Dean&Francis

ISSN 2959-6157

Predicting Stock Market Trends and Analyzing Daily Returns Using Statistical Modeling: A Case Study on Google Stock

Leying Chang

Department of statistics, University of California Davis, Davis, United States

Corresponding author: leychang@ucdavis.edu

Abstract:

A key component of financial planning is stock market forecasting, which assists investors in choosing how to allocate their assets and manage their risk. In particular, Auto Regressive Integrated Moving Average (ARIMA) and linear regression are two statistical and mathematical models that are being evaluated for their predictive power in relation to daily returns and stock prices. Accurate forecasting is difficult due to the inherent volatility and unpredictability of financial markets, which highlights the necessity for reliable models. To ensure accurate predictions, the methodology consists of data preprocessing, model implementation, and diagnostic assessments. Results indicate that the ARIMA model effectively captures long-term trends in stock prices, making it suitable for general forecasting. However, the linear regression model exhibits inconsistent performance when predicting daily returns, especially during periods of high volatility, as evidenced by increased residual errors. This research highlights the importance of selecting appropriate predictive models and integrating advanced techniques to enhance accuracy. The findings provide valuable insights into improving financial forecasting practices, ultimately contributing to better decision-making in investment strategies.

Keywords: stock market forecasting, ARIMA model, linear regression

1. Introduction

In order to reduce risk and optimize profits, stock market forecasting is essential to financial decision-making for analysts, investors, and institutions. It helps them predict future fluctuations in stock prices. Accurate predictions are difficult yet necessary in the stock market because of its inherent volatility, which is impacted by a wide range of factors such as political developments, company-specific news, and macroeconomic data. It is challenging to predict stock prices accurately since doing so requires com-

plex statistical and mathematical models that take into account both short- and long-term trends. Many methods for predicting stock prices have been studied in a number of studies. For instance, forecast accuracy has been increased by combining contemporary machine learning techniques with conventional financial analysis approaches [1]. Deep learning models have also been used to improve the prediction of price trends in the Forex and stock markets [2]. To improve the precision and efficacy of stock price forecasting, a number of strategies have been used, such as sentiment analysis based on social media data, machine learning techniques like Support Vector Machines (SVM) and Artificial Neural Networks (ANN), and fundamental and technical analysis [3].

For time series forecasting, the Auto Regressive Integrated Moving Average (ARIMA) model is a commonly used technique in the stock market. By using differencing to account for non-stationarity, it combines moving average and autoregressive components. It has been shown that ARIMA models are useful for identifying underlying trends in stock price data and for generating accurate short-term projections. Beyond the stock market, the ARIMA model can be applied not only to predict disease outbreaks and trends in health data [4] but also to forecast oil and gas production by capturing linear patterns in complex time-series data [5]. This makes it a versatile and powerful tool for analyzing various types of time-dependent data across different fields. Aside from ARIMA, linear regression models are widely used to analyze stock market returns by considering factors such as past returns, moving averages, and other market indicators. Linear regression also proves to be a versatile tool in various fields. For instance, in gene regulation, it can help identify relationships between transcription factors and gene expression [6]. In building energy assessments, it provides a straightforward yet reliable way to estimate thermal energy demand based on selected parameters, making it useful for preliminary energy evaluations [7]. Additionally, in energy demand forecasting, it has been applied to predict building energy requirements and estimate industrial energy consumption using variables like production levels, macroeconomic factors, and energy efficiency measures [8]. By effectively modeling complex systems in these diverse domains, linear regression demonstrates its strength and adaptability. When combined with ARIMA, these models offer a more comprehensive approach to stock market forecasting, capturing both price trends and daily return fluctuations.

This study builds on established theories in time series analysis, which use linear regression and autoregressive models to describe market behavior. The goal is to contribute to existing research in stock market forecasting by taking a closer look at these methods and evaluating their effectiveness and accuracy. This paper will explore how these predictive techniques can be applied to real-world financial data, providing both theoretical and practical insights into stock market forecasting.

2. Methodology

The methodology applied to forecast stock market trends and analyze daily returns is based on a structured approach that integrates time series modeling with linear regression, aiming to reveal patterns in stock prices and assess short-term volatility. This approach emphasizes both predictive precision and a deep comprehension of market dynamics. The following sections provide a detailed explanation of each step involved in the process.

Data preparation is the initial stage, which involves importing historical stock data into the analysis platform. Typical variables in this dataset include Date, Open, High, Low, Close, and Volume. To maintain data consistency, any missing values are filled in utilizing interpolation or forward-filling techniques. Calculated as the percentage change in closing prices from one trading day to the next, daily returns are an essential component. This variable is the foundation of the study of linear regression that follows. Preprocessing, like previous research, consists on clustering related data points to facilitate pattern recognition by the model, perhaps resulting in more precise predictions [9].

The study will use the function to calculate daily returns as the percentage change in closing prices from one day to the next:

$$DailyReturn = \frac{Close_t - CLose_{t-1}}{Close_{t-1}} \times 100 \quad (1)$$

Making sure the data is stationar after preprocessing is an important part of time series analysis. The Augmented Dickey-Fuller (ADF) test, which looks for the existence of a unit root, is used to evaluate stationarity. The ADF test's null hypothesis suggests non-stationarity. In the event that the p-value exceeds 0.05 and the null hypothesis is not rejectable, the time series is deemed non-stationary. In such instances, transformations are applied to stabilize variance and remove trends. Common techniques include logarithmic transformation and differencing to make the data suitable for forecasting models. Logarithmic transformation is important in various studies because it helps normalize data distributions, making them more suitable for statistical tests like Analysis of Variance (ANOVA) and t-tests, especially for small sample sizes. For example, in a study examining EC50 values for fungicide sensitivity, applying a logarithmic transformation greatly improved the normal-

ISSN 2959-6157

ity and homogeneity of the datasets, thereby increasing the power and precision of statistical tests. This method is similarly beneficial for transforming skewed financial data, ensuring that subsequent analyses are accurate and reliable [10]. Once the data is rendered stationary, it is decomposed into three main components: trend, seasonality, and residual. This decomposition helps identify the long-term movement of the stock (trend), periodic patterns occurring at specific intervals (seasonality), and irregular variations (residual). By separating these components, the model gains a clearer understanding of the different factors driving stock price fluctuations.

To forecast stock prices, an ARIMA model is utilized, which integrates three key components: autoregression, differencing, and moving averages. Autoregression examines the relationship between current values and their historical values, differencing is used to remove trends and stabilize the data, and moving averages smooth out fluctuations by analyzing past errors. These components work together to give ARIMA an accurate representation of both trends and anomalies in the behavior of stock prices over time. Metrics like Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are used in ARIMA to assess how accurate the predicted stock prices are. Additionally, the residuals—the differences between actual and predicted values—are examined to make sure there are no observable patterns and that they roughly follow a normal distribution, which indicates that the model has captured all significant trends and randomness in the data. This process validates the accuracy of the model. The ARIMA model can be used for forecasting when it has been validated, and its dependability is assessed by contrasting expected values with actual results over time. Further performance evaluation of the ARIMA model involves diagnostic plots, including standardized residuals to verify if residuals are centered around zero, histograms to visualize residual distribution, Q-Q plots to assess normality, and autocorrelation plots to detect any remaining correlations. The residuals' lack of considerable autocorrelation is proof positive that the underlying data patterns have been adequately captured by the ARIMA model.

For daily return analysis, a linear regression model is employed, with daily stock returns as the dependent variable. Lagged returns are included as the main feature to capture temporal dependencies in the return series. The dataset is first split into training and testing subsets to assess the model's performance on unseen data. After fitting the model to the training set, predicted values are plotted against actual returns to visually evaluate the model's accuracy. Additionally, residual plots are analyzed to examine the error distribution and detect any patterns or inconsistencies, helping to validate the assumptions underlying linear regression.

3. Result

3.1 Visualize the Data

The author aims to analyze Google's stock performance over time, focusing on trends and market behavior from 2013 to 2018. Fig. 1 displays the closing prices of Google's stock, demonstrating a clear upward trend over these years. Although there are periods of increased volatility, particularly between 2015 and 2016, the overall trajectory remains positive, indicating sustained growth. The brief dips suggest occasional market corrections, but the stock consistently rebounds, reflecting strong market fundamentals and investor confidence.

Fig. 1 Google Closing Price

Besides, the Google Daily Returns chart provides several insights into the stock's volatility. Most daily returns are centered around 0%, suggesting that the stock experiences relatively minor fluctuations on typical trading days. How-

ever, there are significant spikes, particularly in 2014 and 2016, where daily returns exceeded 10%, indicating sharp movements likely driven by market or company-specific events. Although the majority of daily changes fall within ±5%, these spikes highlight periods of increased volatility in Google's stock performance. The distribution of daily returns shows a sharp peak around 0%, indicating that small fluctuations are the norm. The distribution is slightly right-skewed, with a long tail extending towards higher positive returns, suggesting occasional large gains, while negative returns are less extreme. Overall, the returns are largely centered around zero, reflecting stable daily movements with some instances of larger variations.

3.2 Stationarity Check and Log Transformation

The outcomes of the Augmented Dickey-Fuller (ADF) test are displayed in Table 1. Since the p-value is much greater than 0.05, the non-stationarity null hypothesis cannot be ruled out. Furthermore, at every significance level, the ADF statistic exceeds the critical values, indicating that the time series is non-stationary. This implies that before using a forecasting model like ARIMA, adjustments like logarithmic transformation and differencing are required to stabilize the mean and make the series stationary.

Table 1. ADF Test Results

ADF Statistic:	-0.540057
p-value:	0.883960
Critical Value (1%):	-3.4355671297788666
Critical Value (5%):	-2.8638438984080117
Critical Value (10%):	-2.5679966213893057

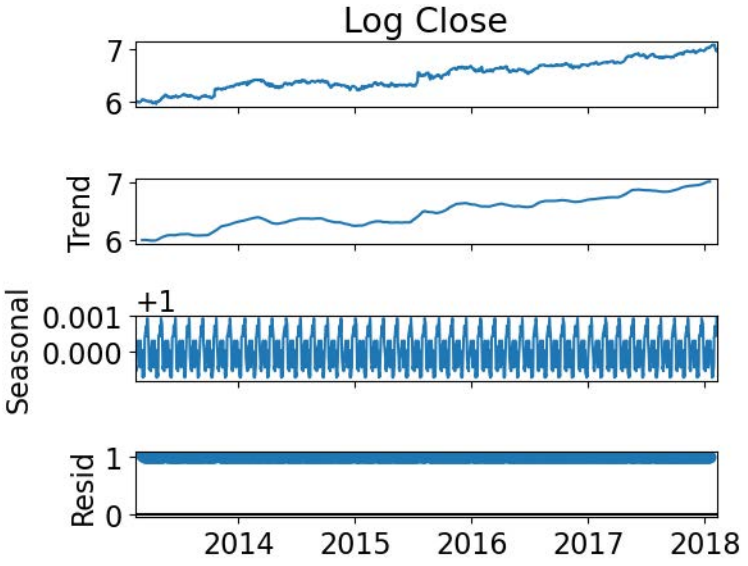

Fig. 2 Decompose Log-transformed Closing Price

For the benefit of clarity, the author will briefly introduce some key concepts in the following sections, as presented in Fig. 2. The top plot in Fig. 2 shows the log-transformed closing prices, which stabilizes the variance and makes the overall trend more apparent. The second plot highlights the underlying trend in the log-transformed closing prices, revealing a consistent upward movement throughout the analyzed period, with slight fluctuations observed around 2015 and 2016. This trend reflects the general growth in Google's stock value over time. The third plot captures the seasonal component, indicating periodic, repeating patterns. Although the seasonal variations are relatively small (around 0.001), they suggest a degree of regularity, which could be attributed to recurring market or trading patterns. The bottom plot represents the residual component, which shows the noise or randomness in the data after removing the trend and seasonal effects. The flat distribution of the residuals indicates that the decomposition model has effectively captured the significant components, leaving minimal unexplained variance. Overall, Fig. 2 illustrates that Google's stock prices follow a strong upward trend, with only minor seasonal effects, and that the residuals

ISSN 2959-6157

are relatively stable, meaning the trend and seasonality account for most of the variability in the data.

3.3 ARIMA Model for Time Series Forecasting

The Google stock price forecast produced by an ARIMA model is shown in Fig. 3, which also shows the comparison between the anticipated and real stock prices over time. The training data, which comprises of past stock prices used to create the ARIMA model, is represented by the blue line. The overall rising trend of Google's stock from 2013 till late 2017 is represented by this data. The yellow line shows the model's projected values for the same time period, and the orange line shows the actual stock prices during the test period. The ARIMA model successfully captures the overall trend in the data, as seen by the model's predictions closely matching the actual stock values. The shaded grey area depicts the confidence interval, which serves as an indication of the uncertainty associated with the model's predictions. As expected, the confidence interval widens over time, reflecting the growing uncertainty as the forecast horizon extends further. Overall, Fig.FF 3 demonstrates that the ARIMA model performs well in forecasting Google's stock prices, with predicted values largely consistent with actual outcomes. The widening of the confidence interval over time highlights the challenges of long-term stock price forecasting, but the relatively narrow range indicates a strong model fit for this particular dataset.

Fig. 3 Google Stock Price Prediction using ARIMA

Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) are among the error metrics for the ARIMA model that was used to forecast Google's stock prices that are shown in Table 2. With only small differences between expected and actual values, the model worked well, as shown by the MSE value of 0.0029, the MAE of 0.0423, and the RMSE of 0.0542. These measures show that the underlying trend in the Google stock data is well captured by the ARIMA model, enabling accurate forecasts. Despite the NaN value obtained from the MAPE, which could have been caused by division by zero, the remaining error metrics validate the general accuracy of the model. This analysis demonstrates that the ARIMA model fits the data quite well, demonstrating its efficacy in predicting Google's stock price.

Table 2: Error Metrics for ARIMA Model

MSE:	0.002935315749945799
MAE:	0.04229100192231715
RMSE:	0.054178554336063625
MAPE:	NaN

Fig. 4 presents diagnostic plots for evaluating the ARIMA model's residuals. The residuals fluctuate around zero without showing any clear patterns or trends, suggesting that the model has effectively captured most of the information in the time series. However, a few spikes indicate the presence of outliers or periods of increased volatility. The histogram of residuals, along with the kernel density estimate (KDE), shows that the distribution deviates slightly from a normal distribution. While centered around zero, the residuals display some skewness, and the tails are heavier than expected under a normal distribution, suggesting some non-normality that could impact model

assumptions. The residuals do not exactly follow a normal distribution, as the Q-Q plot shows departures from the red line, especially at the extremities. The tails show the biggest disparities, which suggests that the data may contain outliers or other non-normal features. All lags fall within the confidence ranges, and the autocorrelation plot indicates that there are no significant autocorrelations in the residuals. This shows that the model has successfully eliminated autocorrelation from the time series, as seen by the residuals' resemblance to white noise. Overall, Fig. 4 indicates that although the residuals show some indications of non-normality, particularly in the Q-Q plot and the heavier tails of the histogram, these deviations are not significant. The model remains suitable for forecasting purposes, as it effectively captures the underlying patterns and reduces autocorrelation in the residuals.

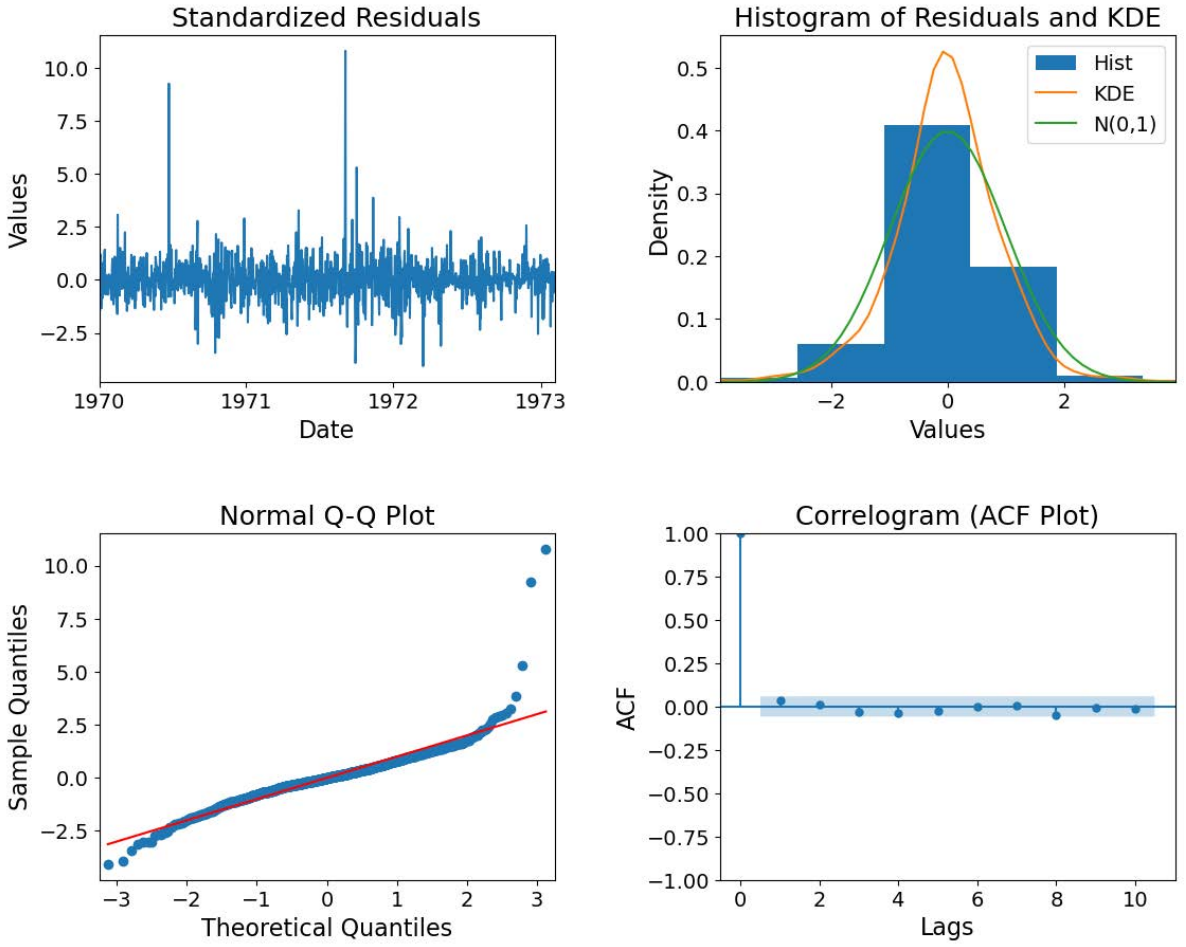

Fig. 4 Diagnostic plots for different quantities.

3.4 Linear Regression for Daily Return Analysis

Fig. 5 illustrates the comparison between predicted and actual daily returns for Google stock. The actual returns show considerable volatility, with significant fluctuations around zero, indicating notable day-to-day variability in returns. Such high volatility is typical in stock markets, where external factors can lead to rapid changes in prices. In contrast, the predicted returns exhibit a much smoother pattern, with less volatility and a tendency to remain close to zero. This suggests that the model effectively captures the overall direction of returns but struggles to account for the high degree of variability observed in the actual data. While the model does well in predicting the general trend of daily returns, it falls short in capturing the short-term volatility present in the actual returns. The predicted values remain relatively stable near zero, whereas the actual returns show more pronounced fluctuations, indicating that the model might be better suited for long-term trend forecasting rather than accurately predicting short-term variations in daily returns.

ISSN 2959-6157

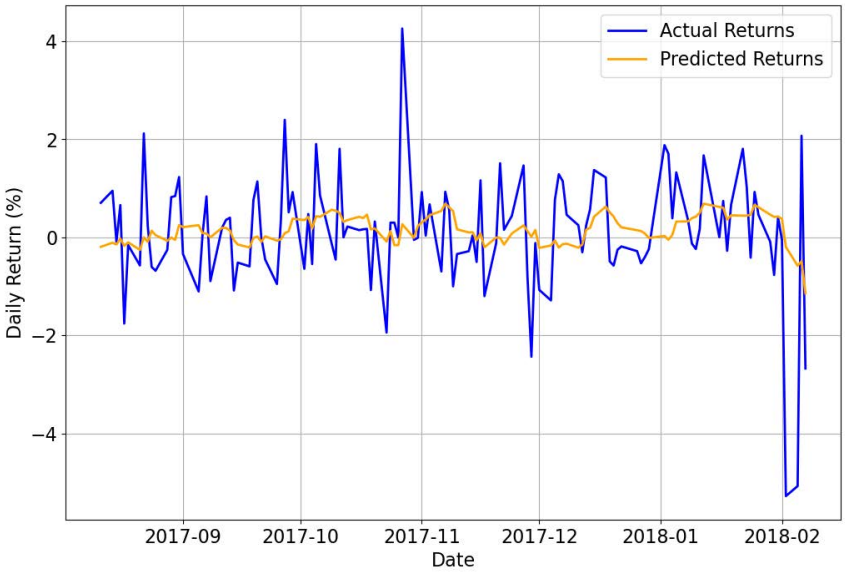

Fig. 5 Predicted vs Actual Daily Returns

Now the author performs a residual analysis to assess the linear regression model's performance in predicting daily returns and to validate the underlying assumptions of the model, providing additional insights to complement the previous evaluation. The residual plot displays the relationship between predicted daily returns (x-axis) and residuals (y-axis), which represent the differences between actual and predicted values. Most of the residuals are clustered around zero, suggesting that the model generally produces small and evenly distributed errors. However, as predicted returns increase, the residual spread widens, indicating greater variance and reduced accuracy for higher values. This suggests that while the model effectively captures smaller returns, it struggles to accurately predict larger or more volatile returns, revealing inefficiencies in capturing the extremes of the data. Overall, the residual plot helps identify the limitations of the model, indicating that it is more suitable for general predictions than for capturing high volatility in daily returns.

4. Conclusion

This research explores the use of statistical models, specifically ARIMA and linear regression, to predict stock market trends and daily returns. By conducting an in-depth analysis of Google's stock price data, the study reveals both the strengths and limitations of these models in capturing the inherent dynamics and volatility of financial markets. The ARIMA model is shown to be effective at identifying overall trends and mitigating short-term fluctuations in stock prices, making it well-suited for long-term forecasting. Conversely, the linear regression model exhibits mixed performance when predicting daily returns, as reflected by increased residual errors for more volatile periods. These findings suggest that while ARIMA and linear regression offer valuable insights into stock movement patterns, they may fall short in addressing the complexities of extreme market conditions. In the case of Google, the ARIMA model effectively captures the company's long-term growth trend, reflecting its stable market position and consistent performance over the analyzed period. However, the inability of linear regression to accurately predict the more volatile daily returns highlights the challenges of forecasting short-term movements for a tech giant like Google, whose stock is influenced by a variety of factors, including market sentiment, regulatory changes, and rapid innovation. The fluctuating residuals observed in the daily return predictions emphasize the need for more sophisticated modeling techniques to better account for these dynamic influences on Google's stock price.

In order to improve model robustness and forecast accuracy, future research could address these constraints by including other variables, such as macroeconomic data or market sentiment indicators. Furthermore, a deeper understanding of machine learning techniques like Long Short-Term Memory (LSTM) networks and other deep learning architectures may be able to better capture the non-linear correlations and temporal dependencies seen in financial data. Leveraging the advantages of several prediction frameworks, extending the scope to include ensemble techniques or hybrid models may also improve the overall performance. In the end, these improvements are required to make predictive models more responsive to the dynamic and intricate structure of financial markets, which will help to improve the quality and accuracy of decisions

made regarding investment strategies and financial planning.

References

[1] Shah D, Isah H, Zulkernine F. Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. International Journal of Financial Studies. 2019, 7(2):26.

[2] Hu Z, Zhao Y, Khushi M. A Survey of Forex and Stock Price Prediction Using Deep Learning. Applied System Innovation. 2021, 4(1):9.

[3] Rouf N, Malik MB, Arif T, Sharma S, Singh S, Aich S, Kim H-C. Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. Electronics. 2021, 10(21): 2717.

[4] Alabdulrazzaq H, Alenezi MN, Rawajfih Y, Alghannam BA, Al-Hassan AA, Al-Anzi FS. On the accuracy of ARIMA based prediction of COVID-19 spread. Results in Physics. 2021, 27: 104509.

[5] Fan D, Sun H, Yao J, Zhang K, Yan X, Sun Z. Well production forecasting based on ARIMA-LSTM model considering manual operations. Energy. 2021, 220: 119708.

[6] Liu S, Lu M, Li H, Zuo Y. Prediction of Gene Expression Patterns With Generalized Linear Regression Model. Frontiers in Genetics. 2019, 10.

[7] Ciulla G, D'Amico A. Building energy performance forecasting: A multiple linear regression approach. Applied Energy. 2019, 253: 113500.

[8] Maaouane M, Zouggar S, Krajačić G, Zahboune H. Modelling industry energy demand using multiple linear regression analysis based on consumed quantity of goods. Energy. 2021, 225: 120270.

[9] Patel J, Shah S, Thakkar P, Kotecha K. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications. 2015, 42(1):259-268.

[10] Liang H, Li J, Di Y, Zhang A, Zhu F. Logarithmic Transformation is Essential for Statistical Analysis of Fungicide EC50 Values. Journal of Phytopathology. 2015, 163(6):456-464.