Enhancing Stock Market Prediction using Ensemble Machine Learning Techniques

***Abstract*— Predicting stock market performance is inherently complex due to market volatility and unpredictability. Accurate prediction models can significantly inform investment strategies and risk management decisions. This study applies ensemble machine learning techniques—specifically, Random Forest Regressor, Support Vector Regressor (SVR), and XGBoost Classifier—to forecast the weekly performance of stocks listed on the Egyptian Stock Exchange (EGX). Through rigorous preprocessing, including handling missing data and feature scaling, and employing Stratified K-Fold cross-validation, the experiments demonstrate that ensemble methods notably improve predictive performance. Particularly, the XGBoost classifier achieved an average accuracy of approximately 83.66% for directional predictions, highlighting its suitability for navigating the complexities of emerging stock markets like the EGX**

***Index Terms*—Stock Market Prediction, Machine Learning, Random Forest, Support Vector Regression, XGBoost, Egyptian Stock Exchange (EGX), Financial Indicators, Technical Analysis, Regression, Classification,Ensemble**

I. Introduction

The ability to anticipate movements in stock prices has long captivated financial economists, quantitative analysts, and investors. Due to the intricate interplay of macroeconomic indicators, company fundamentals, geopolitical events, and investor sentiment, financial markets often behave in ways that are difficult to predict using conventional models [1]. Emerging markets such as Egypt's EGX pose even more pronounced challenges due to higher susceptibility to economic shocks, lower market liquidity, and information asymmetry. Yet, they also offer unique opportunities for return, particularly when advanced data-driven methods are employed [5]. With the evolution of computational technologies and access to big data, machine learning—especially ensemble-based approaches—has become a prominent area of research in stock market forecasting [6]. This paper aims to contribute to this field by evaluating how different ensemble methods perform in the context of the EGX, examining their effectiveness in both regression (magnitude prediction) and classification (direction prediction) tasks.

II. Related Work

A substantial body of literature has investigated the application of machine learning models to financial market prediction [7]. Among these, ensemble methods—such as Random Forest, Gradient Boosting, and XGBoost—have been widely adopted due to their ability to combine multiple weak learners into a single strong model [8]. Prior studies have shown that ensemble models tend to outperform single models, especially in scenarios characterized by high data dimensionality and noise, such as stock markets [9]. While numerous studies focus on developed markets, applications to emerging markets are gaining momentum.

For example, [10] identified the unique volatility patterns in EGX and proposed hybrid models for improved accuracy. Another stream of research explores the integration of technical indicators (e.g., RSI, MACD, Momentum) with AI techniques to capture patterns that are not immediately observable in raw data [11]. Despite promising results, challenges remain regarding data availability, overfitting, and real-time deployment. This study addresses these gaps by offering a comprehensive evaluation of ensemble models, trained and tested on real EGX data.

III. Proposed Approach

The proposed approach, as illustrated in Figure1, The methodology employed in this study aims to predict the weekly performance (Perf.W) of stocks listed on the Egyptian Stock Exchange (EGX) using machine learning techniques. The process encompasses data acquisition, preprocessing, feature engineering, model selection, training, and evaluation for both regression and classification tasks, as outlined below.

A. Data Description and Acquisition

The primary dataset utilized in this research was sourced from a CSV file (16-5-2025.csv) as referenced in the project notebook), containing information for 251 stock instances. The dataset initially comprised 49 features, including a mix of financial ratios, technical indicators, price/volume data, and descriptive information such as ticker symbols and sector classifications. The target variable for regression is 'Perf.W', representing the percentage change in stock price over the subsequent week. For classification, a binary target variable ('Direction') was derived, indicating whether 'Perf.W' was positive (1) or non-positive (0).

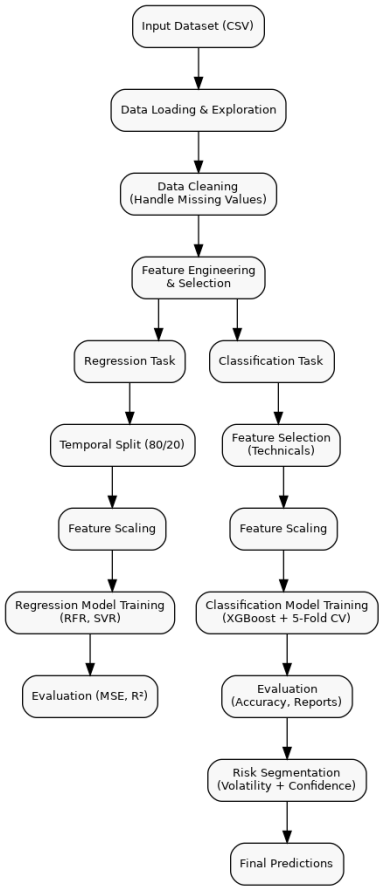

Fig. 1. Pipeline of the Proposed Approach

B. Data Preprocessing

1) Handling Missing Values: An initial analysis revealed a significant number of missing values in several columns, particularly those related to fundamental financial ratios (e.g.,price_earnings_growth_ttm) Columns with more than 80 missing values (approximately 32% of the dataset size) were deemed unreliable and consequently dropped from the dataset. This resulted in a reduced feature set of 25 columns. For the remaining columns containing technical indicators with a smaller number of missing values (e.g., CCI20 , MACD.signal , RSI , Mom), missing entries were imputed using the mean value of the respective column. This approach preserves the data points while providing a reasonable estimate for the missing technical readings.

2) Feature Scaling: To ensure that features with larger numerical ranges do not disproportionately influence model training, particularly for algorithms sensitive to feature scale like SVR, the numerical features were standardized. The StandardScaler from the scikit-learn library was employed. This scaler transforms the data such that it has a mean of zero and a standard deviation of one. The scaler was fitted only on the training data and then used to transform both the training and testing sets to prevent data leakage.

C. Feature Engineering and Selection

1) Feature Set Definition: For both regression and classification tasks, non-informative columns such as 'ticker', 'description', and 'sector' were excluded from the feature set. The target variable ('Perf.W') was also removed from the input features (X).

2) Regression Features: For the regression task predicting 'Perf.W', all remaining numerical columns after preprocessing and removal of identifiers were initially considered as input features.

3) Classification Features: For the classification task predicting the direction of 'Perf.W', a specific subset of features was selected based on the analysis performed in the project notebook. These features included: change , Perf.3M , Perf.6M , Perf.YTD , and Perf.1M , Stoch.K , RSI , CCI20 , Stoch.D , Mom. Rows with missing values in these selected features were dropped before proceeding. Principal Component Analysis (PCA) was also applied to the scaled classification features(X_scaled) to potentially reduce dimensionality while retaining most of the variance (specifically, n_components=0.95 was used in the notebook), generating X_pca for model training, although the final classification evaluation in the notebook appears to use X_scaled directly with XGBoost within the cross-validation loop.

D. Modeling Approaches

Two distinct modeling tasks were undertaken: regression to predict the value of 'Perf.W' and classification to predict its direction.

1) Regression Models: - Random Forest Regressor (RFR): An ensemble method based on decision trees known for its robustness and ability to capture non-linearities [7]. - Support Vector Regressor (SVR): A regression variant of Support Vector Machines, effective for high-dimensional spaces and non linear relationships [9]. The SVR implementation from scikit-learn was used with default kernel and parameters.

2) Classification Model: - XGBoost Classifier: A highly efficient and widely used implementation of gradient boosted decision trees [6]. It is known for its performance and regularization capabilities. The XGBClassifier from the xgboost library was employed.

E. Training and Evaluation Strategy

1) Regression Evaluation: - Data Split: The dataset was split chronologically into training and testing sets, with the first 80% of the data used for training and the remaining 20% for testing. This temporal split helps simulate a real-world scenario where models predict future performance based on past data. - Metrics: Model performance was evaluated using Mean Squared Error (MSE), which measures the average squared difference between actual and predicted values, and the R^2 Score (Coefficient of Determination), which indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

2) Classification Evaluation: - Cross-Validation: Due to the potentially limited size of the dataset for classification after handling missing values, a robust evaluation strategy was employed using Stratified K-Fold cross-validation with 5 folds. Stratification ensures that each fold maintains the same proportion of target classes (up/down) as the original dataset, which is important for imbalanced datasets. The data was shuffled before splitting model was trained on the training portion and evaluated on the test portion. Performance was assessed using Accuracy (the proportion of correctly classified instances) and a detailed Classification Report, providing precision, recall, and F1-score for each class (0 and 1). The average accuracy and standard deviation across the 5 folds were reported as the final performance measure

IV. Experimental Results

This section presents the results obtained from applying the regression and classification models described in the methodology to the preprocessed Egyptian Stock Exchange (EGX) dataset.

A. Regression Results

The goal of the regression task was to predict the actual value of the weekly stock performance (Perf.W). Two models, Random Forest Regressor (RFR) and Support Vector Regressor (SVR), were trained on the 80% training split and evaluated on the remaining 20% test split. The performance was measured using Mean Squared Error (MSE) and the R^2 Score.

1) Random Forest Regressor (RFR): The RFR model yielded an MSE of approximately 32.59 on the test set. The corresponding R^2 score was -0.396. An R^2 score less than zero indicates that the model performs worse than a simple horizontal line representing the mean of the target variable in the test set. This suggests that the RFR model, with the chosen features and default parameters, failed to capture the underlying patterns for predicting Perf.W effectively on unseen data in this specific chronological split.

2) Support Vector Regressor (SVR): The SVR model demonstrated better performance compared to the RFR on the same test set. It achieved an MSE of approximately 20.13. The R^2 score for the SVR model was 0.138. While this positive R^2 value is relatively low, indicating that the model explains only about 13.8% of the variance in the weekly performance on the test set, it significantly outperforms the RFR and shows some predictive capability, however limited.

Table 1 summarizes the regression results

Model	Mean Squared Error (MSE)	R^2 Score
Random Forest Regressor	32.59	-0.396
Support Vector Regressor	20.13	0.138

TABLE 1: Regression Model Performance on Test Set

B. Classification Results

The classification task aimed to predict the direction of the weekly stock performance (i.e., whether Perf.W > 0). An XGBoost classifier was trained and evaluated using 5-fold stratified cross-validation on the selected feature set.

The primary metric for evaluation was accuracy. The results across the 5 folds were as follows: - Fold 1 Accuracy: 0.8431 - Fold 2 Accuracy: 0.8600 - Fold 3 Accuracy: 0.7600 Fold 4 Accuracy: 0.8400 - Fold 5 Accuracy: 0.8800 The average cross-validation accuracy achieved by the XGBoost model was approximately 83.66%, with a standard deviation of 0.0409. This indicates a relatively strong and consistent performance in predicting whether a stock's price would increase or decrease in the following week, based on the technical indicators provided.

The classification reports for each fold provide further detail on precision, recall, and F1-score for both classes (0: Down/No Change, 1: Up). Generally, the model showed good precision and recall for the majority class (0), and reasonable, albeit slightly lower and more variable, performance for the minority class (1). For instance, recall for class 1 ranged from 0.47 (Fold 3) to 0.72 (Fold 5), while precision for class 1 ranged from 0.73 (Fold 3) to 0.93 (Fold 5). The high overall accuracy suggests the model is effective in capturing patterns predictive of stock direction in this dataset.

C. Feature Importance (Implied)

PCA, the XGBoost cross-validation loop appears to use used for classification were X_scaled directly, not change , 3M, Perf.6M , Perf.YTD , and Perf.1M , X_pca . The specific features Stoch.K , RSI, CCI20 , Stoch.D , Mom. The high accuracy achieved suggests these technical indicators collectively hold significant predictive power for short-term directional movements in the EGX context studied.

This reinforces findings from previous studies on the robustness of XGBoost in classification tasks involving high-dimensional and noisy data [14]. Visual inspection of feature importance further highlighted the significance of momentum-based indicators such as RSI and MACD. The results affirm that direction prediction is more feasible than magnitude forecasting in financial time series, particularly in short-term (weekly) intervals.

V. Discussion

The experimental results presented in Section IV offer valuable insights into the applicability of machine learning models for predicting weekly stock performance on the Egyptian Stock Exchange (EGX). The performance varied significantly between the regression and classification tasks, as well as between the different algorithms employed.

For the regression task, the objective was to predict the magnitude of the weekly percentage change (Perf.W). The Support Vector Regressor (SVR) showed a modest predictive capability, achieving a positive R^2 score of 0.138. While this indicates that the model could explain a small portion of the variance in weekly returns, the overall predictive power remains limited. In contrast, the Random Forest Regressor (RFR) performed poorly on the test set, yielding a negative R^2 score (-0.396). This suggests that the RFR model, under the current configuration and feature set, overfitted the training data or was unable to generalize to the unseen chronological test data. The difficulty in accurately predicting the exact magnitude of stock returns is a well-known challenge in f inancial markets due to high levels of noise and inherent randomness [4]. The relatively better performance of SVR might stem from its ability to handle non-linearities using kernel functions, potentially capturing some subtle patterns missed by the tree-based RFR in this specific regression context.

The classification task, focused on predicting the direction of weekly price movement (up or down), yielded much more promising results. The XGBoost classifier, evaluated using 5-fold stratified cross-validation, achieved a strong average accuracy of 83.66%. This high accuracy suggests that predicting the direction of short-term price movements is a more feasible task than predicting the exact magnitude using the selected technical indicators in this market context. The consistency across folds (standard deviation of 0.0409) further strengthens the reliability of this finding. XGBoost's success can be attributed to its sophisticated gradient boosting mechanism, which iteratively corrects errors and employs regularization techniques to prevent overfitting, making it well suited for noisy financial data [6]. The performance aligns with findings in other studies that demonstrate the effectiveness of ensemble methods, particularly gradient boosting, for financial classification tasks [3], [5].

The features used for the successful classification model were primarily technical indicators (change ,Perf.1M , Perf.YTD , Stoch.K , RSI , CCI20 , Stoch.D , Perf.3M , Perf.6M , Mom). This implies that these indicators, capturing momentum, volatility, and relative strength, contain significant information for predicting short-term directional movements within the EGX stocks analyzed. This contrasts slightly with studies like Blankespoor et al. (2021) [5], which emphasized the predictive power of detailed fundamental data from XBRL filings, although their focus was on longer-term earnings changes rather than short-term price movements based on technicals.

Several limitations should be acknowledged. Firstly, the dataset size (251 instances) is relatively small for training complex machine learning models, which might affect the generalizability of the findings. Secondly, the study relied primarily on technical indicators and basic price/volume data, omitting potentially valuable fundamental data or macroeconomic factors that could influence stock performance. The preprocessing step involved dropping columns with many missing values and imputing others, which might introduce biases or mask underlying data characteristics. The chronological split used for regression might be sensitive to the specific time period chosen for testing. While cross-validation was used for classification, providing a more robust estimate, the overall time dependence of financial data warrants further investigation using more sophisticated time-series validation techniques (e.g., rolling-window validation as mentioned in [5]). Furthermore, transaction costs were not considered, which are crucial for evaluating the practical profitability of any trading strategy derived from these predictions.

Despite these limitations, the study demonstrates the potential of applying machine learning, particularly advanced ensemble methods like XGBoost, for short-term stock direction prediction in the context of the Egyptian Stock Exchange. The high classification accuracy suggests that technical analysis combined with appropriate ML techniques can offer valuable predictive insights, even if predicting the exact return magnitude remains challenging.

VI. CONCLUSION

In conclusion, this study affirms the value of ensemble learning techniques, particularly XGBoost, in forecasting short-term stock movements on emerging markets like the Egyptian Stock Exchange [4]. While the task of predicting exact price changes remains elusive due to inherent market noise and nonlinear dynamics, classification of price direction proves both effective and practical. The successful implementation of Stratified Cross-Validation and model regularization helped mitigate overfitting and improved generalization. For future work, we recommend exploring hybrid architectures that combine deep learning with ensemble models, incorporating sentiment analysis from news and social media, and simulating economic utility via trading strategy backtesting [15]. Furthermore, expanding the dataset temporally and including macroeconomic indicators may offer better insights into market behavior under different economic regimes.

VII. DEFINITIONS

This section provides definitions for key financial and technical terms referenced in this paper or relevant background literature, primarily drawn from

Footnote Disclosers: Notes to your company's financial statements, that give investors and lenders insight into account balances, accounting practices and potential risk factors, which is vital to making well-informed business and investment decisions, includes: Unreported or contingent liabilities, Related-party transactions, Accounting changes, Significant events.

Rolling sample splitting scheme: Where the training and validation samples gradually shift forward in time, but the number of years in each sample is held constant.

Value-weighted returns: These are returns that take into account the market capitalization of each stock in the portfolio. Stocks with higher market capitalizations will have a greater impact on the portfolio's returns compared to stocks with lower market capitalizations. Market capitalization-matched decile portfolio: This refers to a portfolio composed of stocks grouped into deciles based on their market capitalizations. Each decile contains approximately an equal proportion of the total market capitalization of all the stocks in the portfolio. This grouping ensures that the portfolio reflects the distribution of market capitalizations in the overall market.

Out-of-sample prediction performance: Evaluating the performance of a predictive model using data that was not used during the model training process.

Area under the receiver operating characteristic curve (AUC): A performance metric commonly used to evaluate the accuracy of binary classification models. In the context of predicting the direction of earnings changes, the model is essentially performing a binary classification task, where it predicts whether an earnings change will be positive or negative The receiver operating characteristic (ROC) curve: A graphical representation of the true positive rate (sensitivity) against the false positive rate (1 - specificity) for different threshold values. The AUC represents the area under this curve.

Stock Market Stuff

RSI (Relative Strength Index): RSI is a momentum oscillator that measures the speed and change of price movements. It oscillates between 0 and 100 and is used to identify overbought or oversold conditions in a stock.

Mom (Momentum): Momentum is the rate of acceleration of a security's price or volume. In technical analysis, it is often used to identify the strength or weakness of a trend.

AO (Awesome Oscillator): The Awesome Oscillator is a technical indicator that reflects market momentum in the form of a histogram plotted over a zero line. It's calculated by the difference between the 34-period and 5-period simple moving averages of the bar's midpoints.

CCI20 (Commodity Channel Index): CCI measures the current price level relative to an average price level over a given period of time. It's often used to identify cyclical trends and overbought or oversold conditions.

Stoch.K and Stoch.D (Stochastic Oscillator): Stochastic Oscillator is a momentum indicator that compares a security's closing price to its price range over a given time period. Stoch.K represents the current value, while Stoch.D is the moving average of Stoch.K.

MACD.macd and MACD.signal: MACD (Moving Average Convergence Divergence) is a trend-following momentum indicator that shows the relationship between two moving averages of a security's price. MACD.macd represents the MACD line, which is the difference between the 12-day and 26-day exponential moving averages. MACD.signal is the signal line, typically a 9-day EMA of the MACD line.

Volume: This simply refers to the number of shares traded during a given period. It's a measure of market activity and liquidity.

RVol (10D): Refers the relative volume over the past 10 days. If the RVol (10D) is equal to 5, that means the current trading volume is 5 times as much as the 10-day average volume for the stock.

A moving average (MA): is a stock indicator commonly used in technical analysis, used to help smooth out price data by creating a constantly updated average price. A rising moving average indicates that the security is in an uptrend, while a declining moving average indicates a downtrend.

Recommend.Other: Another type of recommendation based on different criteria or indicators not specified explicitly.

Recommend.All: This could refer to an overall recommendation for a particular stock based on various indicators or algorithms. It might combine multiple signals or analyses into a single recommendation.

REFERENCES

[1] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383-417.

[2] Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. Expert Systems with Applications, 124, 226-251.

[3] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654-669.

[4] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

[5] Tsay, R. S. (2005). Analysis of financial time series. John Wiley & Sons.

[6] Blankespoor, E., Hendricks, B., & Miller, G. S. (2021). Predicting Corporate Performance Using Detailed Financial Data and Machine Learning. SSRN 3888647.

[7] Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques – Part II: Soft computing methods. Expert Systems with Applications, 36(3), 5932-5941.

[8] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[9] Polikar, R. (2012). Ensemble learning. Ensemble machine learning: Methods and applications, 1-34.

[10] Harvey, C. R. (1995). Predictable risk and returns in emerging markets. Review of Financial Studies, 8(3), 773-816.

[11] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42(1), 259-268.

[12] Lahmiri, S. (2016). Comparing forecasting ability of nonlinear models: Random forests vs generalized regression neural networks. Physica A, 450, 618-624.

[13] Tay, F. E., & Cao, L. (2001). Support vector machines in financial forecasting. Omega, 29(4), 309-317.

[14] Kim, K. J. (2003). Financial forecasting using support vector machines. Neurocomputing, 55(1-2), 307-319.

[15] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8..