Bitcoin price prediction based on fear & greed index

Qichuan Huang*

Computer Science, CSE Department, The Chinese University of Hong Kong, 999077, Hong kong, China

Abstract. This study investigates the Fear & Greed Index, an indicator designed to reflect market sentiment regarding Bitcoin price, intending to utilize it as a predictive parameter for future price fluctuations. Due to the substantial volatility in Bitcoin prices and its significant influence on prediction outcomes, the dataset was preprocessed through monthly filtering and normalization. To forecast Bitcoin prices, an array of machine learning algorithms, including linear regression, random forest, and XGBoost, as well as their enhanced counterparts, were employed. The optimal model was identified by comparing the Grid Search XGBoost analysis results. This research holds implications for accurately predicting Bitcoin prices and underscores the impact of market sentiment on its valuation.

1 Introduction

The cryptocurrency market has witnessed substantial growth and volatility in recent years, with Bitcoin emerging as the most prominent and widely adopted digital currency. The erratic nature of Bitcoin's price has sparked heightened interest in comprehending and forecasting its price movements.

The cryptocurrency market has witnessed substantial growth and volatility in recent years, with Bitcoin emerging as the most prominent and widely adopted digital currency. The erratic nature of Bitcoin's price has sparked heightened interest in comprehending and forecasting its price movements.

Numerous studies have approached Bitcoin price prediction from various perspectives. For instance, some research has employed blockchain data and machine learning techniques to develop models considering factors such as Bitcoin properties, network, transactions, market, attention, and gold spot price [1]. Other studies have analyzed the sources and influences of Bitcoin price volatility, encompassing market liquidity, trading volume, market sentiment, policy events, and cyberattacks [2-5]. Furthermore, investigations have explored the relationship between Bitcoin prices and financial crises and the potential of Bitcoin as an early warning indicator or hedge against financial crises [6,7]. Additional research has examined the bidirectional causality between Bitcoin price and network effects, alongside the impact of these network effects on Bitcoin price volatility [8,9]. Lastly, some studies have scrutinized the long-run equilibrium relationship between Bitcoin price and energy consumption, as well as the influence of energy consumption on Bitcoin price stability [10, 11]. These diverse perspectives and approaches to Bitcoin price forecasting underscore the multi-faceted nature of factors influencing its valuation.

Aside from these dimensions, another potential driver of Bitcoin's price that has garnered attention in recent years is market sentiment. Market sentiment refers to the collective attitude of investors and traders towards a specific asset or market, which various factors, including economic news, social media discourse, and individual investor behavior, can shape.

Numerous studies have investigated the relationship between market sentiment and Bitcoin's price. For instance, a study by Nadarajah and Chu discovered a robust positive correlation between sentiment and price clustering within the Bitcoin market [12]. Another study by Dimpfl and Kleiman employed transaction-level data from Coinbase to derive a measure of investor sentiment, uncovering a significant and robust relationship between escalating sentiment and price increases and vice versa [13].

Chiah et al. investigated the influence of investor sentiment on Bitcoin returns and conditional volatilities during the COVID-19 era, revealing that investor sentiment positively impacted Bitcoin returns and their volatility, particularly after the COVID-19 outbreak [14]. Moreover, BeinCrypto and the International Monetary Fund offer overviews of the primary factors that affect Bitcoin's price, encompassing supply and demand, news and public opinion, emotions, innovation and competition, regulation and security, institutional adoption, and market manipulation.

This study's objective is to expand upon the existing literature by examining the relationship between market sentiment and Bitcoin's price, utilizing the Fear & Greed Index as a proxy for market sentiment. The study will assess the predictive capability of this index in forecasting Bitcoin's price fluctuations and investigate potential enhancements to the predictive accuracy of the study's models by incorporating additional relevant factors and scrutinizing the interactions between these factors and the Fear & Greed Index.

*Corresponding author: 1155173886@link.cuhk.edu.hk

2 Data and Method

2.1 Data

2.1.1 Sample and Variables

This study utilizes a dataset from Bitcoin & Fear and Greed on Kaggle, originating from the alternative. Me. The dataset comprises five variables: Date, Value, Value_Classification, BTC_Closing, and BTC_Volume. Date records the date information, spanning from February 1, 2018, to the present day. Value is a numerical representation of the Fear & Greed Index, ranging from 0 to 100. Value_Classification categorizes the index into five levels: Extreme Fear, Fear, Neutral, Greed, and Extreme Greed, reflecting the intensity of market sentiment concerning Bitcoin's price. BTC_Closing denotes the closing price of Bitcoin on the specified date, while BTC_Volume represents the market volume of Bitcoin on that date.

2.1.2 First Preprocesse

The dataset is initially preprocessed to guarantee data quality. The study begins by checking for missing values. Upon inspection, it is found that data for only three days are missing. Given that the study's dataset consists of 1,885 rows in total, removing these three rows does not significantly impact the results. Therefore, to preserve the accuracy of the study's analysis, these three rows are removed from the dataset.

2.1.3 Descriptive Statistics

To gain a deeper understanding of the study's dataset, the study provides some descriptive statistics for the Value and BTC_Closing variables in a table. Table 1 presents the descriptive statistics for the dataset's two variables of interest.

Table 1. Statistics of dataset

Statistics	Value	BTC_Closing
Mean:	42.39	20598.99
Median	39	11,464.51
Mode	24	6,741.75
Variance	487.85	279,449,800
Max	95	67566.82813
Min	5	3236.761719

These statistics offer a preliminary understanding of the distribution and central tendencies of the Fear & Greed Index and Bitcoin closing prices. In the subsequent sections, the study will delve deeper into the relationship between these variables and employ machine-learning techniques to forecast Bitcoin prices based on the Fear & Greed Index.

2.1.4 Correlation Analysis

To explore the relationship between the fear and greed index (Value) and the closing price of Bitcoin (BTC_Closing), this study conducted a correlation analysis and visualized it in Figure 1 and Figure 2.

Figures 1 and 2 show that the correlation between BTC_Closing and Value is not strong, with a correlation coefficient of 0.24. However, the visualization shows that the data points cluster around these three lines. This lower correlation may be due to the significant ups and downs of the Bitcoin price, as shown in Figure 3, which results in significant value changes over a long time horizon. Also, although the Bitcoin price and the index are correlated, the different bases produce a stratified relationship.

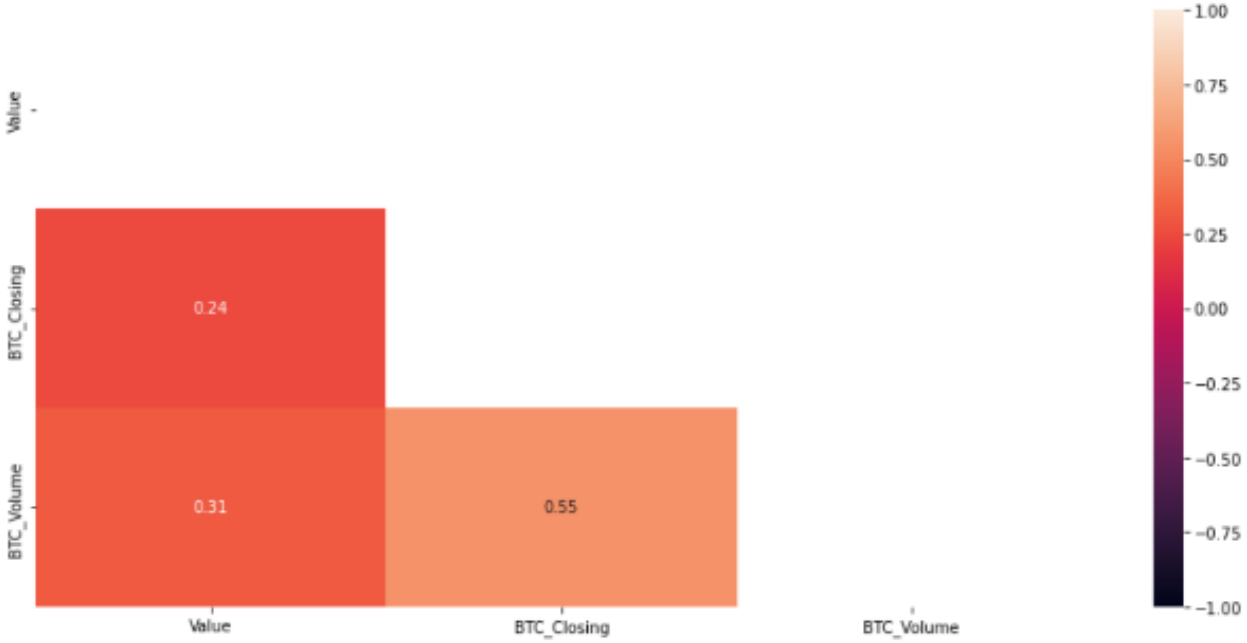

Fig. 1. The correlation coefficient of each index

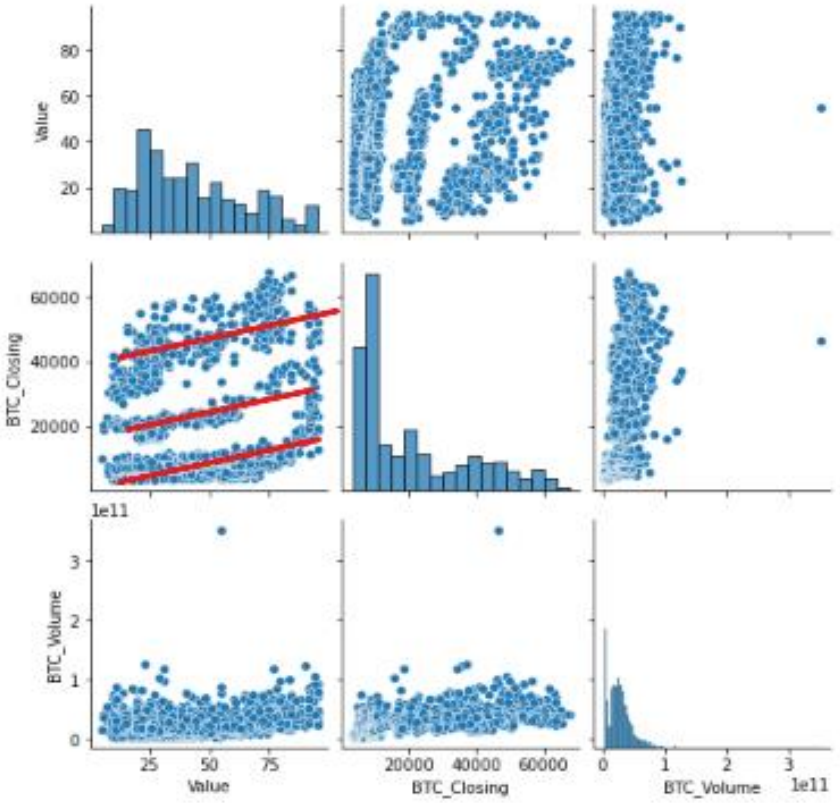

Fig. 2. Correlation between visualization processing indices

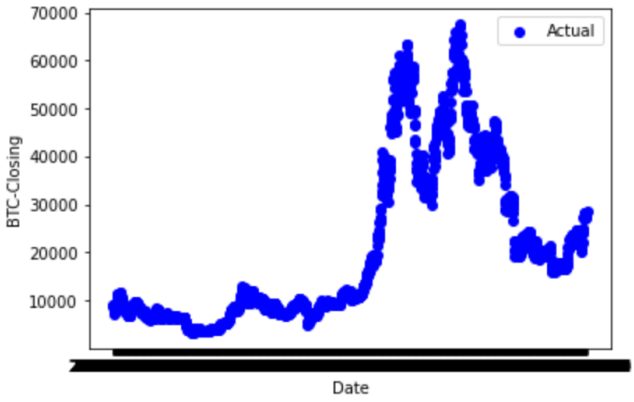

Fig. 3. The correlation coefficient of each index

2.1.5 Additional Data Preprocessing

Considering these findings, the study further preprocesses the data to account for the large fluctuations in Bitcoin prices on specific days. Three additional data preprocessing steps are performed:

Divide data by month and select months with higher correlations.

To mitigate the impact of large fluctuations in Bitcoin prices, this study decided to split the dataset by month and investigate the correlation between Bitcoin prices and the Fear & Greed Index for each month. By excluding individual months where prices fluctuate significantly due to other factors, we aim to establish a correlation between Bitcoin prices and the index as long as we can achieve a sufficient correlation for most months.

The final results revealed that the correlation was more significant than 0.6 for a total of 40 weeks out of 62 months or 64.52%. Additionally, a total of 55 out of 62 months exhibited a correlation greater than 0.3, amounting to a rate of 88.71%.

These findings demonstrate that in the majority of months (64.52%), there is a strong relationship between the Fear & Greed Index and Bitcoin prices, as indicated by a correlation greater than 0.6. In most months (88.71%), a moderate relationship exists, as suggested by a correlation greater than 0.3. By excluding months with significant fluctuations in Bitcoin prices, the study can conclude that there is a certain degree of correlation between Bitcoin prices and the Fear & Greed Index. The study retains data from months with a correlation greater than 0.6 for further analysis. Consequently, the study can establish that there is a correlation between Bitcoin prices and the Fear & Greed Index.

2.1.5.1 Bitcoin price normalization

Although the Bitcoin prices are separated by month, in practice, there are still considerable variations in the price of Bitcoin between each month. To harmonize the range of values and reduce errors, the study employs the first day of each month as the base and normalizes the Bitcoin price for the remaining days.

Normalization is a technique used to transform the scale of variables so that they can be compared on a common ground. One standard normalization method is the Min-Max scaling. In this study, the study can use the first day of each month as the base and normalize the

Bitcoin price for the rest of the days using the following formula:

$$P_{normalized\ price} = \frac{(P_{price} - P_{base\ price})}{P_{base\ price}} \quad (1)$$

Applying this formula transforms the Bitcoin price for each day within a month into a normalized value relative to the base price (i.e., the price on the first day of the month). This normalization process helps to reduce the impact of significant price variations between months. It allows for more accurate comparisons and analysis of the relationship between the Fear & Greed Index and Bitcoin prices.

2.1.5.2 Split Data into Training and Test Sets

After normalizing the dataset and dividing it by month, the next step is to split the dataset into a training set and a test set in an 8:2 ratio. This division allows us to train machine learning models on most data (80%) and evaluate their performance on the remaining unseen data (20%).

To recap, the following preprocessing steps have been completed:

1. Checking for and removing missing values
2. Calculating descriptive statistics
3. Smoothing the data using a moving average or other smoothing techniques
4. Removing outliers
5. Dividing the data by month and selecting months with higher correlations
6. Normalizing Bitcoin prices using the first day of each month as the base
7. Splitting the dataset into a training set and a test set in an 8:2 ratio

With the preprocessed dataset, the study can apply different machine-learning models to build and evaluate their effectiveness in predicting Bitcoin prices based on the Fear & Greed Index and other relevant factors. Standard models include linear regression, decision trees, random forests, and neural networks. After training and evaluating each model, the study can compare their performance and select the most suitable model for its specific use case.

2.2 Methodology

2.2.1 Linear Regression

Linear Regression (LR) is a foundational supervised learning algorithm employed to ascertain a linear association between input features (independent variables) and continuous output targets (dependent variables). The primary objective of the algorithm is to discover a linear function that minimizes the sum of squared prediction errors and effectively models the underlying data points.

2.2.1.1 Definition

Linear regression is a predictive model for estimating a continuous target variable based on input features. The model posits a linear relationship between input features and the target variable. The primary aim of linear regression is to determine a linear function that minimizes the sum of squared prediction errors.

2.2.1.2 Formula

The linear regression formula can be denoted as:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_n \times x_n + \varepsilon \quad (2)$$

In this formula, y is the target variable, xi represents input features, β i denotes model parameters, β 0 is the intercept term, and ε is the error term.

2.2.1.3 Build model

The study creates a linear regression model, fits it with training data, produces predictions using test data, calculates performance metrics, and visualizes predicted results compared to actual values.

2.2.2 Polynomial Regression

Polynomial Regression is an extension of linear regression, allowing for modelling non-linear relationships between input features and continuous output targets. By introducing higher-degree polynomial terms into the linear regression equation, polynomial regression can effectively capture non-linear patterns in the data.

After observing satisfactory performance from linear regression, the study applies polynomial regression for further improvement. In this analysis, the study chooses a 5-degree polynomial for fitting.

Combining higher accuracy and simplicity of calculation, this study creates a polynomial regression model with a degree of 5, fits it with training data, produces predictions using test data, calculates performance metrics, and visualizes predicted results compared to actual values.

2.2.3 Random Forest

This section shows the improvement achieved using polynomial regression compared to linear regression. However, the results are still unsatisfactory, leading to exploring Random Forest regression as a further alternative.

Random Forest is a supervised learning algorithm that can be employed for both classification and regression tasks. It is an ensemble method that combines multiple decision trees to enhance prediction performance.

2.2.3.1 Definition

Random Forest is an ensemble learning technique based on decision trees. It constructs multiple decision trees

and aggregates their predictions (through voting or averaging) to make a final prediction. Random Forest employs bootstrap sampling to generate multiple training datasets from the original data and builds a decision tree on each training set. Additionally, it randomly selects a subset of features during the decision tree construction process to increase model diversity.

2.2.3.2 Formula

Random Forest has no specific formula since it is an ensemble method based on multiple decision trees. However, the prediction process for Random Forest can be represented as follows:

$$y_{pred} = \frac{1}{N} \times \sum y_i \quad (3)$$

Where y_i is the prediction of the ith decision tree, and N is the number of decision trees. The Random Forest prediction is the average of all decision tree predictions.

2.2.3.3 Improvement

Although the Random Forest regression results show improvement over polynomial regression, the performance is still unsatisfactory. Therefore, Grid Search is employed to find the optimal parameters and perform Random Forest regression fitting once again.

Grid Search is a hyperparameter optimization technique that aims to identify the best combination of hyperparameters for a given model by systematically searching through the parameter space. It performs an exhaustive search over the specified parameter grid with cross-validated performance metrics, selecting the optimal parameters that maximize the model's performance.

A Random Forest Regressor object is created in the subsequent implementation, and a parameter grid is defined to search for the best hyperparameters. A Grid Search CV object is created, and the model is fitted using the training data. After identifying the best parameters, predictions are made using the test data, and performance metrics are computed. Finally, the predicted and actual values are visualized.

By incorporating Grid Search into the approach, the aim is to optimize the Random Forest regression model's performance and achieve more accurate predictions.

2.2.4 XGBoost

Although the results achieved through random forest fitting are pretty satisfactory, with an R^2 value of 0.95, which is very close to 1, the study still wants to try one last method: XGBoost.

XGBoost is a supervised learning algorithm that can be employed for classification, regression, and ranking tasks. It is an ensemble method that combines multiple decision trees using gradient-boosting techniques to enhance prediction performance.

2.2.4.1 Definition

XGBoost, or eXtreme Gradient Boosting, is an ensemble learning technique based on gradient boosting and decision trees. It constructs multiple decision trees and optimizes the loss function by iteratively adding weak learners. XGBoost aims to enhance computational performance, reduce the risk of overfitting, and allow customization of loss functions and evaluation metrics.

XGBoost is a supervised learning algorithm that can be employed for classification, regression, and ranking tasks. It is an ensemble method that combines multiple decision trees using gradient-boosting techniques to enhance prediction performance.

2.2.4.2 Formula

XGBoost does not have a specific formula, as it is an ensemble method based on gradient boosting and decision trees. However, the prediction process for XGBoost can be represented as follows:

$$y_{pred} = \sum_{k=1}^{k} f_k(x_i) \quad (4)$$

Where $f_k(x_i)$ is the prediction of the kth decision tree, and K is the number of boosting rounds. The XGBoost prediction is the sum of all decision tree predictions, with each tree weighted by its learning rate.

2.2.4.3 Improvement

Although the results from Random Forest regression show improvement over polynomial regression, XGBoost may further enhance the performance. To optimize the XGBoost regression model, hyperparameter tuning techniques, such as Grid Search or Randomized Search, can be employed.

In the subsequent implementation, an XGBoost Regressor object is created, and a parameter grid is defined to search for the best hyperparameters. A Grid Search CV or Randomized Search CV object is created, and the model is fitted using the training data. After identifying the best parameters, predictions are made using the test data, and performance metrics are computed. Finally, the predicted and actual values are visualized.

By incorporating hyperparameter tuning into the approach, the aim is to optimize the XGBoost regression model's performance and achieve more accurate predictions.

3 Results and Discussion

3.1 Results of LR

The Mean Squared Error (MSE) of 0.01 and R-squared value of 0.52 suggest a moderate fit for the model. The visualization results (figure 4), with the test set values roughly around the predicted straight line, further support this conclusion. While the fit is imperfect, it demonstrates a good relationship between the observed

and predicted values. Further improvements to the model can be explored through additional feature engineering, hyperparameter tuning, or trying alternative machine learning algorithms.

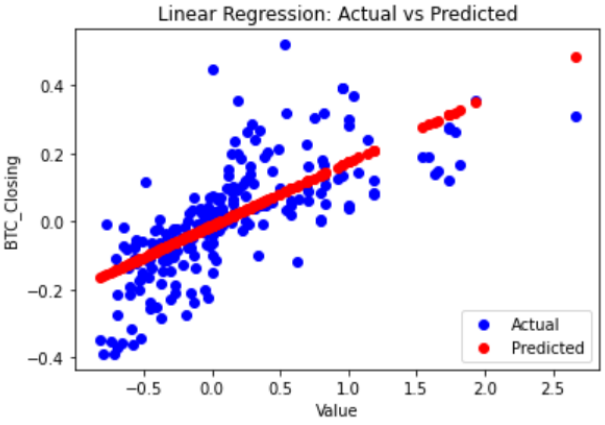

Fig. 4. The result of Linear Regression

3.2 Results of Polynomial Regression

The Mean Squared Error (MSE) of 0.01 and R-squared value of 0.60 for the fifth-order polynomial regression indicate a slightly better fit than the previous model. The visualization results (figure 5) also show that the curve fits the test set more closely, suggesting an improved relationship between the observed and predicted values. Although the model has progressed, there may still be room for further enhancements, such as exploring higher-order polynomials, additional feature engineering, or experimenting with other machine-learning algorithms.

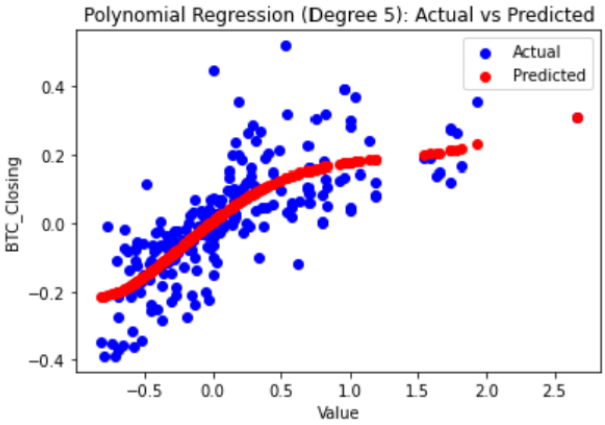

Fig. 5. The result of Polynomial Regression

3.3 Results of Random Forest Regression

The Mean Squared Error (MSE) of 0.01 and R-squared value of 0.63 for the Random Forest model indicate an even better fit than the previous models. The visualization results (figure 6) further support this observation, as the prediction results and the test set values are more closely aligned. This improvement suggests that the Random Forest model is more effective in capturing the underlying patterns in the data. While the model's performance has increased, additional enhancements could still be explored, such as hyperparameter tuning, feature engineering, or trying alternative machine learning algorithms.

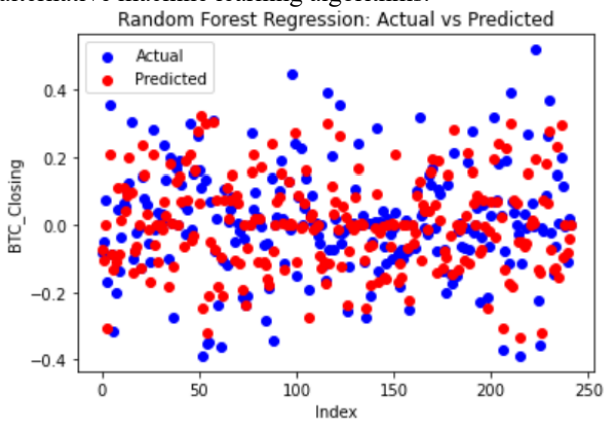

Fig. 6. The results of Random Forest Regression

After applying Grid Search, the results show a significant improvement in the R-squared value (0.95), indicating a solid fit between the predicted and actual values. The substantial increase in R-squared suggests that the model has benefited from the hyperparameter optimization, which may have led to a better understanding of the underlying patterns in the data (figure 7).

However, the Mean Squared Error (MSE) has increased to 12,293,304, which is unexpected given the improvement in R-squared. This discrepancy could be due to the increased complexity after the Grid Search or potential overfitting, where the model may perform exceptionally well on the training data but poorly on unseen data.

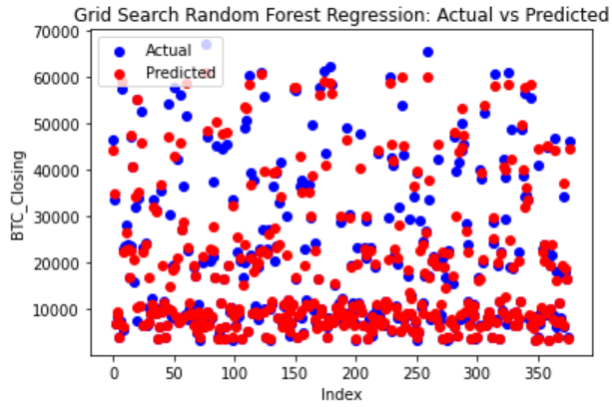

Fig. 7. The results of Grid Search Random Forest Regression

3.4 Results of XGBOOST

The Mean Squared Error (MSE) of 9,888,501.7 and R-squared value of 0.96 for the Random Forest model indicate a strong fit between the predicted and actual values. The visualization results (figure 8) further support this observation, as the prediction results and the test set values are closely aligned. This suggests that the model is highly effective in capturing the underlying patterns in the data.

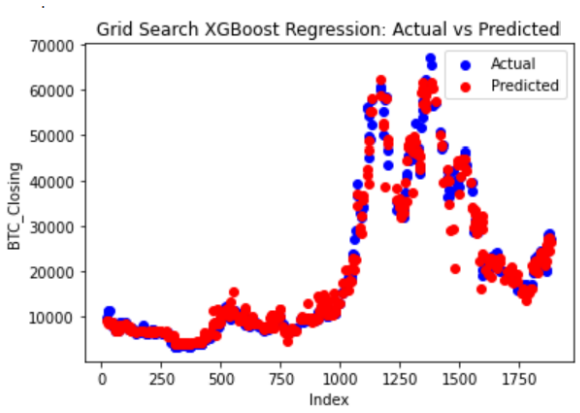

Fig. 8. The results of Grid Search XGBoost Regression

3.5 Discussion

Table 2 are summary of the results. When comparing the results of different models, it is essential to consider both the Mean Squared Error (MSE) and the Coefficient of Determination (R^2). A smaller MSE indicates a minor prediction error, while a larger R^2 suggests better descriptive performance. Due to the significant fluctuations in Bitcoin prices, the MSE results may be relatively large, making R^2 values particularly important.

Table 2. Results of models

	Linear Regression	Polynomial regression	Random Forest Regression	Grid Search Random Forest Regression	Grid Search XGBoost Regression
MSE	0.01	0.01	0.01	12293304.44	9888501.70
R^2	0.52	0.60	0.63	0.95	0.96

The XGBoost and Random Forest models with hyperparameter tuning show considerably higher R^2 values than the other models, indicating better descriptive performance. Among these two, XGBoost has a relatively smaller MSE, making it superior in prediction accuracy. Therefore, considering both its descriptive ability (R^2) and prediction accuracy (MSE), the XGBoost model performs better in predicting Bitcoin prices.

In summary, the study concludes that the Grid Search XGBoost model is the best model for predicting Bitcoin prices in this analysis.

4 Conclusion

In summary, the study concludes that the Grid Search XGBoost model is the best model for predicting Bitcoin prices based on the Fear & Greed Index using a machine learning model in Python. The study found that the Fear & Greed Index strongly correlates with Bitcoin prices. However, due to the high volatility of Bitcoin prices, it is not sufficient to study the relationship between the index and Bitcoin prices in isolation; the study must also consider their relationship over time. The study ultimately determined that Grid Search XGBoost yields the best prediction results by utilizing various models and model improvements for Bitcoin price prediction.

This study provides strong evidence that market sentiment, as represented by the Fear & Greed Index, is closely correlated with Bitcoin prices. This finding offers a foundation for future research on Bitcoin prices using alternative approaches influenced by market sentiment and a direction for incorporating market sentiment when constructing predictive models for Bitcoin prices.

It is essential to acknowledge that Bitcoin prices are influenced by various factors, which means the study's predictions are subject to a certain degree of error. The high MSE values of the optimal model also suggest a potential risk of overfitting. Therefore, given its highly volatile nature, the study advises exercising caution when investing in Bitcoin.

Further models and analyses could be improved by considering additional factors, incorporating more sophisticated features, or applying advanced machine-learning techniques. This would help to enhance the reliability and generalization capabilities of the models, allowing for more accurate predictions and better decision-making in the context of Bitcoin investments.

References

1. Z. Chen, C. Li, & W.Sun, Journal of Computational and Applied Mathematics, **365**, (2020)
2. A. H. Dyhrberg, Finance Research Letters, **16**, (2016)
3. Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D. N., & Giaglis, G. M. SSRN 2607167, (2015)
4. A. Hayes, SSRN 2579445, (2015)
5. L. Kristoufek, Scientific Reports, **3**, 1(2013)
6. E. Bouri, R. Gupta , A. K. Tiwari, & D. Roubaud, Finance Research Letters, **23**, (2017)
7. P. Ciaian, M. Rajcaniova, & D. A. Kancs, The economics of BitCoin price formation. Applied Economics, 48,19 (2016)
8. J. Bouoiyour, & R. Selmi, Annals of Economics and Finance, **16**, 2(2015)
9. D. Koutmos, Annals of Operations Research, **10**, (2018)
10. D.Malone, & K. J. O'Dwyer, *Bitcoin mining and its energy footprint*. In 25th IET Irish Signals & Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies , (2014)
11. H. Vranken, Current Opinion in Environmental Sustainability, **28**, (2017).

12. S. Nadarajah, & J. Chu, Economics Letters, **174**, (2019).
13. T. Dimpfl & V. Kleiman, Journal of Risk and Financial Management, **15**, 2 (2022).
14. M. Chiah, A. Zhong & L. Yao, Applied Economics Letters, **1**, (2021).