating the overreaction of shareholders and comparing it in small and large companies in the Tehran Stock Exchange," *Journal of Financial Management Perspectives,* vol. 25, no. 5, pp. 59-76, 2019.

[10] S. D. Shams and A. T. Esfandiari Moghadam, "The relationship between herd behavior and the performance and characteristics of investment companies," *Research in Financial Accounting and Auditing,* vol. 10, no. 38, pp. 47-66, 2018.

[11] A. Zarei and R. Darabi, "The impact of investors' emotional biases on voluntary disclosure in the Iranian capital market," *Research in Financial Accounting and Auditing,* vol. 10, no. 37, pp. 131-157, 2018.

[12] Y. Xia and G. R. Madni, "Unleashing the behavioral factors affecting the decision making of Chinese investors in stock markets," *PLOS ONE,* vol. 19, no. 2, p. e0298797, 2024, doi: 10.1371/journal.pone.0298797.

[13] S. Dehghan Khavari and S. H. Mir Jalili, "The interaction of systematic risk with stock returns on the Tehran Stock Exchange," *Financial Economics,* vol. 13, no. 49, pp. 257-282, 2019.

[14] M. Taheri Nia, M. Mohammadi, and B. Bahramian, "Analyzing the effect of analytical paralysis on investors' decisions in the Tehran Stock Exchange," *Accounting and Auditing Reviews,* vol. 31, no. 1, pp. 95-122, 2024.

[15] M. Zia Qasemi, O. Pourheidari, and A. Khodami Pour, "Investigating the overreaction of initial public offerings in the Tehran Stock Exchange: Evidence from the effect of quota allocation," *Applied Research in Financial Reporting,* vol. 12, no. 2, pp. 39-71, 2023.

[16] M. Khorram Abadi, H. Abdi, and S. Lashgar Ara, "Investigating investor sentiment and the Iranian capital market's reaction to the COVID-19 pandemic (event study approach)," *Advances in Accounting,* vol. 16, no. 1, 2024, doi: 10.22099/jaa.2024.48558.2375.

[17] M. Shojaei Nasir Abadi, H. Piri, and R. Sotoodeh, "Comparative analysis of artificial neural networks and linear regression in predicting the persistence of overreaction among shareholders," *Accounting and Auditing Reviews,* vol. 31, no. 3, pp. 547-572, 2024, doi: 10.22099/acctgrev.2024.373333.1008920.

[18] M. Taheri Nia, "Understanding the phenomenon of information prominence: Identifying and prioritizing its consequences in the stock and securities market," *Journal of Accounting and Management Audit,* vol. 12, no. 45, pp. 71-83, 2023.

[19] G. Abdoli and M. Heydari, "Investigating the disproportionate reaction of investors in the Iranian stock market," *Financial Management Strategy,* vol. 9, no. 4, pp. 87-106, 2021.

[20] Y. Kamiyabi and A. Javadinia, "Investigating the moderating effect of management ability on the relationship between investor sentiment and accounting conservatism," *Accounting and Auditing Reviews,* vol. 28, no. 1, pp. 102-134, 2021.

[21] S. M. Hasanzadeh Diva and M. Bozorg Asl, "The impact of behavioral biases of investors on the risk of future stock price drops with an emphasis on the role of financial reporting quality," *Journal of Accounting and Management Audit,* 2021.

[22] Y. Faghfor Maghrebi, S. H. Sajadi, H. Eshteh Ashari, and A. Rezaeian, "Explaining the mediating effect of investors' perception, processing dominance, and credibility of disclosures on the relationship between verbal sentiment, readability, expertise, and investors' judgment and decision-making," *Accounting and Auditing Reviews,* vol. 27, no. 1, pp. 87-113, 2020.

[23] A. Davood Abadi, "Investigating the informational content of the persistence of overreaction based on trading volume and specific volatility and its impact on momentum profitability in price-restricted markets," Islamic Azad University, Arak Unit, 2020.

[24] P. Mohanty and S. Mishra, "Did the Indian stock market overreact to Covid-19?," *The North American Journal of Economics and Finance,* vol. 70, p. 102072, 2024, doi: 10.1016/j.najef.2023.102072.

[25] R. Galvão, R. Dias, C. Palma, P. Alexandre, and S. Gonçalves, "Stock Market Instability and Investor Overreaction," *[Publication Title],* p. 49, 2023.

[26] V. Zakamulin, "Stock price overreaction: evidence from bull and bear markets," *Review of Behavioral Finance,* vol. 16, no. 6, pp. 998-1011, 2024, doi: 10.1108/RBF-03-2024-0088.

[27] L. Quy Duong and P. Bertrand, "Overreaction and momentum in the Vietnamese stock market," *Managerial Finance,* vol. 49, no. 1, pp. 13-28, 2023, doi: 10.1108/MF-01-2022-0013.

[28] L. D. Truong, G. N. Cao, H. S. Friday, and N. T. Doan, "Overreaction in a Frontier Market: Evidence from the Ho Chi Minh Stock Exchange," *International Journal of Financial Studies,* vol. 11, no. 2, p. 58, 2023, doi: 10.3390/ijfs11020058.

[29] P. Handayati, L. Indrayani, and Y. Indrayono, "The effect of underreaction and overreaction of firm stocks to firm valuation: data from the Indonesian Stock Exchange," *International Journal of Business and Systems Research,* vol. 17, no. 2, pp. 121-142, 2023, doi: 10.1504/IJBSR.2023.129195.

[30] D. R. Riadi, E. Garnia, and T. T. Tahmat, "Impact of The Covid 19 Pandemic: Is There Overreaction in LQ 45 Stock?," *Jurnal Manajemen dan Keuangan,* vol. 12, no. 1, pp. 178-194, 2023, doi: 10.33059/jmk.v12i1.4727.

[31] O. K. Loang, "Overreaction, investor sentiment and market sentiment of COVID-19," *Vision,* p. 09722629221087386, 2022, doi: 10.1177/09722629221087386.

[32] A. Assan and T. Thomas, "Stock returns and trading volume: does the size matter?," *Investment Management and Financial Innovations,* vol. 10, no. 3, pp. 76-88, 2013.

[33] A. Salem Al Ghazali, "The Relationship between Trading Volume and Stock Returns Index of Amman Stocks Exchange Analytical Study," *Global Journal of Management and Business Research: Economics and Commerce,* vol. 14, no. 7, pp. 8-28, 2014.

[34] b. Ralf, M. Glaser, S. Schaarschmidt, and S. Stankiewicz, "The stock return - trading volume relationship in European countries: evidence from asymmetric impulse responses," 2014. [Online]. Available: www.unikonstanz.de.
[35] R. Kant, "TESTING OF RELATIONSHIP BETWEEN STOCK RETURN and trading volume in india," *International Journal of Multidisciplinary Research,* vol. 1, no. 6, pp. 371-393, 2011, doi: 10.3923/rjbm.2012.30.39.