

Analysis of Bitcoin Prediction using Transparent and Opaque Machine Learning Algorithms

Kavita R. Singh¹, Anushri V. Jamar², Pracheta P. Khadgi³, Smita Kapse⁴ and Ajay Kumar⁵

¹⁻⁴Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India
Email: singhkavita19@gmail.com, shrijamar@gmail.com, prachetakhadgi@gmail.com, smitarkapse@gmail.com

⁵School of Big Data and Computer Science, Guiyang City, Guizhou Province, China
Email: ajay.phdcse@gmail.com

Abstract—The unpredictable nature of cryptocurrency markets, particularly Bitcoin, has attracted significant attention from investors, researchers, and financial institutions seeking to understand and predict price movements. This study aims to analyse the effectiveness of both transparent and opaque machine learning algorithms in predicting Bitcoin prices. Transparent algorithms, such as linear regression and decision trees, provide clear insights into the factors influencing predictions. On the other hand, opaque algorithms, like ensemble methods, offer complex, non-linear models that may capture intricate patterns in the data.

The main analysis is to lower the mean absolute error and enhance the r^2 square, both key indicators of forecast accuracy. The study presents that linear regression and random forest outperformed others, achieving the smallest error and highest variance explanation, suggesting their effectiveness in Bitcoin price prediction with and without outliers respectively. In contrast, the support vector machine underperforms compared to all other machine learning algorithms under consideration. This paper sheds light on the most capable machine-learning techniques for Bitcoin forecasting, aiding trader investment choices and guiding academic study in the field of prediction.

Index Terms— Bitcoin, forecasting, cryptocurrency.

I. INTRODUCTION

Bitcoin, as the first cryptocurrency, operates on a peer-to-peer network, enabling secure transactions without the need for intermediaries. This decentralized nature contributes to its security. Alongside Bitcoin, other prominent cryptocurrencies by market capitalization include Litecoin, Ethereum, and Ripple. To ensure secure transactions, Bitcoin utilizes a binary data format and employs blockchain technology, which involves complex mathematical formulas and encryption techniques. These measures guarantee the integrity and confidentiality of transactions.

The future of Bitcoin is influenced by various factors, which can be broadly categorized into three groups: economic reasons, external sentiment, and blockchain statistics. Blockchain statistics encompass metrics such as block size, transaction volume, median confirmation time, hash rate, transaction cost, mining difficulty, total number of unique bitcoins, average confirmation time, number of unique addresses, total bitcoin supply, difficulty in finding a block, market capitalization, and number of orphaned blocks. These statistics have an impact on the price of Bitcoin. Machine learning algorithms offer a potential solution to this problem, as they can be trained to detect fake news automatically.

Economic factors also play a significant role in shaping Bitcoin's future. These factors include closing price, trade volume, volatility, opening and closing prices, as well as the highest and lowest prices observed in the market.

Researchers often analyse data from various sources, including Wikipedia, Twitter, Reddit, and Google Trends, to explore potential associations between human sentiment and Bitcoin's performance. By examining these sources, researchers aim to gain insights into how external sentiment can influence the cryptocurrency's price and market trends. By studying the interplay blockchain statistics, researchers can develop models and algorithms to forecast Bitcoin's future price movements. In this paper, we present a deeper understanding of analysis of different machine learning algorithms for the bitcoin predictions that can assist traders and investors in making informed decisions.

The organization of the paper is as follow: section 2 presents the literature survey followed by work done in section 3. Finally, the experimentation and result analysis has been discussed in section 4 followed by conclusion in section 5.

II. LITERATURE REVIEW

The paper analysis finds that the predictions are largely based on either daily intervals or high-frequency data (5-minute prediction). In his research, he implemented Logistics Regression which gives 64.84% for daily prediction, and XG Boost high-frequency predictions with an accuracy of 59.4% [1]. The paper analysed various models such as Theil-Sen Regression, Huber Regression, long short-term memory (LSTM), and Gated Recurrent Unit (GRU). Some libraries used by them are Keras, Scikitlearn, and TensorFlow. Their research presents that among all algorithms tested, GRU demonstrated superior performance [2]. In this paper, the dataset is split into two chunks serving as a dataset based on time periods of days frequency. The Linear Regression model trained with chunk 1 shows a higher prediction rate than chunk 2. The chunk 1 exhibits a percentage error rate of 3.03% and a mean prediction accuracy of 96.97%. On the other side, chunk 2 shows a mean prediction accuracy of 55.11%

This research in [3] utilized deep learning algorithms like RNN (Recurrent Neural Network), LSTM (Long short-term memory), and ARIMA (Autoregressive Integrated Moving Average). The authors conducted a thorough analysis by comparing different and created an evaluation table with parameters such as temporal length, sensitivity, specificity, precision, accuracy, and RMSE. Their finding indicated that RNN and LSTM are highly effective for bitcoin prediction [4]. This research used Bayesian Regression, and Random Forest and concluded that both performed well and were chosen as the best method to solve bitcoin prediction problems. They also demonstrated the data using graphs by log transform normalization technique, MATLAB in-built function and boxcox normalization technique

The authors have done research on [5] the bitcoin price prediction is based on Root Mean Square Error (RMSE), using various deep learning models like convolutional neural network (CNN), Long short-term memory and Grated Recurrent Unit (GRU). The RMSE of the LSTM model is the least. Thus, LSTM proves more capable of predicting long-term dependencies as compared to CNN and GRU [6] The researchers explore the relationship between the features of Bitcoin and next-day change in the price of Bitcoin using an Artificial Neural Network ensemble approach called genetic algorithm-based selective neural network ensemble, constructed using a Multilayered perceptron as the base model for each of the neural network in the ensemble. Out of all these MLP performed best in the ensemble generating approximately 53% in returns [7] This research was undertaken during the COVID-19 pandemic. In this research, they implemented Random Forest, Support Vector Machine, AdaBoost, and Decision Tree. Hence, found that: Twitter data can improve the performance of models; People consider information within 5 days when they make decisions on investments; Support vector machine does not perform well in predicting Bitcoin return rate or price trend [8] The predictions provided in this analysis were generated using the Autoregressive Integrated Moving Average (ARIMA) method, known for its ability to achieve high accuracy when predicting short-term outcomes. Specifically, the ARIMA (4,1,4) models yielded the smallest MAPE (Mean Absolute Percentage Error) values of 0.87 for the next one-day prediction and 5.98 for the next seven days. Hence, ARIMA (4, 1, 4) is a reliable predictive tool for Bitcoin fluctuations within a time frame of one to seven days in this paper [9]. This paper describes the Multilayer Perceptron (MLP) and the Recurrent Neural Networks (RNN) models. The result shows that long-term prediction has a better result than short-term prediction. Multilayer Perceptron outperforms Recurrent Neural Networks with accuracy of 81.3%, precision 81%, and recall 94.7%.

III. METHODOLOGY

In this paper, we present an analysis of Bitcoin Prediction using machine learning algorithms to understand the explainability of the algorithms versus accuracy of the algorithm. Explainability refers to the ability to understand and interpret the decisions made by a machine learning model. It involves providing clear and understandable reasons for why the model produces a particular output. In many real-world applications, such as stocks, healthcare, finance, and criminal justice, the need for model transparency and accountability is critical. Stakeholders, including end-users, regulators, and domain experts, may require explanations for the decisions made by the model. Some models, like decision trees, k-NN and linear models such as linear regression, are inherently more interpretable. So, such models can be categorized as transparent model. Random forest, SVM and XgBoost, can be applied to complex models to provide local or global explanations. Such machine learning models often achieve high accuracy but may lack transparency. Such models can be categorized as opaque model. On the other hand, accuracy is a measure of how well a model performs in terms of correctly predicting the target variable. It reflects the overall correctness of the model's predictions. Time and again, opaque models, such as ensembling models, often achieve high accuracy but may lack transparency. Ensuring accuracy involves using appropriate evaluation metrics and optimizing the model based on the specific problem requirements. There is often a trade-off between explainability and accuracy. Models that are more interpretable may sacrifice some predictive performance, while highly accurate models might be more complex and harder to interpret. Striking the right balance depends on the application. For instance, in healthcare, a more interpretable model might be preferred to ensure that medical professionals can trust and understand the decisions made by the model.

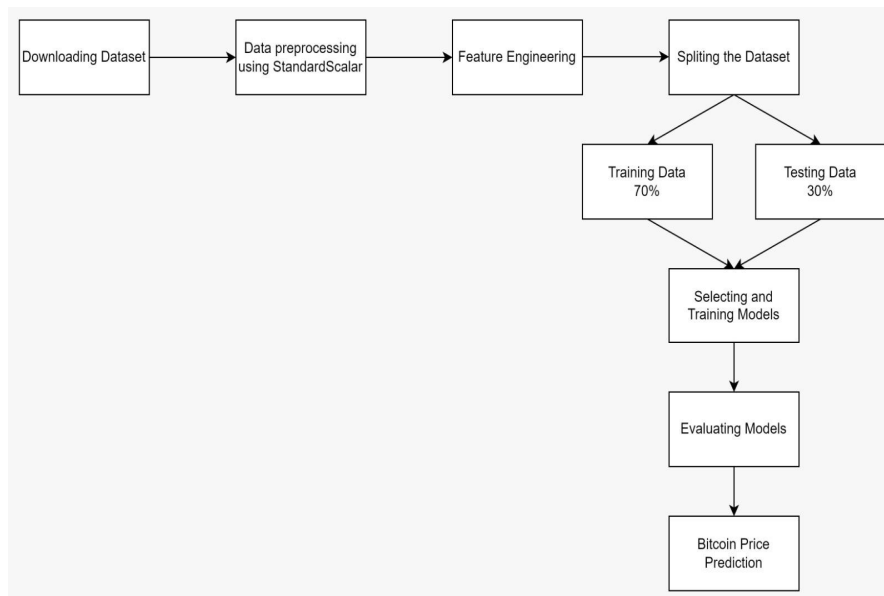


Figure 1: Flowgraph of the work done

IV. EXPERIMENTATION AND RESULT

This section of the paper presents an exhaustive experimentation to perform analysis of different machine learning algorithm for Bitcoin prediction. We performed the experimentation on machine with a configuration of Intel (R) Core (TM) i5 processor with Windows 7 operating system. The dataset used in the experimentation to find optimal algorithm machine learning is Bitcoin Dataset. Bitcoin Dataset contains the data spanning from 2014-2021. The different attributes of the datasets are 'Date', 'Open', 'Low', 'High', 'Close', 'Adj Close', 'Volume'. 'Open' attribute is the price from the first transaction of a trading day. 'High' attribute is the maximum price in a trading day. 'Low' attribute is the minimum price in a trading day. 'Close' attribute is the price from the last transaction of a trading day. 'Adj Close' attribute is a closing price adjusted to reflect the value after accounting for any corporate actions. 'Volume' attribute is the number of units traded in a day. To understand the relation between each of the attributes, a correlation-matrix is computed between all the attributes as depicted in figure 2.

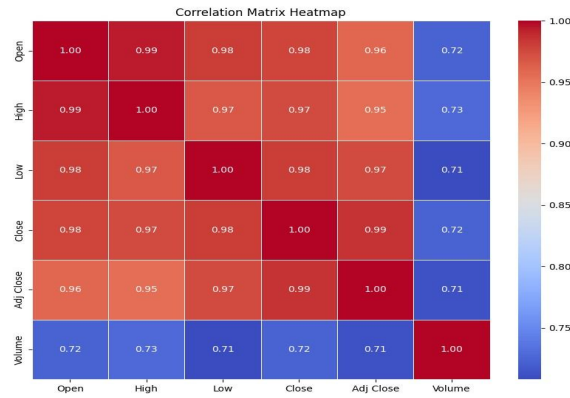


Figure 2: Correlation matrix

From the figure 2 it is evident that correlation coefficient lies between 0 to 1. For instance, for any give two variables let's say x and y , the correlation coefficient 1, indicates that x and y are highly correlated. This can be interpreted as direct proportionality between the attributes x and y . Based on the computed correlation coefficient values between all the mentioned attributes, it was observed that the correlation coefficient value between 'Volume' attribute and other attributes is comparatively lesser. Therefore, attribute 'Volume' was dropped from the dataset. Secondly, the main aim of the proposed work was to predict the 'Close Price' based on 'Open', 'High', 'Low' and 'Adj Close' values. Currently, the paper doesn't focus more on the timelines and therefore the 'Date' attribute was also dropped. The resultant Bitcoin dataset with only following attributes 'Open', 'Low', 'High', 'Close', 'Adj Close' are further used for data pre-processing. At the data pre-processing stage, the null and the duplicate values were removed. Next standardization was performed on the resultant Bitcoin dataset as it was very scattered with large variability and many algorithms are sensitive to the scale of data. To evaluate the different machine learning algorithms, the dataset was divided in 70-30 ratio. The closing price prediction experimentation was performed with machine learning algorithms like Decision Tree, k-NN, SVM, Linear Regression, Random Forest, and XgBoost.

Decision tree was implemented and experimented for different parameters namely, (i) error terms like "absolute error", "Poisson", "Friedmans", "squared error", (ii) depth of the tree plays a major role in the prediction of the desired output. We experimented to find the optimal error criteria and the optimal depth of the tree for the concise desirable Bitcoin prediction.

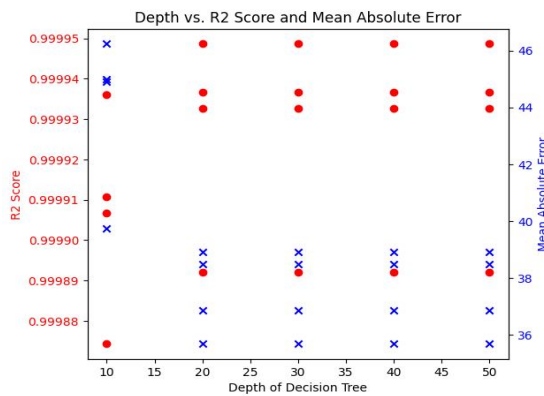


Figure 3: Variation of R2 score and mean absolute error with respect to depth of tree

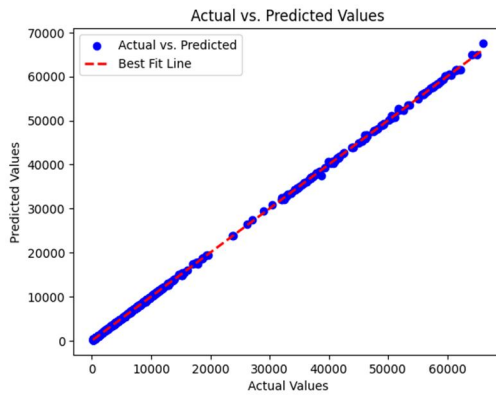


Figure 4: Representation of an actual and a predicted output along the best fit line of the decision tree

Variance reduction is a critical metric in decision tree algorithms, especially for regression tasks. It measures the decrease in variance of the target variable when a node is split into child nodes. By maximizing variance reduction, decision trees create splits that result in subsets with lower variance in the target variable, enhancing predictive accuracy. This concept underscores the importance of selecting informative splits to achieve accurate predictions in regression tasks. Figure 3 depicts the mean absolute error and r^2 score for various depths and we find that for depth greater than equal to 20 has given the optimal results due to less mean square error and significantly more r^2 score hence 20 or more than it is optimal here we are choosing 20.

Next, for k -NN this paper focuses on the calculation of the optimal number of nearest neighbor to have the desirable output prediction as demonstrated in figure 5. Further, we trained the model using linear regression and predicted the output as depicted figure 6.

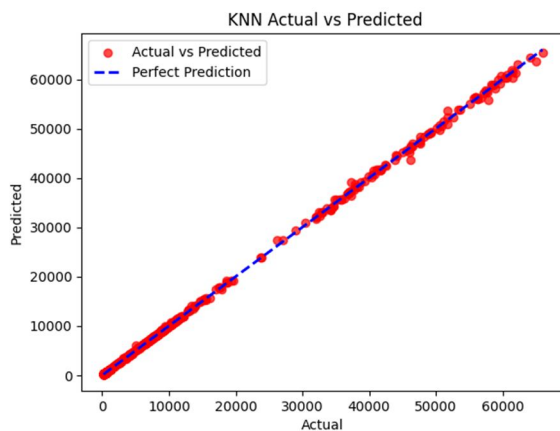


Figure 5: Representation of an actual and a predicted output along the best fit line of k -NN

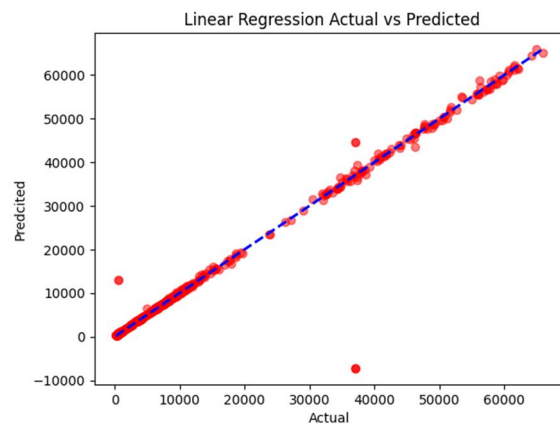


Figure 6: Representation of an actual and a predicted output along the best fit line of the linear regression

Next, we implemented opaque and non-linear machine learning algorithms such as Random Forest, XgBoost and SVR to understand the behaviour for bit coin prediction. Random Forest model computes the optimal number of trees required for a desirable output as shown in figure 7. Similarly, results for XgBoost and SVR is shown in figure 8 and figure 9. From the graph shown in figure 10, it can be observed that SVR has least r^2 score and other transparent and opaque algorithms have almost similar r^2 score for Bitcoin Predictions. Due to high non-linearity of Bitcoin price data, SVR may have struggle to capture it effectively, leading to lower R^2 scores. Although, during experimentation we fine-tuned SVR's hyperparameters still it was observed that it can be a challenging task. If the hyperparameters are not tuned properly, SVR might not be able to capture the underlying patterns in the Bitcoin price data effectively.

V. CONCLUSIONS AND FUTURE SCOPE

In conclusion, the choice between transparent and opaque machine learning algorithms for Bitcoin prediction depends on various factors such as interpretability requirements, data complexity, model performance, and risk tolerance. A careful evaluation of these factors is necessary to select the most suitable approach for a given application. Ultimately, the balance between explainability and accuracy is context-dependent and requires careful consideration of the specific requirements and constraints of the application at hand. Additionally, exploring hybrid approaches and integrating domain knowledge can further enhance the effectiveness of Bitcoin prediction models.

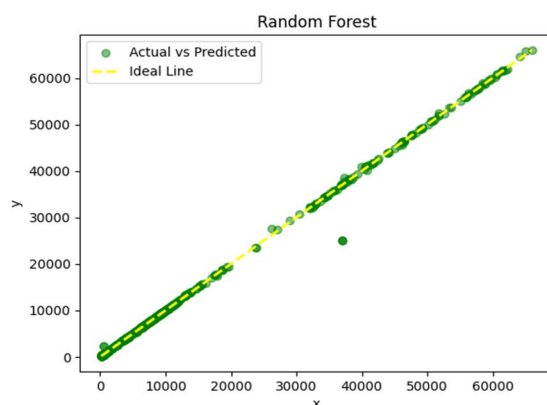


Figure 7: Representation of an actual and a predicted output along the ideal line of the Random Forest

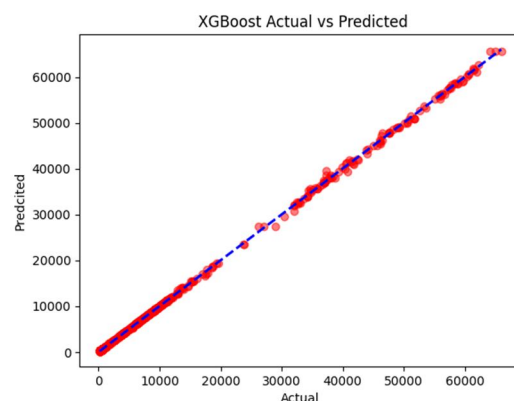


Figure 8: Representation of an actual and a predicted output along the Best Fit Line of the XgBoost

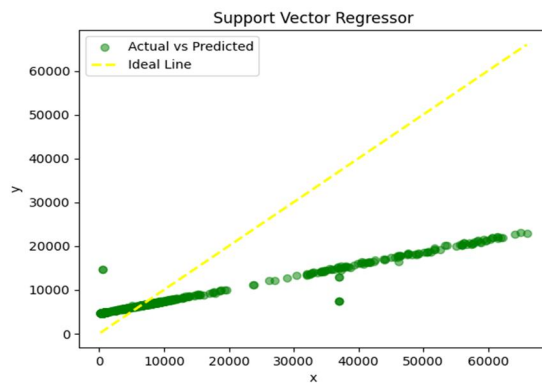


Figure 9: Representation of an actual and a predicted output along the Ideal line of the Support Vector Reg

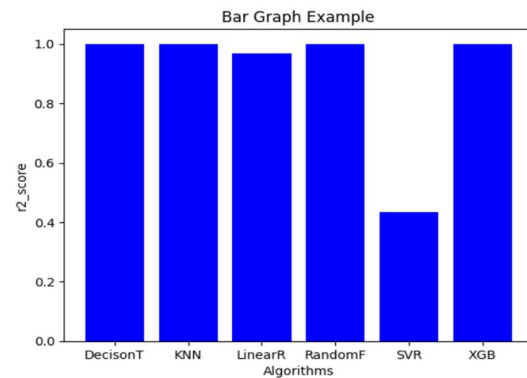


Figure 10: Comparative Analysis of r2 score of all machine learning algorithms under consideration

REFERENCES

- [1] Ranjan, Sumit & Kayal, Parthajit & Saraf, Malvika. (2022). Bitcoin Price Prediction: A Machine Learning Sample Dimension Approach. Computational Economics. 61. 1-20. 10.1007/s10614-022-10262-6.
- [2] T. Phaladisailoed and T. Numnonda, "Machine Learning Models Comparison for Bitcoin Price Prediction," 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE), Bali, Indonesia, 2018, pp. 506-511, doi: 10.1109/ICITEED.2018.8534911
- [3] M. Ali and S. Shatabda, "A Data Selection Methodology to Train Linear Regression Model to Predict Bitcoin Price," 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), Dhaka, Bangladesh, 2020, pp. 330-335, doi: 10.1109/ICAICT51780.2020.9333525
- [4] S. McNally, J. Roche and S. Caton, "Predicting the Price of Bitcoin Using Machine Learning," 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), Cambridge, UK, 2018, pp. 339-343, doi: 10.1109/PDP2018.2018.00060.
- [5] S. Velankar, S. Valecha and S. Maji, "Bitcoin price prediction using machine learning," 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea (South), 2018, pp. 144-147, doi: 10.23919/ICACT.2018.8323676.
- [6] A. Aggarwal, I. Gupta, N. Garg and A. Goel, "Deep Learning Approach to Determine the Impact of Socio-Economic Factors on Bitcoin Price Prediction," 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 2019, pp. 1-5, doi: 10.1109/IC3.2019.8844928.
- [7] E. Sin and L. Wang, "Bitcoin price prediction using ensembles of neural networks," 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, China, 2017, pp. 666-671, doi: 10.1109/FSKD.2017.8393351.
- [8] J. Luo, "Bitcoin price prediction in the time of COVID-19," 2020 Management Science Informatization and Economic Innovation Development Conference (MSIED), Guangzhou, China, 2020, pp. 243-247, doi: 10.1109/MSIED52046.2020.00050.
- [9] M. Wirawan, T. Widiyaningtyas and M. M. Hasan, "Short Term Prediction on Bitcoin Price Using ARIMA Method," 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2019, pp. 260-265, doi: 10.1109/ISEMANTIC.2019.8884257.
- [10] R. Albariqi and E. Winarko, "Prediction of Bitcoin Price Change using Neural Networks," 2020 International Conference on Smart Technology and Applications (ICoSTA), Surabaya, Indonesia, 2020, pp. 1-4, doi: 10.1109/ICoSTA48221.2020.1570610936.