Improving the accuracy of Stock Price Prediction: A Comparative Study of Statistical models and Machine Learning Algorithms with application to Panel Data

By

Dr. Heba Mahmoud Elsegai

Lecturer of Statistics

Faculty of Commerce, Mansoura University

dr.heba.elsegai@mans.edu.eg

Hanem Salah El-Metwally

Master Researcher

Higher institute of administrative science El-Menzalah Egypt

hanemsalahsoker@gmail.com

Dr. Ramy Mohamed Tayea

Lecturer of Statistics

Higher institute of administrative science El-Menzalah Egypt

ramy.tayea@yahoo.com

Scientific Journal for Financial and Commercial Studies and Research (SJFCSR)

Faculty of Commerce – Damietta University

Vol.6, No.2, Part 1., July 2025

APA Citation

Elsegai, H. M.; **El-Metwally**, H. S. and **Tayea**. R. M. (2025). Improving the accuracy of Stock Price Prediction: A Comparative Study of Statistical models and Machine Learning Algorithms with application to Panel Data, ***Scientific Journal for Financial and Commercial Studies and Research***, Faculty of Commerce, Damietta University, 6(2)1, 627-676.

Website: https://cfdj.journals.ekb.eg/

Improving the accuracy of Stock Price Prediction: A Comparative Study of Statistical models and Machine Learning Algorithms with application to Panel Data

Dr. Heba Elsegai; Hanem Salah and Dr. Ramy Tayea

Abstract:

This study conducts a comparative analysis of random effects panel models and random forest algorithms in stock price prediction, exploring the trade-off between econometric interpretability and machine learning's predictive performance. The research contributes to ongoing discussions regarding financial modeling strategies that balance accuracy with explanatory power.

Using panel data from stock markets, we apply two methodological approaches: (1) random effects models to account for unobserved heterogeneity, and (2) random forests to capture complex nonlinear patterns. Model performance is evaluated using mean squared error (MSE) and R^2 metrics, alongside assessments of computational efficiency, data requirements, and interpretability.

The results indicate that random forests achieve marginally superior predictive accuracy in certain scenarios, whereas random effects models retain advantages in interpretability and robustness, particularly in modeling heterogeneity. These findings highlight the inherent tension between predictive power and transparency in financial analytics.

The study demonstrates that random effects models remain a valuable tool for stock price prediction, despite the slight accuracy gains offered by machine learning techniques. Each approach exhibits distinct strengths: statistical models provide clearer economic insights, while algorithmic methods excel in predictive performance.

For researchers and practitioners, we propose a selection framework based on analytical priorities. When interpretability is critical, random effects models are preferable. Conversely, when maximizing predictive accuracy is the primary objective—and sufficient computational resources are available—random forests may be more suitable. The optimal choice depends on research objectives and dataset characteristics, with our findings offering empirically grounded guidance.

Key Words: Stock market prediction for Panel Data; Random Forest; Fixed Effects Model; Random Effects Model; Accuracy Measures,

Introduction:

Stock price prediction remains a pivotal area of financial research, driven by its critical role in investment decision-making, risk assessment, and portfolio management. Traditional econometric approaches, particularly Fixed Effects Models (FEM) and Random Effects Models (REM), have been extensively employed in panel data analysis due to their capacity to address unobserved heterogeneity and temporal dependencies (Vijayarani et al., 2020). These models effectively capture entity-specific and time-specific variations, making them particularly suitable for financial market data. However, their dependence on linear specifications and predetermined functional forms often constrains their ability to model the complex, non-linear relationships characteristic of stock price movements.

The advent of machine learning techniques has introduced new paradigms in financial forecasting, with Random Forest (RF) algorithms emerging as particularly promising alternatives. Developed by Breiman (2001), RF algorithms demonstrate notable advantages in handling high-dimensional data, capturing non-linear patterns, and mitigating overfitting through ensemble learning. These characteristics make RF especially appropriate for analyzing the noisy and volatile nature of financial markets, where traditional statistical models may underperform.

This study contributes to the literature by conducting a rigorous comparative analysis of these competing methodologies for stock price prediction. We systematically evaluate the performance of FEM, REM, and RF algorithms in addressing the distinctive challenges of panel data, including cross-sectional and temporal heterogeneity, multicollinearity, and data incompleteness. Through comprehensive empirical testing on an extensive stock price dataset, we aim to delineate the relative strengths and limitations of each approach, thereby providing actionable insights for financial forecasting applications (Royo & Guijarro, 2020).

Our methodological framework is specifically designed to assess these models' capabilities in handling panel data complexities. Following established practices in panel data analysis (Stone et al., 1954), we employ a robust dataset encompassing multiple entities across various time periods. The FEM implementation controls for time-invariant entity-specific characteristics, while the REM accommodates both time-varying and time-invariant unobserved factors (Raudenbush et al., 2002). The RF approach,

by contrast, utilizes its ensemble-based architecture to model intricate predictor relationships without restrictive parametric assumptions. Model performance is evaluated using standard metrics including Mean Absolute Error (MAE) and R-squared values (Cohen, 1988), ensuring a balanced and methodologically sound comparison.

The anticipated contributions of this research are threefold. First, it advances academic discourse by providing a systematic evaluation of competing prediction methodologies. Second, it offers practical guidance for financial analysts and researchers in selecting appropriate modeling techniques. Third, it highlights the ongoing tension between model interpretability and predictive power in financial analytics. Our findings underscore the necessity of judiciously balancing traditional econometric approaches with modern machine learning techniques to enhance prediction accuracy in today's data-intensive financial environment.

Data:

This study employs both statistical and machine learning approaches to analyze two comprehensive datasets of Egyptian real estate sector stocks. The analysis encompasses:

1. **Daily Data**: 31,010 observations per company
2. **Monthly Data**: 1,669 observations per company

Data Sources and Composition:

The datasets were systematically collected from Investing.com's Egyptian market portal (https://www.investing.com/markets/egypt), covering a 10-year period from 2013 to 2022. The sample includes 14 prominent real estate companies listed on the Egyptian Exchange:

1. El Taamir for Engineering Consultations
2. Gulf Canadian Real Estate Investment
3. SCC for Contracting and Real Estate Investment
4. Amer Group Holding
5. Global Investment and Development
6. United Housing and Development
7. Egyptians for Housing, Development, and Construction

8. National Housing for Professional Syndicates
9. Palm Hills Development
10. Recap Financial Investments
11. Zahraa El Maadi for Investment and Development
12. Talaat Moustafa Group Holding
13. New Cairo for Housing and Development
14. Mina for Tourism and Real Estate Investment

Predictive Framework:

The study focuses on forecasting closing stock prices using the following explanatory variables:

- Previous period's closing price (lagged)
- Opening price
- Daily/monthly high price
- Daily/monthly low price
- Company valuation metric (from preceding day/month)

This temporal autoregressive structure captures essential price dynamics while maintaining computational tractability. The selected variables represent fundamental market indicators that typically influence stock price movements in the real estate sector.

Methodological Considerations:

The dual timeframe analysis (daily and monthly) enables examination of both short-term market fluctuations and longer-term trends. The 10-year window encompasses various market conditions, including periods of economic stability and volatility, thereby enhancing the robustness of our findings.

Methodology

Panel Data Framework

This study employs panel data methodology to analyze the stock price dynamics of Egyptian real estate companies. Panel data, alternatively termed longitudinal data, offers a robust analytical framework by combining:

1. Cross-sectional dimension: Observations across multiple entities (14 real estate companies)
2. Time-series dimension: Daily and monthly observations spanning 2013-2022

The dual-dimensional structure enables simultaneous examination of:

- Entity-specific heterogeneity (company-level variations)
- Temporal patterns (market dynamics over time)
- Interaction effects between cross-sectional and time-series components

Mathematical Representation

The panel data structure can be formally expressed as (Semykina & Wooldridge, 2010):

$$Y\_it = \alpha + \beta X\_it + \mu\_i + \lambda\_t + \varepsilon\_it \quad (1)$$

Where:

- Y_it: Dependent variable (stock price) for entity i at time t
- X_it: Vector of explanatory variables
- μ_i: Entity-specific effects (unobserved heterogeneity)
- λ_t: Time-specific effects
- ε_it: Idiosyncratic error term

Analytical Advantages

The panel data approach particularly benefits financial market analysis by (Stone et al., 1954),:

1. Controlling for unmeasured company-specific characteristics
2. Capturing temporal market trends and cycles
3. Mitigating omitted variable bias through double dimensionality

4. Allowing for more nuanced modeling of dynamic processes

Data Structure Characteristics

Our dataset exhibits the following panel data features:

- Balanced structure (equal time periods for all entities)
- High-frequency temporal dimension (daily observations)
- Lower-frequency validation (monthly observations)
- Complete cases across all variables of interest

$$D = \{(x_{it}, y_{it}) \mid i = 1,2, \dots, N; t = 1,2, \dots, T\} \quad (2)$$

where:

- x_{it} is the feature vector for entity i at time t,
- y_{it} is the target variable for entity i at time t,
- i represents the entity (i.e., company),
- t represents the time period.

The primary objective of this study is to develop predictive models for y_it (stock prices) using x_it (predictor variables) while properly accounting for the inherent dependencies in panel data. As noted by Wooldridge (2003), panel data analysis presents three fundamental challenges that require careful methodological consideration:

1. **Temporal Dependence**: Autocorrelation in observations for individual entities across time periods
2. **Cross-sectional Dependence**: Correlation among entities due to common market influences or shocks
3. **Unobserved Heterogeneity**: Entity-specific characteristics that affect outcomes but are not directly measured.

To address these analytical challenges, we implement a comprehensive methodological framework that integrates both traditional econometric techniques and modern machine learning approaches for panel data analysis. Building upon the foundational

work of Cohen (1988), our study employs Machine Learning Approaches and Econometric Panel Data Models.

(1) Machine Learning algorithms:

Machine Learning Framework for Panel Data Analysis

Machine learning, a core domain of artificial intelligence, develops systems that autonomously improve their predictive performance through data-driven learning rather than explicit programming (Weinblat, 2018).

While these techniques have demonstrated remarkable success across diverse applications - including computer vision, natural language processing, and financial forecasting - our investigation focuses specifically on the Random Forest algorithm due to its particular suitability for panel data analysis (Weinblat, 2018).

Rationale for Algorithm Selection:

The Random Forest approach offers several advantages for financial panel data:

1. **Native Handling of Complex Dependencies**: Capable of capturing nonlinear relationships and interaction effects without restrictive parametric assumptions
2. **Robustness to Overfitting**: Ensemble-based structure mitigates variance through bootstrap aggregation
3. **Panel Data Adaptability**: Can incorporate temporal dynamics through:
    - Lagged variable engineering
    - Entity-specific fixed effects encoding
    - Time-series aware cross-validation schemes (Weinblat, 2018)

- Random Forest Approach:

Ensemble learning methods, such as Random Forests, are made up of a set of classifiers—e.g. decision trees—and their predictions are aggregated to identify the most popular result. The most well-known ensemble

methods are bagging, also known as bootstrap aggregation, and boosting (Bhardwaj & Ansari, 2019).

Historical Development and Core Concepts

The methodological foundation of this study traces back to Leo Breiman's seminal work on bagging (bootstrap aggregating). This ensemble technique involves:

1. Generating multiple bootstrap samples (with replacement) from the training data
2. Training independent models on each resampled dataset
3. Aggregating predictions through averaging (regression) or majority voting (classification)

Algorithmic Architecture

Random Forest (Breiman, 2001) extends this paradigm by constructing an ensemble of decision trees with two key modifications:

- Feature randomness (random subspace method)
- Complete decorrelation of individual trees

The algorithm's widespread adoption stems from three principal advantages (Liu et al., 2012):

1. **Robustness**: Reduced variance through ensemble averaging
2. **Versatility**: Native handling of both continuous and categorical variables
3. **Regularization**: Built-in protection against overfitting via:
    - Out-of-bag error estimation
    - Feature importance thresholds
    - Maximum depth constraints

Mathematical Framework for Applying Random Forests to Panel Data

While the core Random Forest (RF) algorithm remains consistent when applied to panel data, the preparation and adaptation of the data are crucial. Below is a detailed outline of the process tailored for regression tasks (Liu et al., 2012):

1. Problem Formulation

Consider a panel dataset comprising N entities (i.e., companies) observed over T time periods. Let:

- y_{it} denote the dependent variable for entity i at time t.
- $\boldsymbol{x}_{it}$= (x_{it}^1, x_{it}^2,…, x_{it}^p) represent a p-dimensional vector of predictor variables for entity i at time t.
- $\boldsymbol{X} = \{x_{it}\}_{i=1,t=1}^{N,T}$ be the full set of predictor variables.
- $y = \{y_{it}\}_{i=1,t=1}^{N,T}$ be the full set of observed outcomes.

The objective is to model the relationship between **X** and y, capturing both cross-sectional and temporal dependencies inherent in the panel data.

2. Core Random Forest Algorithm

Random Forests are an ensemble learning method that constructs multiple decision trees and aggregates their predictions. For regression tasks, the algorithm proceeds as follows:

(a) Sampling (Bagging):

For each tree k ($k = 1,2, \dots, K$), draw a bootstrap sample D_k from the original dataset $D = \{(\boldsymbol{X}_{it} y_{it})\}$.

(b) Decision Tree Construction:

For each bootstrap sample D_k, grow a decision tree $Tree_k$ by partitioning the data, recursively:

At each node, select a random subset of m features:

- Choose the feature and split point that minimizes the mean squared error (MSE), a common criterion for regression tasks.
- Repeat until a stopping criterion is met (e.g., maximum tree depth or minimum samples per leaf).

(c) Prediction Aggregation:

For regression, the final prediction $\hat{y}_{it}$ is the average of predictions from all K trees:

$$\hat{y}_{it} = 1/y \sum_{k=1}^{K} \hat{y}_{it}^{k}, \qquad (3)$$

where,

- $\hat{y}_{it}$ is the final predicted value for entity i at time t.
- K is the total number of trees in the Random Forest.
- $\hat{y}_{it}^{k}$ is the predicted value for entity i at time t from the k-th tree (i.e., the prediction of the k-th decision tree in the Random Forest), and is presented as:

  $\hat{y}_{it}^{k} = Tree_k(x_{it}, x_{it-1}, x_{it-2}, \dots)$ (Liu et al., 2012).

3. Adaptation of Random Forests for Panel Data

To account for the temporal and cross-sectional structure of panel data, the following adaptations are implemented:

(a) Feature Engineering:

- Incorporate temporal dependencies by creating lagged variables. For example:

  $\boldsymbol{x}_{it}$=($x_{it}^{1}, x_{it-1}^{1}, x_{it-2}^{1}, \dots, x_{it}^{2}, x_{it-1}^{2}, x_{it-2}^{2}, \dots$).
- Include entity-specific fixed effects by adding dummy variables for each entity *i.*
- Ensure each observation corresponds to an entity *i* at time *t*, with lagged variables and fixed effects included as additional features.

(b) Data Structure:

This framework leverages the Random Forest algorithm's ability to handle complex datasets and reduce overfitting, making it particularly suitable for regression tasks in panel data settings. By incorporating temporal dependencies and entity-specific effects, the model captures the unique structure of panel data, enhancing its predictive performance.

(1) Statistical Methods:

Analyzing panel data requires specialized statistical models to account for both cross-sectional and temporal variations. Two widely used approaches are the Fixed Effects Model (FEM) and the Random Effects Model (REM). These models differ in how they handle entity-specific characteristics and their assumptions about the relationship between these characteristics and the independent variables.

(a) Fixed Effects Model (FEM)

The Fixed Effects Model (FEM) accounts for entity-specific, time-invariant characteristics by including entity-specific intercepts (dummy variables). This approach assumes that any unobserved heterogeneity is constant over time and correlated with the independent variables (Honore,١٩٩٨).

Mathematically, the FEM is specified as:

$$y_{it} = \alpha_i + \boldsymbol{\beta x_{it}} + \varepsilon_{it}, \qquad (4)$$

where:

- y_{it} is the dependent variable for entity *i* at time *t*.
- α_i represents the entity-specific fixed effect, capturing time-invariant characteristics of entity i.
- $\boldsymbol{\beta}$ is the coefficient vector for the independent variables.
- $\boldsymbol{x_{it}}$ is a vector of independent variables for entity *i* at time *t*.
- ε_{it} is the error term, assumed to be independently and identically distributed (*i.i.d.*) with mean zero and constant variance.

The Key Features of this model are:

- The entity-specific fixed effects αi control for all time-invariant characteristics, whether observed or unobserved.
- FEM eliminates bias caused by unobserved heterogeneity that is correlated with the independent variables.
- The model is estimated to use techniques such as the Least Squares Dummy Variable (LSDV) approach or the within transformation.

(b) Random Effects Model (REM)

The Random Effects Model (REM) treats entity-specific effects as random variables, assuming they are uncorrelated with the independent variables. This approach is more efficient than FEM when the assumption holds, as it uses both within-entity and between-entity variation (Kmenta, et al., 1986).

The REM is specified as:

$$y_{it} = \alpha + \boldsymbol{\beta x_{it}} + u_i + \varepsilon_{it}, \qquad (5)$$

where:

- y_{it} is the dependent variable for entity i at time t.
- α is the overall intercept.
- $\boldsymbol{\beta}$ is the coefficient vector for the independent variables.
- $\boldsymbol{x}_{it}$ is a vector of independent variables for entity *i* at time *t*.
- u_i represents the entity-specific random effect, assumed to be *i.i.d.* with mean zero and constant variance σ_u^2.
- ε_{it} is the error term, assumed to be *i.i.d.* with mean zero and constant variance σ_ϵ^2.

Equation (4) can be re-written as:

$$y_{it} = \alpha + \boldsymbol{\beta x_{it}} + W_{it}, \qquad (6)$$

where, W_{it} is the composite error term such that: $W_{it} = u_i + \varepsilon_{it}$, which represents two components:

- ε_{it} represents the cross-sectional error component, which is specific to each cross-sectional unit (e.g., individual companies).
- u_i represents the time-series error component, which arises from combining time-series data with cross-sectional data.

In addition, the Expected Value and the Variance of W_{it} are such that: E(W_{it})=0, Var(W_{it})=σ_u^2+σ_ϵ^2.

We can summarize the Key Features of this model as follows:

- The entity-specific random effects u_i are uncorrelated with the independent variables $\boldsymbol{x}_{it}$.
- REM is more efficient than FEM because it uses both within-entity and between-entity variation.
- The model can be estimated using Generalized Least Squares (GLS) or Maximum Likelihood Estimation (MLE) or Ordinary

Least Squares (OLS). In the context of this manuscript, the Generalized Least Squares (GLS) method is the preferred estimation technique in this context, as it accounts for the correlation structure in the error terms and provides efficient estimates (Wooldridge, et al., 2003)

Here, we use both models as this leads to Improved Predictive Accuracy, due to:

- **FEM:** Captures entity-specific trends and patterns, which are crucial for predicting stock prices of individual firms.
- **REM:** Incorporates both within-entity and between-entity variations, providing a more generalized prediction across the entire dataset.
- Comparing predictions from FEM and REM allows to identify which model better fits the data.
- Then, we combine insights from both models to improve the overall predictive accuracy (Taylor, etal ,1980) .

Empirical Results:

(1) For Monthly stock market price data:

In this section, we perform both machine learning algorithms as well as statistical approaches to the monthly stock price dataset.

(a) Machine Learning Algorithms - Random Forest Approach:

Prior to analyzing the collected raw data, it is critical to conduct data preprocessing, as several challenges must be resolved to ensure accurate interpretation and analysis using Python version 3.11. This includes managing missing values by identifying and handling null or incomplete entries to maintain data integrity and quality, as well as converting data types to suitable formats that align with the requirements of specific operations or analytical procedures. Additionally, eliminating irrelevant data by removing columns or values that do not meaningfully contribute to the analysis helps streamline the dataset, enhance its manageability and focusing on the most relevant information. This preprocessing phase is vital for ensuring the precision, dependability, and effectiveness of subsequent data analysis and modeling efforts.

Furthermore, the problem related to the date is resolved by converting the date data into a time series format. Columns containing null values are removed, and the dataset is cleaned of irrelevant entries. In Table 1, we present the data after this preprocessing step, as well as the closing column is relocated to the last position to facilitate easier partitioning and analysis.

Table (1)

Preprocessed Monthly Dataset after rearranging and moving the closing price column to the last column position

	DATE	OPEN	HIGH	LOW	SIZE	VARIANCI	LAST_PRICE
0	012022/12	25.04	25.75	20.00	224000.0	-0.0712	23.22
1	012022/11	22.61	26.50	21.01	187000.0	0.0941	25.00
2	012022/10	21.00	24.19	18.51	145000.0	0.0876	22.85
3	012022/09	24.06	24.10	20.40	105000.0	-0.1365	21.01
4	012022/08	21.95	31.50	20.06	244000.0	0.1253	24.33
...	...	...	...	...	...	...	...
1665	012013/05	0.71	0.77	0.66	877000.0	0.0000	0.69
1666	012013/04	0.68	0.76	0,66	644000.0	0.0122	0.70
1667	012013/03	0.81	0.83	0.67	764000.0	-0.1633	0.69
1668	012013/02	0.85	0.89	0.80	757000.0	-0.0101	0.82
1669	012013/01	0.92	0.96	0.77	1174000.0	-0.0748	0.83
1669 rows × 7 columns							

Table (2)

Preprocessed Monthly Dataset after a column of predicting the next month is added

	OPEN	HIGH	LOW	SIZE	VARIANCE	Close_plus_t
0	25.04	25.75	20.00	224000.0	-0.0712	23.20
1	22.61	26.50	21.01	187000.0	0.0941	23.22
2	21.00	24.19	18.51	145000.0	0.0876	25.00
3	24.06	24.10	20.40	105000.0	-0.1365	22.85

4	**21.95**	**31.50**	**20.06**	**244000.0**	**0.1253**	**21.01**
...	**...**	**...**	**...**	**...**	**...**	**...**
1665	**0.71**	**0.77**	**0.66**	**877000.0**	**0.0000**	**0.51**
1666	**0,68**	**0.76**	**0,66**	**644000.0**	**0.0122**	**0.69**
1667	**0,81**	**0.83**	**0.67**	**764000.0**	**-0.1633**	**0.70**
1668	**0,85**	**0,89**	**0,80**	**757000.0**	**-0.0101**	**0.69**
1669	**0.92**	**0.96**	**0.77**	**1174000.0**	**-0.0748**	**0.82**
1669 rows × 6 **columns**						

In Table 2, a new column of next month's price prediction is added replacing the last price (closing price).

Now, we use the preprocessed data to perform RF analysis. The random forest model comprises numerous decision trees, where the final prediction is based on the average of the results from all the trees. During training, the data is selected randomly, and subsets of variables are chosen at random when splitting nodes. At each node in every decision tree, only one subset of all available variables is used to make the split. Consequently, each decision tree in the random forest is built using a random sample from the dataset, enhancing the model's robustness, and reducing the risk of overfitting. We use the following Random Forest equation:

$$\text{Gini Inder} = \mathbf{1} - \sum_{i=1}^{n} (\boldsymbol{P_i})^2 \qquad (7)$$

$$= 1 - [(\boldsymbol{P_+})^2 + (\boldsymbol{P_-})^2],$$

where,

Gini Index: is a measure used in tree analysis and study in machine learning techniques as it shows the degree of imbalance or variance in the distribution of classes

$\boldsymbol{P_i}$: It is the probability of class i within the data set

$\boldsymbol{P_+}$: It is the probability of the positive class

$\boldsymbol{P_-}$: It is the probability of the negative class

The dataset is, then, partitioned into training and test subsets to facilitate model development and evaluation. The models are trained on the training data, evaluated, and subsequently assessed using performance metrics applied to the test data. In this study, an 80:20 train-test split ratio was employed, where 80% of the dataset (training data) was utilized to train the

classifier, while the remaining 20% (test data) was reserved for final evaluation to ensure robust and reliable model performance.

The results of implementing the Random Forest model are presented in Table 3. The table compares the predicted Y_test values with the actual Y_test values. It is observed that the predicted value for Y_test is (25.5764), which aligns closely with the actual Y_test value in the sample (23.620).

Table (3)

Random Forests prediction Results

Actual (Y_test)	**Predicted (Y_test)**
23.620	25.5674
1.072	0.96003
0.576	0.5784
0.571	0.56733
0.530	0.53115
24.290	49.491
2.140	2.14868
6.980	6.9509
3.390	3.36985
5.680	5.9185

In Table 3, the Random Forest model demonstrates strong performance, with most predicted values closely aligning with the actual values (e.g., 0.5784 vs. 0.576), indicating effective capture of underlying data patterns. However, a few significant errors (e.g., 49.6491 vs. 24.290) suggest potential issues such as outliers, noisy data, or regions where the model struggles to generalize. Despite these outliers, the model is generally reliable and robust, as evidenced by the high accuracy score of 86.57% according to results presented in the following as a Python code format:

In[ ]:

```
RF_score = rfmodel.score(X_test, y_test)

RF_score
```

Out[ ]:

0.865670871315873

This accuracy score is a common metric used to evaluate the performance of classification models, including Random Forest. It measures the proportion of correctly predicted instances (both true positives and true negatives) out of the total number of instances. Therefore, the resulting accuracy score suggests that around 87% (when rounded) of the variance in the target variable can be explained by the model, which is as indication of a strong model performance, especially for many practical applications like stock markets. Furthermore, this level of accuracy indicates that the model is reliable for making predictions on unseen data.

For evaluating Model Quality, we use Mean Absolute Error (MAE) (Wang & Lu, ٢٠١٨). The Mean Absolute Error (MAE) is a commonly used metric to evaluate the performance of regression models. It measures the average absolute difference between the predicted values and the actual values. MAE is easy to interpret and provides a straightforward understanding of the magnitude of errors in the model's predictions. The Mean Absolute Error (MAE) is calculated and the result is given (Python Output) as follows:

mean_absolute_error=0.3630118820224718

This result indicates the average absolute difference between the predicted values and the actual values in the test dataset. An MAE of 0.363 means that, on average, the model's predictions deviate from the actual values by approximately 0.363 units. This is a relatively small error, suggesting that the model is performing well. In general, a lower MAE indicates better predictive accuracy. In this case, the MAE is close to zero, which reflects the model's ability to make predictions that are very close to the true values.

(b) Statistical Regression Models - Fixed Effects Model (FEM) and Random Effects Model (REM):

Prior to analyzing the collected raw data, it is critical to conduct various statistical tests first, to ensure accurate interpretation and analysis using STATA 15.

- Stationarity tests were conducted for the time series, treated as panel data, using the Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root. The test checks for the presence of a unit root in the

data, which implies non-stationarity. Non-stationary data can lead to spurious regression results, making it essential to confirm stationarity before modeling. The Fisher-type ADF test combines p-values from individual ADF tests applied to each cross-sectional unit in panel data. This approach is particularly useful for panel datasets, as it aggregates results across multiple units (e.g., countries, firms, or time series). The Null and Alternative Hypotheses for ADF test are:

- Null Hypothesis (H_0): The data contains a unit root (non-stationary).
- Alternative Hypothesis (H_1): The data does not contain a unit root (stationary).

The test produces a combined p-value based on the individual ADF test results. If the p-value is less than a significance level (e.g., 5%), the null hypothesis is rejected, indicating stationarity. The ADF test is conducted and the results for the 5variables under study are presented in Appendix (1). The results indicate that all time series are stationary, with no presence of a unit root. This is evident from the p-values of all tests, which are less than 5%, leading to the rejection of the null hypothesis of a unit root in the panel data for all study variables. In sum, the Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root is a robust method for testing stationarity in panel data (Johnson & DiNardo, ١٩٩٧). In this study, the test confirmed the stationarity of all time series, providing a solid foundation for further analysis.

- Although ADF test is conducted, we now utilize a panel cointegration test (Campbell & Perron, 1991) to check for spurious regression. The results are presented in Table 4.

Table (4)

Panel cointegration test

.xtcointtest kao close_next close open max min size	
Kao test for cointegration	
Ho: No cointegration	Number of panels = 14
Ha: All panels are cointegrated	Avg. number of periods = 117.29
Cointegrating vector: Same	
Panel means: Included	Kernel: Bartlett

Time trend:	Not included	Lags:	3.14 (Newey-West)
AR parameter:	Same	Augmented lags:	1
		Statistic	p-value
Modified Dickey-Fuller t		-1.2e+02	0.0000
Dickey-Fuller t		-37.9008	0.0000
Augmented Dickey-Fuller t		-19.4662	0.0000
Unadjusted modified Dickey-Fuller t		-1.3e+02	0.0000

The Null and Alternative Hypotheses for the Kao panel cointegration test are:

- Null Hypothesis (H_0): No cointegration exists among the variables.
- Alternative Hypothesis (H_1): Cointegration exists (residuals are stationary).

According to the results, the null hypothesis is rejected as P-value is smaller than 0.05, which means that a cointegration exists. This confirms that a stable, long-term relationship between trending variables exists.

- Normality tests were conducted using Kolmogorov-Smirnov test and Shapiro-Wilk test. These tests are essential for validating assumptions in many statistical analyses, such as regression, ANOVA, and parametric tests. The Null and Alternative Hypotheses for the tests are:
  - Null Hypothesis (H_0): The data follows a normal distribution.
  - Alternative Hypothesis (H_1): The data does not follow a normal distribution.

The null hypothesis is rejected, if the p-value < 0.05, which indicates that the data is not normally distributed, while the null hypothesis cannot be rejected if the p-value ≥ 0.05, suggesting the data may be normally distributed.

The results are shown in Table 5, and it can be noted that the datasets for all variables were found to be non-normally distributed, as determined by the Kolmogorov-Smirnov test, as well as by Shapiro-Wilk test. More precisely, the p-values for both tests were less than 5% for all variables, indicating that they do not follow a normal distribution. However, due to the large sample size in terms of the number of observations (1669 observation) and in accordance with the Central Limit Theorem, the data can be considered approximately normally distributed.

Table (5)The results for Normality tests

(Kolmogorov-Smirnov and Shapiro-Wilk tests)

Tests of Normality						
	Kolmogorov-Smirnov[a]			**Shapiro-Wilk**		
	Statistic	df	Sig.	Statistic	df	Sig.
close	.240	1670	.000	.655	1670	.000
open	.239	1670	.000	.656	1670	.000
max	.244	1670	.000	.641	1670	.000
min	.233	1670	.000	.676	1670	.000
size	.477	1670	.000	.022	1670	.000
close_next	.241	1670	.000	.653	1670	.000
a. Lilliefors Significance Correction						

Table (6) The STATA output result for FEM

.xtreg close_next close open max min value, fe	
Fixed-effects (within) regression	Number of obs = 1,670
Group variable: id	Number of groups = 14
R-sq:	Obs per group:
within = 0.9817	min = 114

between = 0.9991				avg =	119.3	
overall = 0.9924				max =	120	
			F(5,1651)	=	17755.34	
corr(u_i, Xb) = 0.5056			Prob > F	=	0.0000	
---						
close_next	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval[
---						
close	.0993149	.0182249	5.45	0.000	.0635685	.1350613
open	1.017931	.0161903	62.87	0.000	.9861753	1.049687
max	-.0654313	.015992	-4.09	0.000	-.0967979	-.0340646
min	-.0479814	.0219162	-2.19	0.029	-.090968	-.0049949
size	6.53e-07	6.79e-06	0.10	**0.923**	-.0000127	.000014
_cons	.0591652	.0286442	2.07	0.039	.0029825	.1153479
---						
sigma_u	.27112701					
sigma_e	.69990947					
rho	.13047934	(fraction of variance due to u_i)				

F test that all u_i=0: F(13, 1651) = 11.38				Prob > F = 0.0000		

The prediction is, now, made using Fixed Effects Model (FEM) and Random Effects Model (REM) for the regression analysis. Table 6 presents the results for FEM and we interpret the results in the following:

- The coefficients estimate represent the estimated effect of each predictor on close_next, holding other variables constant. We can observe that all variables have a statistically insignificant effect on the next month's closing price. That is EXCEPT the independent variable (size , (6.53e-07, p = 0.923)), where the coefficient is close to zero, and the p-value (0.923) indicates that this variable does not contribute meaningfully to the model.
- The variance component rho (0.1305) indicates that 13.05% of the total variance in close_next is due to company-specific effects, while the remaining variance is due to the idiosyncratic error (sigma_e).
- The overall R^2 (0.9924) indicates that 99.24% of the total variation in close_next is explained by the model.
- The model is statistically significant overall, meaning at least one of the predictors has a significant effect on close_next, due to F-statistic (17755.34, p = 0.0000). Furthermore, the F-test confirms that the company-specific fixed effects are jointly statistically significant.

This justifies the use of the fixed-effects model over a pooled OLS model.

- The Correlation between Fixed Effects and Predictors is such that corr(u_i, Xb) = 0.5056, which indicates a moderate positive correlation (0.5056) between the company-specific effects and the predictors. This suggests that the fixed-effects model is appropriate, as there is evidence of unobserved heterogeneity correlated with the predictors.

Based on the results, we conclude that the fixed-effects regression model provides strong evidence that the opening price, current closing price, maximum price, and minimum price are significant predictors of the next month's closing price. The model is highly effective, with a very high R^2 and statistically significant coefficients for most predictors. However, the variable "size" does not appear to influence close_next in this model. The use of fixed effects is justified, as company-specific heterogeneity plays a significant role in explaining the variation in close_next.

Furthermore, we conduct the analysis again after removing the independent variable (*size*) and the results are shown in Table 7. It can be observed that the Overall R^2 (0.9924), which indicates that 99.24% of the total variation in close_next is explained by the model. This is identical to the result before removing the independent variable (*size*), see Table 6. This means that the exclusion of this value does not affect the model prediction accuracy, however, we can conclude that this variable has no effect on stock price movements.

Table (7)The STATA output result for FEM after removing the independent variable (size)

.xtreg close_next close open max min, fe	
Fixed-effects (within) regression	Number of obs = 1,670
Group variable: id	Number of groups = 14
R-sq:	Obs per group:
within = 0.9817	min = 114
between = 0.9991	avg = 119.3
overall = 0.9924	max = 120
	F(4,1652) = 22207.49
corr(u_i, Xb) = 0.5055	Prob > F = 0.0000

--						
close_next	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval[
---+--------------						
close	.0992964	.0182184	5.45	0.000	.0635627	.1350301
open	1.017911	.0161841	62.90	0.000	.9861678	1.049655
max	-.0654209	.0159868	-4.09	0.000	-.0967775	-.0340644
min	-.0479463	.0219066	-2.19	0.029	-.0909139	-.0049786
cons	.0592331	.0286269	2.07	0.039	.0030843	.1153818
---+--------------						
sigma_u	.27108071					
sigma_e	.69969956					
rho	.13050865	(fraction of variance due to u_i)				
--						
F test that all u_i=0: F(13, 1652) = 11.39				Prob > F = 0.0000		

Next, we conduct analyses of the results for REM, presented in Table 8. We notice that the coefficient estimates for the variable's *min* (0.0086) and *size* (1.33e-07 as both coefficients are close to zero, and the p-values are 0.688 and 0.985, respectively, indicate that these variables do not contribute meaningfully to the model. In addition, overall R^2 (0.9925) indicates that 99.25% of the total variation in close_next is explained by the model. Furthermore, the model is statistically significant overall, as Wald chi^2 is equal to 221639.51with p = 0.0000, meaning at least one of the predictors has a significant effect on close next.

Table (8) The STATA output result for REM

.xtreg close_next close open max min value, fe			
Random-effects GLS regression		Number of obs	= 1,670
Group variable: id		Number of groups	= 14
R-sq:		Obs per group:	
within = 0.9816		min =	114
between = 0.9993		avg =	119.3
overall = 0.9925		max =	120
		Wald chi2(5)	= 221395.18
corr(u_i, X) = 0 (assumed)		Prob > chi2	= 0.0000

close_next	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval[
close	.1101735	.0189014	5.83	0.000	.0731274	.1472196
open	1.011828	.0168235	60.14	0.000	.9788543	1.044801
max	-.0980788	.0162568	-6.03	0.000	-.1299415	-.0662161
min	.0086603	.0215365	0.40	**0.688**	-.0335504	.0508711
size	1.33e-07	7.04e-06	0.02	**0.985**	-.0000137	.0000139
cons	-.0528789	.0224224	-2.36	0.018	-.0968261	-.0089318
sigma_u	0					
sigma_e	.69990947					
rho	0	(fraction of variance due to u_i)				

As a consequence, we re-estimate the model again after removing the independent variables (*min* - *size*) and the results are shown in Table 9. It can be observed that the Overall R^2 (0.9925), which indicates that 99.25% of the total variation in close_next is explained by the model. This is identical to the result before removing the independent variables (*min* - *size*), see Table 8. This means that the exclusion of these variables does not affect the model prediction accuracy, however, we can conclude that these variables have no effect on stock price movements.

Table (9) The STATA output result for REM after removing the independent variables (*min , size*)

.xtreg close_next close open max, fe						
Random-effects GLS regression				Number of obs	=	1,670
Group variable: id				Number of groups	=	14
R-sq:				Obs per group:		
within = 0.9816				min	=	114
between = 0.9993				avg	=	119.3
overall = 0.9925				max	=	120
				Wald chi2(3)	=	221639.51
corr(u_i, X) = 0 (assumed)				Prob > chi2	=	0.0000
close_next	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval[

close	.1157601	.0128045	9.04	0.000	.0906638	.1408565
open	1.015975	.0132774	76.52	0.000	.989952	1.041999
max	-.1001277	.0154255	-6.49	0.000	-.1303612	-.0698942
_cons	-.0515975	.0221566	-2.33	0.020	-.0950236	-.0081713
sigma_u	0					
sigma_e	.70050129					
rho	0	(fraction of variance due to u_i)				

To check which model is more appropriate, Hausman test is conducted (Hausman, 1978). The statistical hypotheses are:

Null Hypothesis (H0): The differences in coefficients between the two models are not systematic (i.e., the random-effects model is preferred).

Alternative Hypothesis (Ha): The differences are systematic (i.e., the fixed-effects model is preferred).

The Hausman test is conducted after removing the variables "*min* - *size*" as suggested, that is according to the insignificant results were obtained in Tables 7 and 8. The results are presented in Table 10.

Table (10) Hausman Test: Random-Effects (RE) vs. Fixed-Effects (FE)

.hausman re fe			
---- Coefficients----			
	(b)	(B)	(b-B) sqrt(diag(V_b-V_B))
	re	fe	Difference S.E.
close	.1101735	.0993149	.0108586 .0050116
open	1.011828	1.017931	-.0061031 .0045724
max	-.0980788	-.0654313	-.0326475 .0029224
corr(u_i, X) = 0 (assumed)		Prob > chi2 =	0.1558
b = consistent under Ho and Ha; obtained from xtreg			
B = inconsistent under Ha, efficient under Ho; obtained from xtreg			
Test: Ho: difference in coefficients not systematic			

chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 13.72
Prob>chi2 = 0.1558
(V_b-V_B is not positive definite)

It can be observed that the p-value of 0.1558 is greater than common significance level 0.05. This means we fail to reject the null hypothesis, which means that there is no evidence of systematic differences between the fixed-effects and random-effects models. Thus, the random-effects model (which is efficient under H0) is preferred for this analysis.

(2) Results for Daily stock market price data:

In this section, we perform both machine learning algorithms as well as statistical approaches to the daily stock price dataset.

(a) Random Forest Approach:

Again, before analyzing the collected raw data, it is essential to perform data preprocessing to address various challenges and ensure accurate interpretation and analysis using Python version 3.11. The same preprocessing procedure utilized before for monthly data is conducted here. Therefore, we display the data after this preprocessing step in Tables 11 and 12 (Elsegai et al.,٢٠٢٥).

Table (11) Preprocessed Daily Dataset (Elsegai et al.,٢٠٢٥).

-	Close	Open	High	Low	Size	Variation	Day	Month	Year
	23.22	22.5	23.3	21.73	152.88K	0.0483	29	12	2022
1	22.15	21.21	22.4	21.05	126.52K	0.0326	28	12	2022
2	21.45	21.37	21.8	21.32	33.58K	0.0056	27	12	2022
3	21.33	21.25	21.38	20.78	18.94K	0.0186	26	12	2022
4	20.94	21.14	21.26	20.72	45.17K	-0.0095	25	12	2022
…	…	…	…	…	…	…..	…..	…..	…..
31026	2.24	2.24	2.28	2.2	423.83K	0.009	9	1	2013
31027	2.22	2.21	2.25	2.17	303.94K	0.0091	8	1	2013
31028	2.2	2.21	2.24	2.18	290.24K	-0.0045	6	1	2013
31029	2.21	2.18	2.24	2.17	645.14K	0	3	1	2013
31030	2.21	2.18	2.25	2.18	487.43K	0.0327	2	1	2013
31010 rows x 9 columns									

Table (12) The closing price prediction using the original data by adding a column (Close + t). (Elsegai et al.2025,).

-	Open	High	Low	Size	Variation	Day	Month	Year	Close	Close_ n_day
0	22.5	23.3	21.73	152.88K	0.0483	29	12	2022	23.22	23.20
1	21.21	22.4	21.05	126.52K	0.0326	28	12	2022	22.15	23.22
2	21.37	21.8	21.32	33.58K	0.0056	27	12	2022	21.45	22.15
3	21.25	21.38	20.78	18.94K	0.0186	26	12	2022	21.33	21.45
4	21.14	21.26	20.72	45.17K	-0.0095	25	12	2022	20.94	21.33
…	…	…	…	…	…..	…..	…..	…..	…	…
31026	2.24	2.28	2.2	423.83K	0.009	9	1	2013	2.24	2.19
31027	2.21	2.25	2.17	303.94K	0.0091	8	1	2013	2.22	2.24
31028	2.21	2.24	2.18	290.24K	-0.0045	6	1	2013	2.2	2.22
31029	2.18	2.24	2.17	645.14K	0	3	1	2013	2.21	2.20
31030	2.18	2.25	2.18	487.43K	0.0327	2	1	2013	2.21	2.21
31010 rows x 10 columns										

Next, the dataset is split into training and test data. Models are then fitted, evaluated, and assessed using metrics on the test data. The study employs an 80:20 train-test ratio, where the larger portion of the dataset (training data) is used to train the classifier.

The results of applying the Random Forest model are presented in Table 13, which compares the predicted Y-test values with the actual Y-test values. It is observed that the predicted value of Y_test is 4.6940514, while the actual Y_test value in the sample is 3.210. Based on the results in Table

13, the accuracy rate of the model is calculated, and the outcome is provided in the following Python code format (Elsegai et al.,٢٠٢٥):

In[ ]:

```
RF_score = rfmodel.score(X_test, y_test)
RF_score
```

Out[ ]:

```
0.632977506046947
```

The result shows that the accuracy rate of the model is 63.4%, which can be considered as a moderate level of accuracy, which is in turn might be relatively acceptable.

Table (13) Random Forest model results (Elsegai et al.2025,).

Actual (Y_test)	**Predicted (Y_test)**
3.210	4.6940514
3.150	4.6940514
5.650	4.6940514
0.491	4.6940514
2.440	4.6940514
.....	.....
9.950	4.6940514
1.700	4.6940514
1.670	4.6940514
9.450	4.6940514
4.080	4.6940514

However, we are having a closer look at Table 13, we observe that all predicted values are identical for all corresponding actual values. Therefore, we run the analysis several times by removing the variables systematically, that is in order to identify the problematic variable. We can, then, conclude that when removing the variables “min - size”, the results show a meaningful prediction as presented in Table 14.

Table (14) Random Forest model results after removing the variables “min - size” from the analysis

Actual (Y_test)	**Predicted (Y_test)**
3.210	4.013
3.150	4.421

5.650	5.14801
0.491	1.3465
2.440	3.26713
.....	.....
9.950	8.7899
1.700	1.96734
1.670	1.89025
9.450	8.4386
4.080	5.24339

In addition, the accuracy rate of the model is, again, calculated, and the outcome is provided in the following Python code format:

In[ ]:

```
RF_score = rfmodel.score(X_test, y_test)
RF_score
```

Out[ ]:

0.749364015438467

The result shows that the accuracy rate of the model is 74.9%, which can be considered reasonably predictive.

(b) Regression models (Fixed Effects Model (FEM) and Random Effects Model (REM))

Before analyzing the collected raw data, it is essential to perform various statistical tests to ensure accurate interpretation and analysis using STATA 15.

- Stationarity tests were conducted for the time series, treated as panel data, using the Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root and the results are shown in Appendix (2). In summary, the results confirm that Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root proves to be a robust method for assessing stationarity in panel data.
- Although ADF test is being conducted, we now utilize a panel cointegration test (Campbell, J. Y., & Perron, P. (1991).) to check for spurious regression. The results are presented in Table 15.

Table (15) Panel cointegration test

.xtcointtest kao close_next close open max min size			
Kao test for cointegration			
Ho: No cointegration		Number of panels	= 14
Ha: All panels are cointegrated		Avg. number of periods	= 2212
Cointegrating vector: Same			
Panel means:	Included	Kernel:	Bartlett

Time trend:	Not included	Lags:	7.64 (Newey-West)
AR parameter:	Same	Augmented lags:	1
		Statistic	p-value
Modified Dickey-Fuller t		-3.8e+02	0.0000
Dickey-Fuller t		-1.1e+02	0.0000
Augmented Dickey-Fuller t		-69.5158	0.0000
Unadjusted modified Dickey-Fuller t		-2.9e+03	0.0000

According to the results, the null hypothesis is rejected as P-value is smaller than 0.05, which means that a cointegration exists. This confirms that a stable, long-term relationship between trending variables exists.

- Normality tests were, again, performed using the Kolmogorov-Smirnov test. The results, presented in Table 16, reveal that the datasets for all variables were found to be non-normally distributed, as indicated by the Kolmogorov-Smirnov test. However, given the large sample size in terms of the number of observations (31010) and in line with the Central Limit Theorem, the data can be treated as approximately normally distributed for analytical purposes.

Table (16) The results for Kolmogorov-Smirnov Normality test

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
close	.217	31010	.000
open	.216	31010	.000
max	.218	31010	.000
min	.215	31010	.000
size	.297	31010	.000
close_next	.217	31010	.000

a. Lilliefors Significance Correction

Next, we show the results of conducting the analysis for FEM presented in Table 17. The interpretation of the results is listed below:

- The coefficient estimates represent the estimated effect of each predictor on close next, holding other variables constant. We can observe that all variables have a statistically insignificant effect on the next month's closing price. That is EXCEPT the independent variable (max, (-0.0033788, p-value = 0.771)), where the coefficient

is close to zero, and the p-value (0.771) indicates that this variable does not contribute meaningfully to the model.

- The Overall R^2 (0.9943) indicates that 99.43% of the total variation in close next is explained by the model.

- The model is statistically significant overall as F-statistics (507529.21, p = 0.0000); meaning that at least one of the predictors has a significant effect on close next. Furthermore, the F-test confirms that the company-specific fixed effects are jointly statistically significant. This justifies the use of the fixed-effects model over a pooled OLS model.

Table (17) The STATA output result for FEM

.xtreg close_next close open max min value, fe						
Fixed-effects (within) regression					Number of obs =	31,010
Group variable: id					Number of groups =	14
R-sq:					Obs per group:	
within = 0.9879					min =	1,093
between = 1.0000					avg =	2,215.0
overall = 0.9943					max =	2,437
					F(5,30991) =	507529.21
corr(u_i, Xb) = 0.5978					Prob > F =	0.0000
close_next	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval[
close	.7324202	.0105638	69.33	0.000	.7117147	.7531258
open	.35466	.011001	32.24	0.000	.3330977	.3762224
max	-.0033788	.0116116	-0.29	**0.771**	-.026138	.0193803
min	-.0911691	.0128593	-7.09	0.000	-.1163739	-.0659642
size	-.0000534	.0000191	-2.79	0.005	-.0000909	-.0000158
_cons	.0560084	.0063362	8.84	0.000	.0435892	.0684275
sigma_u	.09306226					
sigma_e	.67050524					
rho	.01889976	(fraction of variance due to u_i)				
F test that all u_i=0: F(13, 30991) = 13.63					Prob > F =	0.0000

According to these results, we conducted again the analysis for FEM after removing the variable *max* and the results are shown in Table 14. It can be concluded that the Overall R^2 (0.9943), which indicates that 99.43% of the total variation in close next is explained by the model. This is

identical to the result before removing the independent variables (*max*), see Table 18. This means that the exclusion of this variable does not affect the model prediction accuracy, however, we can confirm that this variable has no effect on stock price movements.

Table (18) The STATA output result for FEM after removing the independent variables (*max*)

.xtreg close_next close open min value, fe						
Fixed-effects (within) regression				Number of obs	=	31,010
Group variable: id				Number of groups	=	14
R-sq:				Obs per group:		
within = 0.9879				min	=	1,093
between = 1.0000				avg	=	2,215.0
overall = 0.9943				max	=	2,437
				F(4,30992)	=	634430.22
corr(u_i, Xb) = 0.5975				Prob > F	=	0.0000
---						
close_next	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval[
---						
close	.7311641	.0096414	75.84	0.000	.7122666	.7500616
open	.3530153	.0094378	37.40	0.000	.3345169	.3715137
min	-.0917712	.0126916	-7.23	0.000	-.1166472	-.0668952
size	-.0000535	.0000191	-2.80	0.005	-.000091	-.000016
_cons	.0563402	.0062326	9.04	0.000	.0441239	.0685564
---						
sigma_u	.0936605					
sigma_e	.67049534					
rho	.01913942	(fraction of variance due to u_i(				

F test that all u_i=0: F(13, 30992) = 15.17					Prob > F = 0.0000	

Table (19) The STATA output result for REM

.xtreg close next close open max min value, re						
Random-effects GLS regression					Number of obs =	31,010
Group variable: id					Number of groups =	14
R-sq:					Obs per group:	
within = 0.9879					min =	1,093
between = 1.0000					avg =	2,215.0
overall = 0.9943					max =	2,437
					Wald chi2(5) =	5.39e+06
corr(u_i, X) = 0 (assumed)					Prob > chi2 =	0.0000
close_next	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval[
close	.7716237	.009964	77.44	0.000	.7520946	.7911527
open	.3558558	.0110273	32.27	0.000	.3342427	.377469
max	-.0494123	.0110658	-4.47	0.000	-.0711009	-.0277237
min	-.0796777	.0125803	-6.33	0.000	-.1043346	-.0550209
size	-.0000711	.0000181	-3.93	0.000	-.0001066	-.0000356
_cons	.0239864	.0053867	4.45	0.000	.0134287	.0345442
sigma_u	0					
sigma_e	.67050524					
rho	0	(fraction of variance due to u_i)				

According to the results presented in Table 19, **t**he random-effects GLS regression analysis of panel data with 31,010 observations across 14 groups reveals a highly significant model with excellent explanatory power, as indicated by the overall R-squared of 0.9943. The closing price of the next period (close next) is strongly influenced in a positive way by key indicators such as the current closing price (close) and opening price (open). Conversely, other factors like the maximum price (max), minimum price (min), and overall value show a significant inverse relationship with the next period's closing price, meaning as they increase, the next closing price tends to decrease, and vice versa. The Wald chi-square test confirms the joint significance of the predictors with a p-value of 0.0000. Notably, the random effects do not contribute to the variance (sigma = 0), suggesting that the model might be better specified as a fixed-effects model. The

residual standard deviation (sigma_e) is approximately 0.6705, and the fraction of variance due to random effects (rho) is 0, indicating that all variance is explained by the residual variance. These findings highlight the model's robustness and the significant impact of the predictors on close next.

Lastly, to check which model is more appropriate, Hausman test is conducted (Hausman, 1978) after removing the variables "*max*" as suggested, that is according to the insignificant results were obtained in Table 17. The results are presented in Table 20.

Table (20) Hausman Test: Random-Effects (RE) vs. Fixed-Effects (FE)

.hausman re fe				
----	Coefficients----			
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	re	fe	Difference	S.E.
close	.7324202	.7716237	-.0392034	.003509
open	-.0033788	-.0494123	.0460335	.0035179
min	-.0911691	-.0796777	-.0114913	.0026645
size	-.0000534	-.0000711	.0000177	6.19e-06
corr(u_i, X) = 0 (assumed)			Prob > chi2 =	0.0984
----------				
b = consistent under Ho and Ha; obtained from xtreg				
B = inconsistent under Ha, efficient under Ho; obtained from xtreg				
Test: Ho: difference in coefficients not systematic				
chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)				
= 18.24				
Prob>chi2 = 0.0984				
(V_b-V_B is not positive definite)				

It can be observed that the p-value of 0.0984 is greater than common significance level 0.05. This means we fail to reject the null hypothesis, which means that there is no evidence of systematic differences between the fixed-effects and random-effects models. Thus, the random-effects model (which is efficient under H0) is preferred for this analysis.

Conclusion

Random Forest (RF), a machine learning method, enhances prediction robustness by aggregating the outputs of multiple decision trees, thereby reducing variance and improving the stability of the results. This ensemble approach is particularly effective in capturing non-linear relationships and complex interactions within the data, making it a powerful tool for predictive modeling. Machine learning methods like RF excel in handling large, high-dimensional datasets and identifying intricate patterns that may not be apparent through traditional statistical techniques.

On the other hand, statistical approaches such as the Fixed Effects Model (FEM) and Random Effects Model (REM) provide a structured framework for analyzing panel data, accounting for both time-varying and time-invariant predictors. By integrating FEM and REM, we can comprehensively analyze the effects of these predictors on stock prices, distinguishing between within-group and between-group variations. This dual approach combines the strengths of statistical modeling—such as interpretability, hypothesis testing, and handling unobserved heterogeneity—with the predictive power of machine learning, offering a deeper understanding of the factors driving stock price movements and enabling more nuanced insights into market dynamics.

In this section, we summarize our findings from this manuscript in Table 16. The results compare the performance of three models—Random Forest (RF), Fixed Effects Model (FEM), and Random Effects Model (REM)—on monthly and daily data. RF achieves a high accuracy rate of 86.57% for monthly data but drops to 63.39% for daily data, indicating that it performs better with aggregated data. FEM and REM show excellent goodness-of-fit values (above 99%) for both datasets, suggesting they effectively capture underlying patterns, even after removing certain variables. These approaches are commonly used in predictive modeling (RF) and panel data analysis (FEM, REM) to handle complex, hierarchical, or time-series data, with RF excelling in modeling non-linear relationships and FEM/REM in accounting for unobserved heterogeneity. Furthermore, based on the results, we assert that FEM and REM are prioritized over RF.

Table (16) The summary of results' findings

Analyzed Models	Monthly data	Daily data
RF	Accuracy rate = 86.57% (with MAE = 0.363)	Accuracy rate = 63.39% (with MAE = 3.819)
FEM	Goodness of Fit = 99.24% (before and after removing "*size*" variable	Goodness of Fit = 99.43% (before and after removing "*max*" variable
REM	Goodness of Fit = 99.25% (before and after removing "*min*" & "*size*" variables	Goodness of Fit = 99.43%

However, FEM and REM demonstrate better performance for daily data compared to monthly data, indicating that a larger number of observations improves prediction accuracy. Therefore, we recommend utilizing statistical approaches (FEM, REM) with as many observations as possible to achieve the highest level of accuracy in stock price market prediction. Overall, these models are highly effective in explaining the dependent variable, offering robust frameworks for analyzing stock price behavior and informing decision-making processes.

References

Bhardwaj, N., & Ansari, M. A. (2019). Prediction of stock market using machine learning algorithms. International Research Journal of Engineering and Technology ٦(5), 5994-6005.

Breiman, L. (2001). Random forests. Machine Learning, ٤٥, 5-32.

Campbell, J. Y., & Perron, P. (1991). Pitfalls and opportunities: What macroeconomists should know about unit roots. NBER Macroeconomics Annual, 6, 141-201.

Cervelló-Royo, R., & Guijarro, F. (2020). Forecasting stock market trend: a comparison of machine learning algorithms. Finance, Markets and Valuation, 6(1), 37–49

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587.

Elsegai, H., Al-Mutawaly, H. S., & Almongy, H. M. (2025). Predicting the Trends of the Egyptian Stock Market Using Machine Learning and Deep Learning Methods. Computational Journal of Mathematical and Statistical Sciences,4 (1), 186-221. doi: 10.21608/cjmss.2024.320645.1077.

Hausman, J. A. (1978). Specification Tests in Econometrics. Econometrica, 46(6), 1251–1271. https://doi.org/10.2307/1913827

Honore, Bo E. (1998) . Estimation Of Panel Data Tobit Models With Normal Errors, Department Of Economics, NJ 08544-1021, Princeton University, Princeton.

Johnson, J., & DiNardo, J. (1997). Econometric methods (4th ed.). McGraw-Hill.

Kmenta, J. (1986). Elements of econometrics (2nd ed.). Macmillan, New York.

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random Forest. International conference on information computing and applications. Berlin, Heidelberg: Springer Berlin Heidelberg.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). SAGE Publications.

Semykina, A., & Wooldridge, J. M. (2010). Estimating panel data models in the presence of endogeneity and selection. Journal of Econometrics ، ١٥٧(2), ٣٧٥-٣٨٠.

Stone, J. R. N., Samuelson, P. A., & Koopmans, T. C. (1954). Report of the Evaluative Committee for Econometrica. *Econometrica, 22*(2), 141-146. https://www.jstor.org/stable/1907538.

Taylor, W. E. (1980). Small sample considerations in estimation from panel data. *Journal of Econometrics*, *13*(2), 203-223.

Vijayarani, S., Suganya, E., & Jeevitha, T. (2020). Predicting stock market using machine learning algorithms. IRJMETS, December 2020.

Wang, W., & Lu, Y. (2018). Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. IOP Conference Series: Materials Science and Engineering, 324. IOP Publishing. https://doi.org/10.1088/1757-899X/324/1/012049.

Weinblat, J. (2018). Forecasting European high-growth firms—A random forest approach. Journal of Industry, Competition and Trade, 18, 253–294.

Wooldridge, J. M. (2003). Introductory econometrics: A modern approach (2nd ed.). South-Western College Publishing.

Appendix (1)

The results of Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root for the six variables under study for **monthly stock price dataset**.

Table (1-1) Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

.xtunitroot fisher close, dfuller lags(0)
Fisher-type unit-root test for close
Based on augmented Dickey-Fuller tests

Ho: All panels contain unit roots Number of panels = 14
Ha: At least one panel is stationary Avg. number of periods = 119.29
AR parameter: Panel-specific Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included ADF regressions: 0 lags

Statistic p-value

Inverse chi-squared(28) P 42.5322 0.0386
Inverse normal Z -2.4561 0.0070
Inverse logit t(74) L* -2.3453 0.0108
Modified inv. chi-squared Pm 1.9419 0.0261

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

Table (1-2) Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

xtunitroot fisher open, dfuller lags(0)

Fisher-type unit-root test for open
Based on augmented Dickey-Fuller tests

Ho: All panels contain unit roots Number of panels = 14
Ha: At least one panel is stationary Avg. number of periods = 119.29
AR parameter: Panel-specific Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included ADF regressions: 0 lags

Statistic p-value

Inverse chi-squared(28) P 47.8301 0.0112
Inverse normal Z -2.8380 0.0023
Inverse logit t(74) L* -2.7814 0.0034
Modified inv. chi-squared Pm 2.6499 0.0040

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

Table (1-3) Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

.xtunitroot fisher max, dfuller lags(0)

Fisher-type unit-root test for max			
Based on augmented Dickey-Fuller tests			

Ho: All panels contain unit roots		Number of panels	= 14
Ha: At least one panel is stationary		Avg. number of periods = 119.29	
AR parameter: Panel-specific		Asymptotics: T -> Infinity	
Panel means: Included			
Time trend: Not included			
Drift term: Not included		ADF regressions: 0 lags	

	Statistic	p-value	

Inverse chi-squared(28)	P	43.6722	0.0299
Inverse normal	Z	-2.3430	0.0096
Inverse logit t(74)	L*	-2.3124	0.0118
Modified inv. chi-squared	Pm	2.0943	0.0181

P statistic requires number of panels to be finite.			
Other statistics are suitable for finite or infinite number of panels.			

Table (1-4) Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

xtunitroot fisher min, dfuller lags(0)			
Fisher-type unit-root test for min			
Based on augmented Dickey-Fuller tests			

Ho: All panels contain unit roots		Number of panels	= 14
Ha: At least one panel is stationary		Avg. number of periods = 119.29	
AR parameter: Panel-specific		Asymptotics: T -> Infinity	
Panel means: Included			
Time trend: Not included			
Drift term: Not included		ADF regressions: 0 lags	

	Statistic	p-value	

Inverse chi-squared(28)	P	47.5980	0.0118
Inverse normal	Z	-2.8134	0.0025
Inverse logit t(74)	L*	-2.7513	0.0037
Modified inv. chi-squared	Pm	2.6189	0.0044

P statistic requires number of panels to be finite.			
Other statistics are suitable for finite or infinite number of panels.			

Table (1-5) Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

xtunitroot fisher value, dfuller lags(0)
Fisher-type unit-root test for value
Based on augmented Dickey-Fuller tests

Ho: All panels contain unit roots Number of panels = 14
Ha: At least one panel is stationary Avg. number of periods = 119.29
AR parameter: Panel-specific Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included ADF regressions: 0 lags

Statistic p-value

Inverse chi-squared(28) P 532.2887 0.0000
Inverse normal Z -20.5949 0.0000
Inverse logit t(74) L* -39.4874 0.0000
Modified inv. chi-squared Pm 67.3884 0.0000

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

Table (1-6) Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

.xtunitroot fisher close_next, dfuller lags(0)
Fisher-type unit-root test for close_next
Based on augmented Dickey-Fuller tests

Ho: All panels contain unit roots Number of panels = 14
Ha: At least one panel is stationary Avg. number of periods = 119.29
AR parameter: Panel-specific Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included ADF regressions: 0 lags

Statistic p-value

Inverse chi-squared(28) P 43.0803 0.0342
Inverse normal Z -2.4598 0.0069
Inverse logit t(74) L* -2.3648 0.0103
Modified inv. chi-squared Pm 2.0152 0.0219

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

Appendix (2)

The results of Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root for the six variables under study for **daily stock price dataset**.

Table (2-1)

Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

.xtunitroot fisher close, dfuller lags(0)
Fisher-type unit-root test for close
Based on augmented Dickey-Fuller tests

Ho: All panels contain unit roots Number of panels = 14
Ha: At least one panel is stationary Avg. number of periods =2215.57
AR parameter: Panel-specific Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included ADF regressions: 0 lags

Statistic p-value

Inverse chi-squared(28) P 69.2482 0.0000
Inverse normal Z -4.1920 0.0000
Inverse logit t(74) L* -4.5518 0.0000
Modified inv. chi-squared Pm 5.5120 0.0000

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

Table (2-2) Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

.xtunitroot fisher open, dfuller lags(0)
Fisher-type unit-root test for open
Based on augmented Dickey-Fuller tests

Ho: All panels contain unit roots Number of panels = 14
Ha: At least one panel is stationary Avg. number of periods =2215.57
AR parameter: Panel-specific Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included ADF regressions: 0 lags

Statistic p-value

Inverse chi-squared(28) P 122.4803 0.0000
Inverse normal Z -7.5201 0.0000
Inverse logit t(74) L* -8.8813 0.0000
Modified inv. chi-squared Pm 12.6255 0.0000

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

Table (2-3) Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

.xtunitroot fisher max, dfuller lags(0)
Fisher-type unit-root test for max
Based on augmented Dickey-Fuller tests

Ho: All panels contain unit roots Number of panels = 14
Ha: At least one panel is stationary Avg. number of periods =2215.57
AR parameter: Panel-specific Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included ADF regressions: 0 lags

Statistic p-value

Inverse chi-squared(28) P 81.8734 0.0000
Inverse normal Z -4.6818 0.0000
Inverse logit t(74) L* -5.4353 0.0000
Modified inv. chi-squared Pm 7.1991 0.0000

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

Table (2-4) Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

.xtunitroot fisher min, dfuller lags(0)
Fisher-type unit-root test for min
Based on augmented Dickey-Fuller tests

Ho: All panels contain unit roots Number of panels = 14
Ha: At least one panel is stationary Avg. number of periods =2215.57
AR parameter: Panel-specific Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included ADF regressions: 0 lags

Statistic p-value

Inverse chi-squared(28) P 84.1752 0.0000
Inverse normal Z -4.9983 0.0000
Inverse logit t(74) L* -5.6997 0.0000
Modified inv. chi-squared Pm 7.5067 0.0000

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

Table (2-5) Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

.xtunitroot fisher value, dfuller lags(0)
Fisher-type unit-root test for value
Based on augmented Dickey-Fuller tests

Ho: All panels contain unit roots Number of panels = 14
Ha: At least one panel is stationary Avg. number of periods =2215.00
AR parameter: Panel-specific Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included ADF regressions: 0 lags

Statistic p-value

Inverse chi-squared(28) P 1009.2223 0.0000
Inverse normal Z -30.4043 0.0000
Inverse logit t(74) L* -74.8684 0.0000
Modified inv. chi-squared Pm 131.1213 0.0000

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

Table (2-6) Augmented Dickey-Fuller (ADF) test of the Fisher-type unit root

.xtunitroot fisher close_next, dfuller lags(0)
Fisher-type unit-root test for close_next
Based on augmented Dickey-Fuller tests

Ho: All panels contain unit roots Number of panels = 14
Ha: At least one panel is stationary Avg. number of periods =2215.57
AR parameter: Panel-specific Asymptotics: T -> Infinity
Panel means: Included
Time trend: Not included
Drift term: Not included ADF regressions: 0 lags

Statistic p-value

Inverse chi-squared(28) P 68.9126 0.0000
Inverse normal Z -4.1693 0.0000
Inverse logit t(74) L* -4.5246 0.0000
Modified inv. chi-squared Pm 5.4672 0.0000

P statistic requires number of panels to be finite.
Other statistics are suitable for finite or infinite number of panels.

تحسين دقة التنبؤ بأسعار الأسهم: دراسة مقارنة للنماذج الإحصائية وخوارزميات التعلم الآلي مع تطبيق على بيانات البانل اللوحية(Panel data)

ملخص الدراسة:

تقدم هذه الدراسة تحليلاً مقارنًا بين نماذج البانل اللوحية ذات المكونات العشوائية (Random Effects)وخوارزميات الغابة العشوائية (Random Forest) في التنبؤ بأسعار الأسهم، مع استكشاف المفاضلة بين قابلية التفسير في النماذج الاقتصادية القياسية والأداء التنبئي لخوارزميات التعلم الآلي. وتساهم هذه الأبحاث في النقاشات الجارية حول استراتيجيات النمذجة المالية التي توازن بين الدقة والقدرة التفسيرية.

باستخدام بيانات لوحية (Panel Data) من أسواق الأسهم، قمنا بتطبيق منهجيتين:

١. **نماذج المكونات العشوائية** لمراعاة التباين غير الملحوظ.

٢. **خوارزميات الغابة العشوائية** لاكتشاف الأنماط غير الخطية المعقدة.

تم تقييم أداء النماذج باستخدام مقاييس مثل متوسط مربعات الخطأ (MSE) ومعامل التحديد(R^2) ، إلى جانب تقييم الكفاءة الحسابية، ومتطلبات البيانات، وقابلية التفسير.

أظهرت النتائج أن خوارزميات الغابة العشوائية حققت دقة تنبئية أعلى قليلاً في بعض السيناريوهات، بينما حافظت نماذج المكونات العشوائية على ميزاتها في قابلية التفسير والمتانة، خاصة في نمذجة التباين. وتسلط هذه النتائج الضوء على التناقض الجوهري بين القوة التنبئية والشفافية في التحليلات المالية.

توضح الدراسة أن نماذج المكونات العشوائية تظل أداة قيّمة للتنبؤ بأسعار الأسهم، رغم المكاسب الطفيفة في الدقة التي توفرها تقنيات التعلم الآلي. فكل منهجية تتمتع بمزايا مميزة: النماذج الإحصائية تقدم رؤى اقتصادية أوضح، بينما تتفوق الأساليب الخوارزمية في الأداء التنبئي.

نقترح لإثراء الباحثين والممارسين إطارًا لاختيار النموذج بناءً على الأولويات التحليلية:

- **إذا كانت قابلية التفسير ضرورية**، فإن نماذج المكونات العشوائية هي الخيار الأمثل.
- **إذا كان الهدف هو تعظيم الدقة التنبئية** (مع توفر الموارد الحسابية الكافية)، فقد تكون خوارزميات الغابة العشوائية أكثر ملاءمة.

يعتمد الاختيار الأمثل على أهداف البحث وخصائص البيانات، حيث تقدم نتائجنا توجيهات مستندة إلى أدلة تجريبية.

الكلمات المفتاحية:

التنبؤ بأسواق الأسهم للبيانات اللوحية؛ الغابة العشوائية؛ نموذج الآثار الثابتة؛ نموذج الآثار العشوائية؛ مقاييس الدقة.