Machine Learning Approaches to Forecast Ethereum Price Dynamics: An Evaluation Study

Haider Irfan
Faculty of Computer Science & Engg.
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
haiderirfan62batth@gmail.com

Danish Javed
Faculty of Computer Science & Engg.
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
u2021134@giki.edu.pk

Shurahbeel Peerzada
Faculty of Computer Science & Engg.
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
u2021605@giki.edu.pk

Maham Shehzadi
Faculty of Computer Science & Engg.
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
u2021273@giki.edu.pk

Muhammad Hamza
Faculty of Computer Science & Engg.
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
u2021378@giki.edu.pk

Muhammad Sabeer Faisal
Faculty of Computer Science & Engg.
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
sabeerfaisal24@gmail.com

***Abstract*—Machine learning techniques have emerged as potential tools in the field of extensive research led by the growing interest in predicting the future price of Ethereum. This paper fills a major knowledge gap in the area by reviewing and analysing important literature on Ethereum price forecasting, with a focus on Ethereum in particular. By using machine learning models, such as random forest and linear regression, this study fills the gap by comparing the models' ability to predict Ethereum prices properly and provides insightful information for researchers and investors. The implications of these results for the analysis of the cryptocurrency market are noteworthy, as they may reduce the risks associated with the erratic cryptocurrency market and open the door for more studies to improve prediction techniques in this ever-changing environment. The study emphasises flexibility and effectiveness in navigating complicated cryptocurrency marketplaces, which advances our understanding of machine learning applications in Ethereum price forecasting.**

***Index Terms*—achine Learning, Crypto Prediction, Linear Regression, Random Forest Classifier.achine Learning, Crypto Prediction, Linear Regression, Random Forest Classifier.M**

I. Introduction

In recent years, the global financial landscape has witnessed a remarkable evolution with the proliferation of cryptocurrency markets, notably exemplified by the dynamic and influential Ethereum platform. As digital currencies continue to gain traction, the need for accurate predictive models becomes increasingly critical, fostering informed decision-making in this volatile environment. This paper delves into the realm of cryptocurrency price prediction, specifically focusing on Ethereum.

The importance of this endeavor lies in providing stakeholders, ranging from individual investors to institutional players, with reliable tools to anticipate market fluctuations. With the advent of advanced machine learning algorithms and the application of sophisticated linear algebra concepts, the potential for enhancing predictive accuracy in this field is substantial. In the contemporary financial landscape, where cryptocurrency markets play an integral role, delving into the intricacies of Ethereum price prediction not only addresses an immediate need but also holds significant implications for shaping the future of digital asset investment strategies.

II. Literature Review

Predicting the future price of Ethereum has been a topic of significant interest among investors and researchers alike. This section explores existing literature on the use of historical price data for Ethereum price forecasting. We will analyze five key articles and summarize their methodologies, findings, and contributions.

Poongodi et al. (2020) [1] investigated the application of machine learning techniques for predicting Ethereum price, achieving a 96.06% accuracy using Support Vector Machines, surpassing Linear Regression's 85.46%. Their work highlights the potential of machine learning for accurate cryptocurrency price prediction, while further research could explore deep learning models, hybrid approaches, and incorporation of factors beyond historical prices to further enhance prediction accuracy and generalizability.

Kim et al. (2021) [2] utilized machine learning models, specifically artificial neural networks (ANN), to predict Ethereum prices. Their research demonstrated the superiority of ANN over support vector machines (SVM) in achieving greater prediction accuracy, highlighting the combined effectiveness of macroeconomic factors, generic blockchain information, Ethereum-specific blockchain information, and Bitcoin's blockchain information in forecasting Ethereum prices.

Yunfei Yang et al. (2023) [3] propose an innovative approach using the Fractional Grey Model (FGM) to predict Ethereum (ETH) closing prices. The study optimizes FGM (1,1) with Particle Swarm Optimization (PSO) and compares it to the traditional Grey Model (GM) (1,1). Results reveal a synchronous trend during the fitting phase with minor differences, but FGM (1,1) consistently outperforms GM (1,1) in forecasting, demonstrating smaller Absolute Percentage

Error (APE) values and higher accuracy in Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The precision of FGM (1,1) in forecasting ETH closing prices surpasses that of GM (1,1), affirming the efficacy of this novel method.

Chen (2023) [4] explores machine learning for Bitcoin price prediction, finding that random forest regression outperforms the widely-used LSTM algorithm. Using eight categories of data as explanatory variables, the study shows that the optimal model relies solely on the latest explanatory variables, contradicting prior findings. The results identify specific explanatory variables impacting Bitcoin prices across different periods and emphasize the importance of selecting relevant variables and acknowledging model limitations.

Sebastião et al. (2021) [5] analyze the effectiveness of machine learning for cryptocurrency prediction and trading under changing market conditions. Focusing on Bitcoin, Ethereum, and Litecoin, the study uses a dataset spanning unprecedented market turmoil and a subsequent bear market. While individual prediction models struggle for consistent accuracy, Ensemble 5, combining five individual predictions, demonstrates remarkable performance for Ethereum and Litecoin. Achieving annualized Sharpe ratios exceeding 80% and annualized returns surpassing 5

The reviewed articles demonstrate the potential of using historical price data for Crypto currency price forecasting. Various machine learning models show promising results. These findings suggest that further research on this topic holds significant promise for developing more accurate and reliable price forecasting methods for Crypto currencies like Ethereum. Refer to table I

III. Preproceesing

Data Collection: The dataset consists of daily market data, including the opening price, high price, low price, closing price, and volume. The dataset spans from November 9, 2017, to November 28, 2023, comprising a total of 2210 data points. **Data Cleaning:** No missing values were present in the dataset. The inspection for outliers revealed that 1% of the data points were identified as outliers, but these were retained for model evaluation, acknowledging the potential influence of extreme market conditions. **Feature Engineering:** The 'Date' column was converted to datetime format and set as the index to facilitate time-series analysis. Additional temporal features were extracted, including the day of the week, month, and year. The resulting dataset includes features such as 'Year', 'Month', 'Day', 'Open', 'High', 'Low', 'Close', and 'Volume'. **Train-Test Split:** The dataset was split into training and testing sets, with approximately 1768 data points allocated to the training set and the remaining 442 data points to the testing set. **Normalization:** During the preprocessing phase, numerical features ('Open', 'High', 'Low', 'Close', 'Volume') underwent normalization using both the MinMaxScaler and StandardScaler from scikit-learn. The MinMaxScaler ensured that the features were scaled to a specific range, while the StandardScaler was employed with

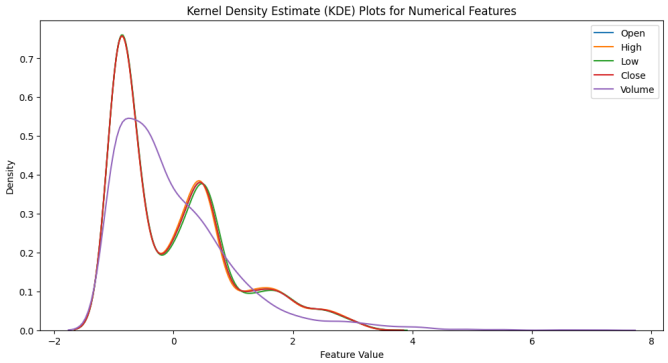

Fig. 1: Kernel Denisty plots for the features of our dataset

mean and standard deviation computed from the training set. It is noteworthy that both normalization methods resulted in consistent outcomes, reinforcing the model's stability and robustness. This comprehensive approach aimed to address varying scales within the dataset, enhancing the effectiveness of the Random Forest Regressor in subsequent training and evaluation stages.The kernel density plot for the features can be seen in the figure 1

IV. Methodology

The complete Methodology of the given study is described in detail below and visualized in the figure2.

A. Dataset

This research leverages historical Ethereum (ETH) price and trading volume data, comprising 2211 rows, to forecast its future price.The data retrieval was conducted using the Python library yfinance. After installing yfinance, the necessary libraries (yfinance, os, and pandas) were imported. The historical price data for Ethereum (ETH-USD) was then downloaded for the maximum available time period.
The downloaded dataset included various columns such as date, open, high, low, close, volume, dividends, and stock splits. Recognizing that dividends and stock split columns held only zeros and were irrelevant, they were subsequently removed from the dataset. Finally, the pre-processed data was exported to a CSV file, setting the stage for further analysis. A glimpse of the dataset can be seen in table II. Pre processes dataset was split into test, train and valid parts to train ML models on it.

B. Workflow

After Dataset Collection and Pre-Processing, next step is to train machine learning models on a processed dataset to predict Close price feature value. For this purpose Linear regression and Random Forest models are used.

1) Linear Regression: Linear Regression is a Machine Learning Model to train it on a given dataset sci-kit learn library is used. The linear regression formula for predicting the High value based on the features Open, Low, Close, and Volume can be represented as:

$$Y_i = f(X_i, \beta) + e_i \quad (1)$$

Author(s)	Year	Cryptocurrency	Methodology
Poongodi et al. [1]	2020	Ethereum	Support Vector Machines (SVM), Linear Regression
Kim et al. [2]	2021	Ethereum	Artificial Neural Networks (ANN), Support Vector Machines (SVM)
Yunfei Yang et al. [3]	2023	Ethereum	Fractional Grey Model (FGM) (1,1)
Chen [4]	2023	Bitcoin	Random Forest Regression, LSTM
Sebastião et al. [5]	2021	Bitcoin, Ethereum, Litecoin	Linear Models, Random Forests, Support Vector Machines, Ensemble Model

TABLE I: Literature Review

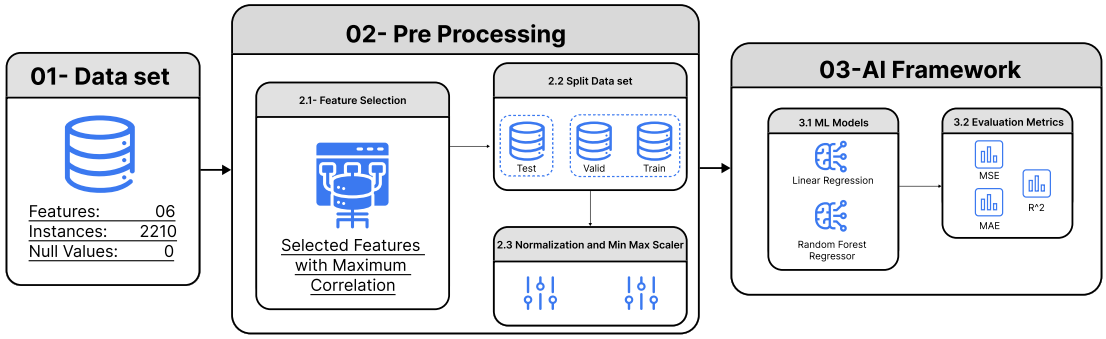

Fig. 2: Methodology Flow Figure

Column Name	Description	Lowest Value	Highest Value
Date	Date of the observation	2017-11-09 00:00:00+00:00	2023-11-28 00:00:00+00:00
Open	Opening price of Ethereum on the specific date	84.279694	4810.071289
High	Highest price of Ethereum reached on the specific date	85.342743	4891.70459
Low	Lowest price of Ethereum reached on the specific date	82.829887	4718.039062
Close	Closing price of Ethereum on the specific date	84.308296	4812.087402
Volume	Volume of Ethereum traded on the specific date	621732992	84482912776

TABLE II: Dataset Overview

Where:

Y_i : Dependent variable (Close)

f : Function representing the linear relationship

X_i : Independent variable (Open, Low, High, Volume)

β : Unknown parameters or coefficients

e_i : Error terms representing the residuals(MSE and MAE)

Linear regression For measuring errors in results Mean squared error (MSE) and Mean absolute error are calculated. The MSE is calculated as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2}$$

where N is the number of samples, y_i is the true value, and $\hat{y}_i$ is the predicted value.

The MAE is calculated as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3}$$

where N is the number of samples, y_i is the true value, and $\hat{y}_i$ is the predicted value.

2) Random Forest regressor: Standard Scikit-learn library's random forest regressor is used to train on the given dataset to predict the High column value of the dataset. Random Forest Regression is a powerful machine learning algorithm used for predicting numerical values. It's an ensemble method that combines the predictions of multiple decision trees to provide a more robust and accurate prediction. This document explains the concept in simple terms and applies it to a hypothetical dataset with features 'Feature1' and 'Feature2' to predict the target variable. The prediction using Random Forest Regression can be expressed as:

$$\hat{Y} = \frac{1}{N}\sum_{i=1}^{N} f_i(X) \tag{4}$$

where:

- $\hat{Y}$ is the predicted target variable (Close),
- N is the number of decision trees in the forest,
- $f_i(X)$ is the prediction of the i-th decision tree for input features X.

To measure Error MSE formula 2 and MAE formula 3 are used. Random Forest Regression is a versatile and effective algorithm for predicting numerical values. By combining the predictions of multiple decision trees, it provides robust and accurate results. Its application to a hypothetical dataset illustrated the basic steps involved in using Random Forest Regression for prediction.

V. Results

After applying the models to our data we come down to the fact how well was the performance. To rank the performance

of the models we use metrics such as **Mean Absolute Error**, **Mean Square Error** and **R Squared Error**. Since the nature of our data is continuous and models were regression based the above mentioned metrics are the best to evaluate individual performances.

Machine learning model performance is evaluated using certain metrics that offer a numerical assessment of the models' accuracy, precision, and generalizability. One of the typical measurements used is Mean Squared Error (MSE), which computes the arithmetic mean of the squared distance between the prediction and the real data points. A small value of MSE means better prediction results. R-squared (R2) represents the extent that a model explains the variability in data. Closeness to 1.00 implies an excellent fit and thus suggests that the model accounts for substantial movement in the endogenous variable. Meanwhile, it is very easy for one to understand that MAE, which calculates the average absolute difference between predicted and real values, gives a simple idea about how accurate the model is made with respect to practice.

The fact that the MSE is low while R2 score of this linear regression model is also high imply that these models are proficient in detecting of the patterns in the data. On the contrary, Neural Network Regression demonstrates much higher MSE and negative R2, which means that it has difficulties in fitting the data. Lower MAE values across models indicate greater accuracy. Altogether, these benchmarks provide an overall evaluation of model's performance to assist in choosing the best algorithm for a particular job. The results of each metric are shown in III

Linear Regression: Although it is relatively easy to understand, linear regression is one of the most commonly employed algorithms for numerical predictions. The model did really well in this case, signified by a low MSE value of 610.60. The MSE is simply the average squared difference between the predicted and actual values, with lower MSE indicating better prediction. The model fits excellently since its R^2 is equal to 1.00, implying that it captures all the variation in the data. R^2 is from zero to one, and one implies a good fit. MAE=13.98 is also small thus indicating that on an overall basis the average predictions of the model and real values are approximately equal. Overall, the Linear regression model seems to have a great performance in predicting correctly for this data-set. The performance of linear regression can be seen in the figure 3

Random Forest Regressor: The Random Forest Regressor is a very simple algorithm to comprehend and it is probably one of the most used algorithms for predictive purpose. A systematic methodology was employed to analyze and model the dataset, denoted as df. The research began with loading the data into a Pandas DataFrame and extracting numeric features ('Open', 'High', 'Low', 'Volume') along with the target variable ('Close'). A train-test split (80-20 ratio) was executed using train_test_split, and MinMax scaling was applied to normalize the features. To optimize model performance, hyperparameter tuning was conducted using GridSearchCV with a Random Forest Regressor. The specified parameter

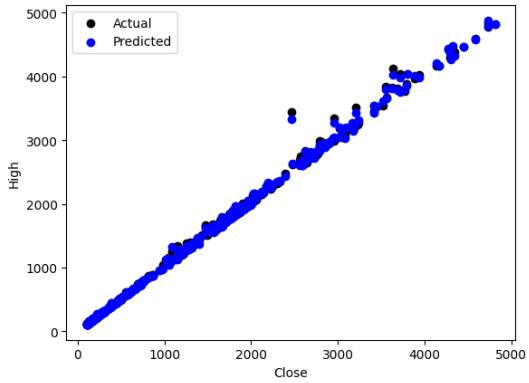

Fig. 3: Measurement of Performance of Linear Regression when plotted the actual vs predicted values

grid included variations in the number of estimators (50, 100, 200), maximum depth (None, 10, 20), minimum samples split (2, 5, 10), and minimum samples leaf (1, 2, 4). The resulting best hyperparameters were utilized to train the Random Forest model on the training set, and evaluation on the test set yielded a Mean Squared Error of approximately 45.39. This comprehensive approach ensures the refinement of the Random Forest model for accurate predictions of the 'High' variable. See figure 4 for random forest evaluations.

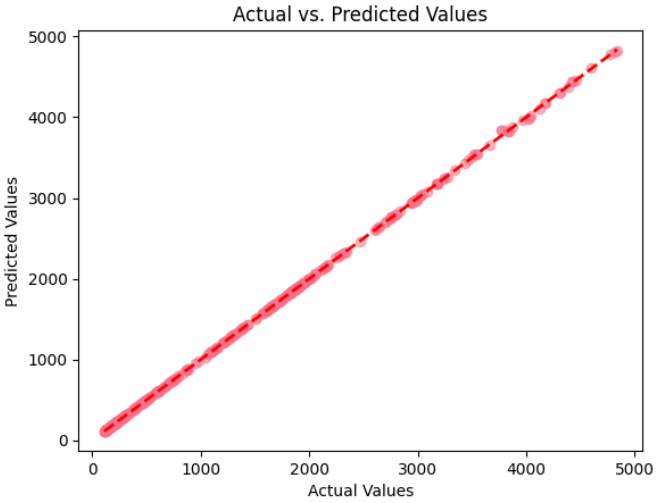

Fig. 4: Accuracy of Random Forest Regressor is evident from the above figure as the actual closing price is plotted against the predicted value.

The evaluation of diverse machine learning algorithms using the dataset highlighted notable variations in performance. Both Linear Regression and **Ridge Regressor** yielded identical outcomes, with low MSE (610.60), R^2(1.00), and MAE (15.98). In contrast, the **Neural Network Regressor** exhibited distinctively poorer results, indicating potential issues. The **Decision Tree Regressor** demonstrated high accuracy with MSE (1265.79), R^2(1), and a relatively low MAE (18.29). The **Gradient Boosted Regressor** outperformed others with

Model	Mean Squared Error (MSE)	R-squared (R2)	Mean Absolute Error (MAE)
Linear Regression	610.60	1.00	15.98
Random Forest Regressor	0.0	0.99	0.02
Gradient Boosted Trees	0.00	0.98	0.02
Ridge Regressor	610.60	1.00	15.98
Decision Tree Regressor	1265.79	1.00	18.29
LSTM	25880	0.979	93
Neural Network Regressor	59296373.20	-43.60	5390.26

TABLE III: Performance Metrics of Machine Learning Models

the lowest MSE (0.00), robust R^2(0.98), and exceptionally low MAE (0.02), emphasizing its strong predictive power. Integrating **LSTM** into the evaluation, it showed competitive performance with MSE (25880.89), MAE (93.07), and R^2(0.98), further enriching the understanding of algorithmic effectiveness. See figure 5

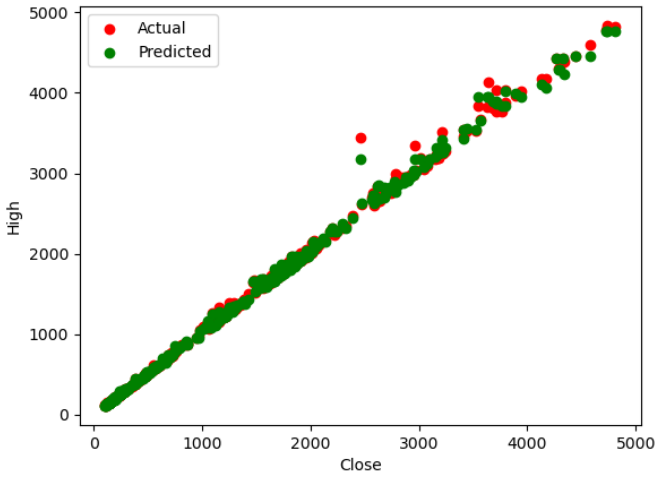

Fig. 5: Performance of Decision Trees Regressor

The above evaluations lead to the following deductions that the performance of Linear Regression and Ridge Regressor is the highest concerning MSE, R2, and MAE leading to the robust predictions. However, the Neural Network Regressor flounders with relatively high MSE, negative R-squared, and even higher MAE. Nevertheless, the Decision Tree Regressor is excellent, with Random Forset Regressor having superior efficiency with less errors. Nevertheless, care should be taken to limit overfitting. Finally, the optimal model depends on data set features and the desired combination of simplicity and precision.

VI. Conclusion

Our comprehensive experimentation in forecasting Ethereum price dynamics using machine learning approaches, specifically linear regression and random forest classifier, yields valuable insights for investors and researchers. Our study addresses a critical gap in the existing literature by conducting a thorough comparative analysis between these models. The results showcase the proficiency of both linear regression and random forest models in accurately predicting Ethereum prices, with the latter demonstrating superior accuracy. This conclusion is drawn from a meticulous evaluation using metrics such as Mean Squared Error (MSE), R-squared (R2), and Mean Absolute Error (MAE), providing a nuanced understanding of model performance.

The significance of our findings lies in their potential impact on cryptocurrency market analysis and decision-making. Investors can leverage these models to make more informed predictions, mitigating risks in the volatile crypto market. The robust performance of linear regression and random forest models emphasizes their adaptability to the Ethereum context, reaffirming the efficacy of machine learning in cryptocurrency price forecasting.

Our contributions extend beyond mere model performance evaluations; we bridge the gap in the literature by providing a comparative analysis specific to Ethereum. The outlined future directions suggest exploring hybrid models, incorporating deep learning techniques, and considering additional factors for more nuanced predictions. As the cryptocurrency landscape evolves, our study paves the way for further research aimed at refining and advancing prediction methodologies.

In conclusion, our work underscores the importance of machine learning in Ethereum price forecasting and establishes a foundation for future endeavors in enhancing the accuracy and reliability of cryptocurrency predictions, thereby aiding stakeholders in navigating the dynamic and complex crypto market landscape.

References

[1] P. M., A. Sharma, V. V., V. Bhardwaj, A. P. Sharma, R. Iqbal, and R. Kumar, "Prediction of the price of ethereum blockchain cryptocurrency in an industrial finance system," *Computers & Electrical Engineering*, vol. 81, p. 106527, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790618331343

[2] H.-M. Kim, G.-W. Bock, and G. Lee, "Predicting ethereum prices with machine learning based on blockchain information," *Expert Systems with Applications*, vol. 184, p. 115480, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421008915

[3] Y. Yang, J. Xiong, L. Zhao, X. Wang, L. Hua, and L. Wu, "A novel method of blockchain cryptocurrency price prediction using fractional grey model," *Fractal and Fractional*, vol. 7, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2504-3110/7/7/547

[4] J. Chen, "Analysis of bitcoin price prediction using machine learning," *Journal of Risk and Financial Management*, vol. 16, no. 1, 2023. [Online]. Available: https://www.mdpi.com/1911-8074/16/1/51

[5] H. Sebastião and P. Godinho, "Forecasting and trading cryptocurrencies with machine learning under changing market conditions," *Financial Innovation*, vol. 7, no. 1, p. 3, January 6 2021. [Online]. Available: https://doi.org/10.1186/s40854-020-00217-x