Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Predicting Bitcoin: An Ensemble Machine Learning Model Based on Social and Financial Features

NIANDI YANG[1], LINGJUN LU[1], FANGZHOU CAO[1], and KEXIN JIE[1]
[1]Department of computing, The Hong Kong Polytechnic University

ABSTRACT Cryptocurrencies are beginning to gain attention owing to the ongoing advancement of blockchain technology and the sharp increase in the price of Bitcoin. Thousands of cryptocurrencies are now available in the decentralized finance market. Due to their unique characteristics, cryptocurrencies can be used for both payment and investing activities, and their price swings are very different from those of fiat currencies and other conventional financial instruments. Therefore, it has become a popular area of study and curiosity to forecast the price patterns of cryptocurrencies.

Numerous financial professionals and academics have focused their attention on predicting the price of bitcoin. They have provided various study findings and discussed what variables affect the average cost of bitcoin and what research problems need to be addressed.

In this study, the random forest model, the gradient boosting model, and a combination of the two are empirically analyzed on time series data derived from the average price of Bitcoin from January 1, 2017, to January 10, 2022. Finally, accuracy and root mean square error are used as an indicator to evaluate each forecasting model. Comparing the experimental findings, the combined models perform better than using a single model for forecasting.

I. INTRODUCTION

A. BACKGROUND AND SIGNIFICANCE OF THE STUDY

Bitcoin and its algorithm were first described by Satoshi Nakamoto in 2008 in "Bitcoin: A Peer-to-Peer Electronic Cash System" [1], which opened the curtain on the development of cryptocurrencies. Bitcoin is the first distributed, anonymous cryptocurrency that is based on blockchain technology for encryption and does not rely on third-party financial institutions or receive central bank controls. Based on a variety of new technologies, Bitcoin has features that are superior to those of traditional currencies, notably transparency, decentralization, and security. These properties have led to it receiving widespread attention and attracting a large number of consumers and investors, with Bitcoin's market capitalization currently standing at approximately US$300 billion [2].

Research into Bitcoin initially focused on its security, legal and technical issues, and initially, the price of Bitcoin was determined by users negotiating with each other. As Bitcoin became better known and accepted by the public, the research direction began to focus on the financial properties of Bitcoin. However, cryptocurrencies have only been around for a short period of time compared to traditional currencies, and the associated price predictions are in their early stages. The price of cryptocurrencies can be influenced by more factors than conventional currencies, generating more significant volatility and a clear difference in openness. As shown by previous research, factors such as speculative bubbles [3]and government regulation of cryptocurrency assets can generate significant volatility in the price of bitcoin. And Bitcoin is a time series forecasting problem, and traditional time series forecasting methods may not be effective in predicting Bitcoin prices.

B. LITERATURE REVIEW

When considering the issue of Bitcoin price trends, the literature has included chiefly empirical work on analyzing the determinants. Sean McNally et al. [4]predicted the price of Bitcoin by selecting the closing price of Bitcoin, the opening price, the daily high and daily low, and blockchain data, i.e., mining difficulty and hash rate. After analyzing the relationship between the problem and HashRate, which are influenced by several factors, these two values will be disregarded in the model of the report. Professor

Bollen [5]has combined information from Wall Street with millions of Twitter feeds and posts for financial performance forecasting. Mai et al. [6]conducted a predictive study of the predictive relationship between social media and Bitcoin by considering different social media platforms such as Internet forums, microblogging, etc. They used vector regression and vector error correction models. Garcia et al. [7] demonstrate a strong dependency between social signals about Bitcoin and price trends, manifesting as social word-of-mouth feedback loops and user-driven adoption loops. Kwon, Do-Hyung, et al. [8] applied a long short-term memory (LSTM) model to classify cryptocurrency price time series. Cryptocurrency price values are collected every 10 minutes. The data collected has five characteristics: the opening price, closing price, high price, low price, and the volume of transactions per time period (i.e., every 10 minutes).In the domain of Finance, Baker and Wurger [9]created an investor sentiment index to examine as an influencing factor, which was generated statistically from variables known to influence the behavior of established investors in the stock market and showed that investor sentiment significantly explained expected stock returns. Sihao Jiang [10]analyzes the relationship between cryptocurrencies and the KOL of Elon Musk, which potentially gave cryptocurrencies a lot of media coverage when Elon Musk focused excessively on Bitcoin, concluding that celebrities have the power to mobilize the public for cryptocurrencies. Znort987's Github blockchain parser [11]contains code to parse Bitcoin binary blockchain data to extract transaction, input, and output addresses for each block, and these have been used in several research efforts to parse blockchain data.

Chih-Hung Wu et al. [12]use a new forecasting framework of a novel hybrid LSTM model, combining a traditional LSTM model and an LSTM with AR(2) model, to perform forecasting of Bitcoin daily price data. The ACF and PACF graphical features of the Bitcoin price are used to derive the price lag and moving average periods, as well as the trading volume as predictor variables, which are trained and predicted in this hybrid model. MSE, RMSE, MAE, and MAPE evaluation metrics were calculated for the final prediction results.

Mohammad Ali et al. [13]applied two financial indicators called the Simple Moving Average (SMA) and the Exponential Moving Average (EMA) to train a simple linear regression forecasting algorithm to predict the 7-day trend of the Bitcoin price. The final error was calculated using the percentage error method, giving a percentage error rate of 3.03% for the linear regression prediction model, which implies a prediction accuracy of 96.97%.

Mahir Iqbal et al. [14]applied three machine learning algorithms, ARIMA, FBProphet, and XG Boosting, for time series prediction of bitcoin prices in the cryptocurrency market. The parameters evaluated for these models were: root means square error (RMSE), mean absolute error (MAE), and R2. ARIMAX was finally found to have an RMSE of 322.4, and it is the best algorithm for predicting bitcoin price changes in the market. In comparison, the FBProp and XGBoost algorithms obtained RMSE scores of 229.5 and 369, respectively, which is much lower than the ARIMAX algorithm.

Aggarwal [15]conducted a comparative study of various parameters that may affect the bitcoin price prediction results using models such as convolutional neural networks, long and short-term memory networks, and gated recurrent units, using RMSE as the evaluation criterion and also explored the effect of gold on bitcoin price, and the experimental results showed that LSTM has the smallest root mean square error and concluded that there is no positive relationship between gold and The experimental results show that the LSTM has the lowest root mean square error and concludes that there is no positive correlation between gold and bitcoin.

C. STRUCTURE

This paper is based on a machine learning approach to predict the price of Bitcoin, which is characterized by high volatility, many influencing factors, and high transaction frequency. Therefore, the selection of the feature values and the selection of the model for bitcoin prediction has a significant impact on the prediction results. Based on previous research and the consequences of their studies, we determined the eigenvalues and model selection for this project. In this project, we selected five data sources: the average daily bitcoin price series, Twitter sentiment, Google searches, daily active addresses, and top 100 addresses' holdings as the feature values that affect the bitcoin price. We also use Random Forest and Gradient Boosting as specific prediction models. Finally, problems in the project are given, as well as future research directions.

The main structure of the paper is as follows.

Section 2: Firstly, a background introduction to the research is given, as well as the importance of the study. And the relevant literature review is given, and the innovation points of the project are presented based on the literature review.

Section 3: Mainly describes the methodology of this project. It mainly includes the project objectives of the project and the relevant methods used in the project. These include two methods for data analysis, Sliding Window and Sentiment Analysis, and two predictive models, Random Forest and Gradient Boosting.

Section 4: The experimental case study, which is the central part of the project. In this section, the main work of the project is described in detail, and the performance results of the models are given.

Section 5: Future research directions. This section gives the expected future research directions based on the shortcomings of the project.

Section 6: Summary. Summarises the main points of the paper and the findings of this project.

II. METHODOLOGY

A. TASK DEFINITION

In order to predict the price trend of Bitcoin, this project indicates the price of Bitcoin based on the historical daily average price of Bitcoin. We will show the bitcoin price by using four regression models: linear regression model, Lasso, Ridge, and Random forest. And the average daily cost, daily active addresses, Google searches, Twitter sentiment, and top 100 holdings are used as the characteristic values that affect the everyday bitcoin price. We set our features as the following:

{ avg_price , active_addresses, google_trends, top100_coin_percent, avg_polarity }

(note: the above features are all one day data)

In this experiment, all models will be based on the time series.

A time series is a chronological sequence of values of indicators of a phenomenon, with the aim of making predictions about the future based on existing historical data. The value in this project is a typical time series with both temporal and numerical elements. Using sliding windows to transform time series data into a regression task in supervised learning. The regression task is to fit a sample of existing data points and then predict future values based on the fitted function. In this project, the features from the first five days are used to analyze the daily average bitcoin price on the sixth day, with the following relationship equation.

[FIGURE 1 about here.]

B. RELATED METHODS

This subsection is used to introduce the data processing methods and predictive models used in the project. The data processing methods include mainly sliding windows and sentiment analysis, and the predictive models used include random forest and gradient boosting.

1) Sliding window

The sliding window method is a method for reconstructing time series datasets as supervised learning problems. In statistics and time series analysis, it is also known as the lag or hysteresis method. This method reconstructs a time series dataset using the previous time step as the input variable and the next time step as the output variable.

2) Sentiment analysis

a:

Definition: Sentiment analysis is often thought of as a classification task that classifies a piece of text into one of several categories of sentiment based on the feelings or emotions expressed by the masses in the text. Sentiment analysis focuses on the polarity of a text (positive, negative, neutral), but it also goes beyond polarity to detect specific feelings and emotions (anger, happiness, sadness, etc.), urgency (urgent, not urgent), and even intention (interested, not interested).

Polarity: TextBlob returns the polarity and subjectivity of the sentence. Polarity lies between [1], with -1 defining negative sentiment and one defining positive sentiment. Negation reverses polarity. textBlob has semantic tags that help with fine-grained analysis. For example - emoticons, exclamation marks, emojis, etc. Subjectivity lies between [0,1]. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. A high subjectivity means that the text includes personal views rather than factual information.

3) Random forest

[FIGURE 2 about here.]

Random Forest is a supervised machine-learning algorithm that is one of the most commonly used today. This algorithm has non-linear properties and can be used for both classification and regression tasks, making it adaptable to a variety of application scenarios. Random forests reduce the number of variances by averaging multiple deep decision trees. The decision trees are trained on different parts of a dataset, and although there is some loss of interpretability and a slight increase in bias, the performance is still substantially better in the final model. Random forests offer the advantages of easy measurement of relative importance, versatility, no overfitting, high accuracy, reduced time spent on data management, and fast training speed.

The generic bagging method is used for tree learning by the random forest training algorithm. The bagging procedure is repeated (B times) with put-back samples from the training set, and the tree model is then trained on these samples, given a training set X = x 1,..., x n and a target Y = y 1,..., y n.

For b = 1, ..., B:

1. Sample n training examples from X and Y with replacement; refer to these as X_b and Y_b.

2. Train on X_b, and Y_b to get a regression tree f_b: After training, the prediction of an unknown sample x can be achieved by averaging the predictions of all individual regression trees on x, with the following relationship equation.

[FIGURE 3 about here.]

4) Gradient boosting

Gradient Boosting is a method of Boosting in which the main idea is that each time a model is built, it is in the direction of the gradient descent of the previously made model loss function. The loss function is used to evaluate the performance of the model and is usually positively correlated with performance. In GBDT (gradient boosting decision tree) each tree learns the conclusions and residuals from all the previous trees, and the residuals are a cumulative sum of the actual values plus the predicted values, and the final result is a joint decision of the decision trees.

Its advantages are mainly in the combination of features and the discovery of essential elements, and the ability to generalize. Feature combination: The original parts are combined to create higher-order features or non-linear mapping after being transformed into high-dimensional sparse features by GBDT. These newly created features are then fitted once again as the input of FM (Factorization Machine) or LR (Logistic Regression). Finding significant features: Since the growth process of decision trees involves the continuous selection and segmentation of features, the inherent advantage of GBDT, which consists of many decision trees, is that the importance ranking of features can be quickly and easily obtained and is highly interpretative. Bias and variance are the two components of the generalization error. 1) Boosting is used to ensure minimal bias, where each step involves fitting the original data more closely to the preceding round. 2) Simple models, such as decision trees with relatively shallow depth, are used to ensure a low variance. Combining the two permits the development of integrated models with strong generalization potential based on learners with relatively subpar generalization abilities.

III. EXPERIMENT CASE STUDY

A. DATASET

1) Feature engineering

In this section, the selection and exclusion of characteristic values are presented. The distinctive value was selected based on the previous study. Finally, the unique value difficulty value and hash rate are excluded, and Twitter sentiment, Google search rate, Bitcoin daily average price, top 100 addresses holding coins, and daily active address volume are selected as relevant eigenvalues.

a: Why not choose difficulty and hashrate

The relationship between difficulty and hashrate: The relationship between the difficulty of the current block and the target value of the current block is difficulty = difficulty_1_targetcurrent_target. Where difficulty_1_target is a vast constant, which indicates the maximum difficulty value allowed for mining pools, with a difficulty value close to 2^(256-32) at the maximum, this formula shows that the relationship between the difficulty value and the target value is inversely proportional. The smaller the target value, the greater the difficulty of generating blocks.

Network hashrate is the number of calculations to find a random number that makes the block hash value lower than the target value. It is used to estimate how many operations each node can do per second during the mining process of the miner. The unit is hash/s. When the computing power is higher, more computers are involved in mining. The relationship between the current target value and the current difficulty value is currentTarget = difficulty_1_target difficulty [1]. Usually, the highest possible target is defined as 0x1d00ffff. When the target is at the maximum allowed value, the minimum difficulty is 1. When the difficulty is 1, the offset is 0xffff * 2**208. So the number of times the hash needs to be calculated to find a block with difficulty D is D22560xffff2208=D232. The current requirement for finding the corresponding nonce value is to find it within ten minutes, that is, to complete the calculation within 600 seconds, which means that the network hashrate must be D232600 at least. Blocks in Bitcoin are set to expect a new block to be generated every 10 minutes on average. To maintain the block generation rate, the block difficulty needs to be dynamically adjusted. Every 2016 block generated will change the target value for the following 2016 blocks, which means that the block generation difficulty changes about every two weeks. The relationship between the network hash rate and the target value is 600R=2256T+1. The formula shows that the network hash rate is inversely proportional to the block target value, and the lower the target value, the higher the network hash rate.

In summary, the network hash rate is inversely proportional to the target value, and the target value is inversely proportional to the mining difficulty value, so the network hash rate is positively proportional to the difficulty value. So the network hash rate and the difficulty value need to choose only one if they are to be used as feature values, but if they cannot be selected, neither can be used as feature values.

Multiple factors influence hashrate: Since the relationship between difficulty and hashrate is linear, when hashrate is influenced by multiple factors, it means that problem is also affected. Within the Bitcoin industry, there is a claim that changes in computing power have some reference value to Bitcoin's market trends. However, by analyzing the multiple factors that lead to changes in computing power, it was found that the coin price is not closely linked to the increase in computing power. In addition, as many factors influence computing power, it is not suitable to be considered an independent variable affecting the price of Bitcoin.

Mining machine prices: To mine, miners need to maintain machines with high computing power. The second half of 2018 saw a significant drop in the price of bitcoin mining machines, with even some mainstream miners, such as Avalon and Ant miners, falling by more than 50%. On used mining websites such as cybtc.com and huobi.com, the used price of the Ant Mining S9 is currently around $1,200-1,500, while the price of a new mining machine on the official website is $3,000. This shows that despite the falling price of bitcoin, many miners are still willing to "take the plunge" and stock up as the cost of mining machines decreases, and the market for mining machines is not shrinking.

Mining machine development: With the rapid evolution of mining machine technology, the major manufacturers are competing to launch high-capacity, low-power mining products. At the end of July 2018, Avalon found the A921 mining machine with a 7nm chip. In August, Ant Mining released the S9 Hydro water-cooled mining machine, and in September, Sleipnir Mining followed suit with the M10

mining machine with a 16nm chip. This indicates that the competition among mining machine manufacturers has entered a feverish phase. The mining industry has seen a rise in shipments as the rate of mining machines updates, while a large number of new mining machines are supporting the vast computing power of the mining machines.

Bitcoin Reliance: Market conditions research is significant for cryptocurrencies. Many experts pointed to a cryptocurrency bubble in 2017 when the price of cryptocurrencies rose by 900%. (åijŢçŤÍhttps://jfin-swufe.springeropen.com/articles/10.1186/s40854-021-00321-6)However, the drop and extremes in cryptocurrency trading caused market panic as mainstream coins such as ETH and ETC plummeted massively in 2018. In this situation, Bitcoin has undoubted stability compared to other cryptocurrencies. This has made miners more inclined to bitcoin mining. The return of miners in droves has also led to a rise in Bitcoin's computing power.

Because various factors influence bitcoin's computing power, there is no apparent relationship between it and its price. It is not considered an independent variable for predicting bitcoin's price trend. Firstly, it is clear from the data that bitcoin's net-wide computing power has been climbing higher this year, but bitcoin's market cap has shrunk significantly. Hence, it is clear that the currency's price is not closely linked to the growth in computing power, at least not in a positive way. Secondly, based on Bitcoin's difficulty adjustment settings, the amount of Bitcoin mined over a fixed period is mainly stable, and an increase in computing power will have no impact on the amount of Bitcoin produced and, therefore, no impact on Bitcoin's price recovery.

In short, computing power does not affect the price of the coin, but the cost of the currency affects computing power. Because when the price of the coin rises, miners' profits grow, and they buy more mining machines. The higher the computing power of new mining machines, the more there are, leading to an overall increase in bitcoin computing power.

2) What features are selected

a:

Twitter sentiment & Google search rate: To study the reactions of individuals and groups to particular topics is to conduct sentiment analysis. Twitter is a relatively new tool for measuring social research. We are talking about millions of people voluntarily expressing their opinions on any topic. It is an organic data source for data collection and research. Sociological, political, economic, and analytical analyses are currently being carried out, especially for science and business. Twitter is probably the best place to get a sample of public opinion. Olivier Kraaijeveld et al. [16]implemented a sentiment analysis method based on a specific dictionary of cryptocurrencies and, in the final results, found that Twitter sentiment could be used to predict price returns for Bitcoin, Bitcoin Cash, and Litecoin.

In today's increasingly digital economy, people's preferences, tastes, or consumption habits are leaving a digital footprint on media platforms. Google search has been found to be used to predict economic indicators. As cryptocurrencies are digital native assets, investors gather market information mainly through online platforms such as community media, exchange forums, etc. Therefore, Google search has become one of the main options for investors. Urquhart [17]was one of the first papers to link market interest in cryptocurrencies to Google trends, finding that realized volatility, trading volume, and returns influence future searches for the term "bitcoin."

Twitter sentiment and Google searches have a strong dependency on Bitcoin trends and are perfect indicators of investor behavior and valuable tool for developing effective trading strategies. Therefore the choice to combine the two in our model allows us to more accurately capture the sentiment trends of investments towards bitcoin and, in turn, examine how this affects the trend direction of bitcoin prices.

In addition, there are different emotions to bitcoin per day. Since we want to predict bitcoin's average price, we need to calculate the mean value of the tweet's polarity. We choose average, not median due to the following reasons: There are almost no outliers in our data; The standard is the estimate of the population mean, while the median is the estimate of the population center; Our data distribution is almost symmetric, so avg can better reflect the population. Here is an example of the polarity of the tweet on 2022.9.15. We can see that the high frequency happens on 0, which is almost symmetric.

Daily average price of bitcoin: Most Bitcoin price trend analysis models use an opening price, a closing price, a high price, and a low price. The daily opening price is the price at which a contract is traded within the first five minutes of the market opening. If no such price is generated, the opening price of the day will be the first sold price of the day. If the contract is not traded throughout the day, yesterday's settlement price will be used as the opening price of the day. The closing price is the last sold price of the contract for the day. If there are no transactions for the agreement throughout the day, the opening price will be the closing price for the day. The high and low prices are also the highest or lowest point of the day in the amount of price produced. These four prices are all fixed prices, the amount generated at a specific time or point, and do not have the property of neutralizing the price fluctuations of the day. The daily average price, on the other hand, reflects the general situation of the day's prices and has an intuitive, concise character. Among the prices that fluctuate throughout the day, the daily average price reflects the relative trend of the data and the average level of the data for the day.

The holding volume of the top 100 address: When the price of Bitcoin is stable at a specific range the supply relationship is balanced. If the holdings of the top one hundred holders of bitcoin were to continue to increase,

individual retail investors would be driven by psychological and emotional biases and would potentially invest based on fundamentals. In this way, their options would unbalance the price of bitcoin. Therefore, the holdings of the top 100 holders of bitcoin have an impact on the bitcoin price float. With this direction in mind, we chose it as the eigenvalue for model testing.

The active address number: Studies have shown that the number of active addresses is the most critical variable affecting the price movements of Bitcoin and Ether. According to Metcalfe's Law, the value of a network is equal to the square of its number of users. The number of users on the Bitcoin and Ether networks is expressed by the number of active addresses (Metcalfe's law) is a law about the value of networks and the development of network technology, proposed by George Gilder in 1993, but named after computer networking pioneer and 3Com founder Robert Metcalfe in recognition of his contributions to the ethereal contribution to the network. It reads that the value of a network is equal to the square of the number of nodes within that network and that the value of that network is proportional to the square of the number of users connected to it. The law states that the greater the number of users of a network, the greater the value of the entire network and each computer within that network.

3) Data collection

This paper selects data from 2017.1.1 to 2022.11.10 as the total sample, with total 2,140 rows, each including five features: Bitcoin daily average price series, Twitter sentiment, Google search, daily active address volume, and coin holdings of top100 addresses. A large amount of data helps to provide a more comprehensive analysis of the characteristics of bitcoin price movements. Other data use the Optimized Selenium Chrome driver to crawl data. The following values are derived from the model in this report.

[FIGURE 4 about here.]

[FIGURE 5 about here.]

[FIGURE 6 about here.]

[FIGURE 7 about here.]

[FIGURE 8 about here.]

B. EXPERIMENT PROCEDURE

1) Data pre-processing

a:

Interpolate: Some data is missing in the active_address column and the top100_coins_percent column. If the rows missing these two features are removed, then the time series is not homogeneous, and the sliding window will keep passing effects in the following predictions. To fill in the missing values, we use the time interpolation method from the panda's library. Since the indices of our data are one day apart and uniformly distributed, our time interpolation is equivalent to linear interpolation.

IV. SLIDING WINDOWS

[FIGURE 9 about here.]

[FIGURE 10 about here.]

The graph shows that we use data 1-5 (window1) to predict the 6th average price, and then use data 2-6 (window2), the 6th data being the previous prediction, to indicate the 7th average price.

We set the window size to 5. By experiment, we demonstrate that the RMSE and accuracy are better when the window size is five than when the window size is 7 and 30. We did not conduct experiments with window sizes larger than 30. This is because we thought it was meaningless to predict quarterly trends for variables like bitcoin prices that can produce changes at any time.

V. TRAIN-TEST SET SPLIT AND NORMALIZATION

We first use 80% of the data as the training set and 20% of the data as the test set. Then, we normalized our train set and tested using Z-score normalization. Z-score normalization performs the normalization process for each value in the data by the following formula so that the mean of all values is 0 and the standard deviation is 1. This normalization method allows de-scaling and eliminates the influence of magnitudes between features. Unifying all elements into a roughly equal value interval allows metrics of different volumes to be compared and weighted. The normalization process can be described as the following steps: create a standard scaler, fit the scaler with the data from the training set, and then use the fitted scaler to transform the data from the training and test sets. It is important to note that we should only use the data from the training set to fit the scaler. Once we use the data from the test set to match the scaler, information about the test set data will be leaked. This will lead to inaccurate prediction results of our model.

A.

1) Modeling

For the modeling part, we used sklearn to implement a random forest regressor, gradient boosting regressor, and a voting regressor model using both of the above as the estimator. To tune the parameters, we used 5-fold cross-validated grid search method to find the optimal hyperparameters. The final hyperparameters of the gradient boosting regressor are listed below: n_estimators=100,min_samples_split=300,max_depth=7,min_sample

Similarly, the optimal hyperparameters of the random forest are n_estimators=200 and max_depth=8. To get the same result every time, we set the random seed to 5. Voting regressors are suitable for combining well-performing models to balance the weaknesses of each. The two-based voting

regressor fits both regressors over the entire data set. It then averages the individual predicted values to form a final expected value. Voting regressors are suitable for combining well-performing models to balance the weaknesses of each. Since both random forest and gradient boosting perform well, we fit both regressors on the entire dataset. The voting regressor then averages the two predictors to form a final prediction.

B. PERFORMANCE EVALUATION

1) Metrics

We did not choose R-square as the performance metric for the following reasons: First, R-square measures how good the fit is, not the extent to which the correct model affects or predicts the wrong model. Second, our task is to predict the price of Bitcoin, so there is no need to use it as a measure of the model. Instead of R-square, we choose RMSE and accuracy. However, RMSE reflects the maximum deviation of the error, and accurately reflects the actual variation of the error.

a:

RMSE: : Root Mean Squared Error (RMSE) is the squared difference between the current predicted value and the actual value. The arithmetic square root of the mean value, which gives relatively high weight to more significant errors, is suitable for undesired large error scenarios. The gradient is simple to calculate and is often used in the prediction process of regression tasks. The formula for the mean squared error is shown in the figure below.

[FIGURE 11 about here.]

Accuracy: : accuracy is the rate of the absolute difference between the actual mean price and the predicted one divided by the actual average price. The scientific definition of accuracy refers to the degree to which the average value of multiple measurements under certain experimental conditions is consistent with the actual value, expressed as an error. It is used to represent the size of the systematic error.

The accuracy rate intuitively reflects how good our prediction is. It uses the absolute value to calculate the percentage of the gap between the actual and predicted value. The formula for the mean squared error is shown in the figure below.

[FIGURE 12 about here.]

[FIGURE 13 about here.]

[FIGURE 14 about here.]

VI. FUTURE RESEARCH DIRECTIONS

Use the state-of-the-art model in NLP like Bertand Transforme to do sentiment analysis

Investigating how to predict bitcoin prices using low-shot deep learning methods Combine graph neural networks and self-attention mechanism for prediction to provide good results and interpretability

Combine graph neural networks and self-attention mechanism for prediction to provide good results and interpretability

VII. CONCLUSION

Cryptocurrencies have received a lot of attention as an emerging thing in recent years. In this paper, we show three different machine learning models to predict the average price of bitcoin.

In terms of data selection, the most representative cryptocurrency for Bitcoin is chosen as the subject of this study. In predicting the average price, the cryptocurrency's price data is affected by historical data, so the prediction is made by using a sliding window. We use the relevant data from the previous 5 days to predict the average price of the following day. For the characteristic values, we selected five deals: the average daily bitcoin price series, Twitter sentiment, Google searches, daily active addresses, and top100 addresses' holdings.

After processing the characteristics and analyzing the model, finally, the best results were achieved in the random forest, Gradient Boosting, and voting models with a prediction accuracy of 95.73%.

REFERENCES

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Decentralized Business Review, pp. 21 260–21 260, 2008.
[2] M. Briere, K. Oosterlinck, and A. Szafarz, "Virtual currency, tangible return: Portfolio diversification with bitcoin," Journal of Asset Management, vol. 16, no. 6, pp. 365–373, 2015.
[3] R. Adcock and N. Gradojevic, "Non-fundamental, non-parametric Bitcoin forecasting," Physica A: Statistical Mechanics and its Applications, vol. 531, pp. 121 727–121 727, 2019.
[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), pp. 339–343, 2018.
[5] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," Journal of computational science, vol. 2, no. 1, pp. 1–8, 2011.
[6] F. Mai, Q. Bai, J. Shan, X. S. Wang, and R. H. Chiang, 2015.
[7] D. Garcia, C. J. Tessone, P. Mavrodiev, and N. Perony, "The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy," Journal of the Royal Society Interface, vol. 11, no. 99, pp. 20 140 623–20 140 623, 2014.
[8] D.-H Kwon, J.-B Kim, J.-S Heo, C.-M Kim, and Y.-H Han, "Time series classification of cryptocurrency price trend based on a recurrent LSTM neural network," Journal of Information Processing Systems, vol. 15, no. 3, pp. 694–706, 2019.
[9] M. Baker and J. Wurgler, "Investor sentiment and the cross-section of stock returns," The journal of Finance, vol. 61, no. 4, pp. 1645–1680, 2006.
[10] S. Jiang, "The Relationship between the Cryptocurrency and the KOL of Elon Musk," 2022 13th International Conference on E-Education, E-Business, E-Management, and E-Learning (IC4E), pp. 454–458, 2022.
[11] Znort987. [Online]. Available: https://github.com/znort987/blockparser. (accessed
[12] C. H. Wu, C. C. Lu, Y. F. Ma, and R. S. Lu, "A New Forecasting Framework for Bitcoin Price with LSTM," 2018 IEEE International Conference on Data Mining Workshops, pp. 168–175, 2018.
[13] 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), pp. 330–335, 2020.
[14] M. Iqbal, M. Iqbal, F. Jaskani, K. Iqbal, and A. Hassan, "Time-series prediction of cryptocurrency market using machine learning techniques," EAI Endorsed Transactions on Creative Technologies, vol. 8, no. 28, 2021.
[15] 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), pp. 1–5, 2019.
[16] O. Kraaijeveld and J. De Smedt, "The predictive power of public Twitter sentiment for forecasting cryptocurrency prices," Journal of International Financial Markets, Institutions and Money, vol. 65, pp. 101 188–101 188.
[17] A. Urquhart, "What causes the attention of Bitcoin?" Economics Letters, vol. 166, pp. 40–44, 2018.

List of Figures

$$f(x_t, x_{t+1}, x_{t+2}, x_{t+3}, x_{t+4}) = y_{t+5}$$

FIGURE 1. Target formula

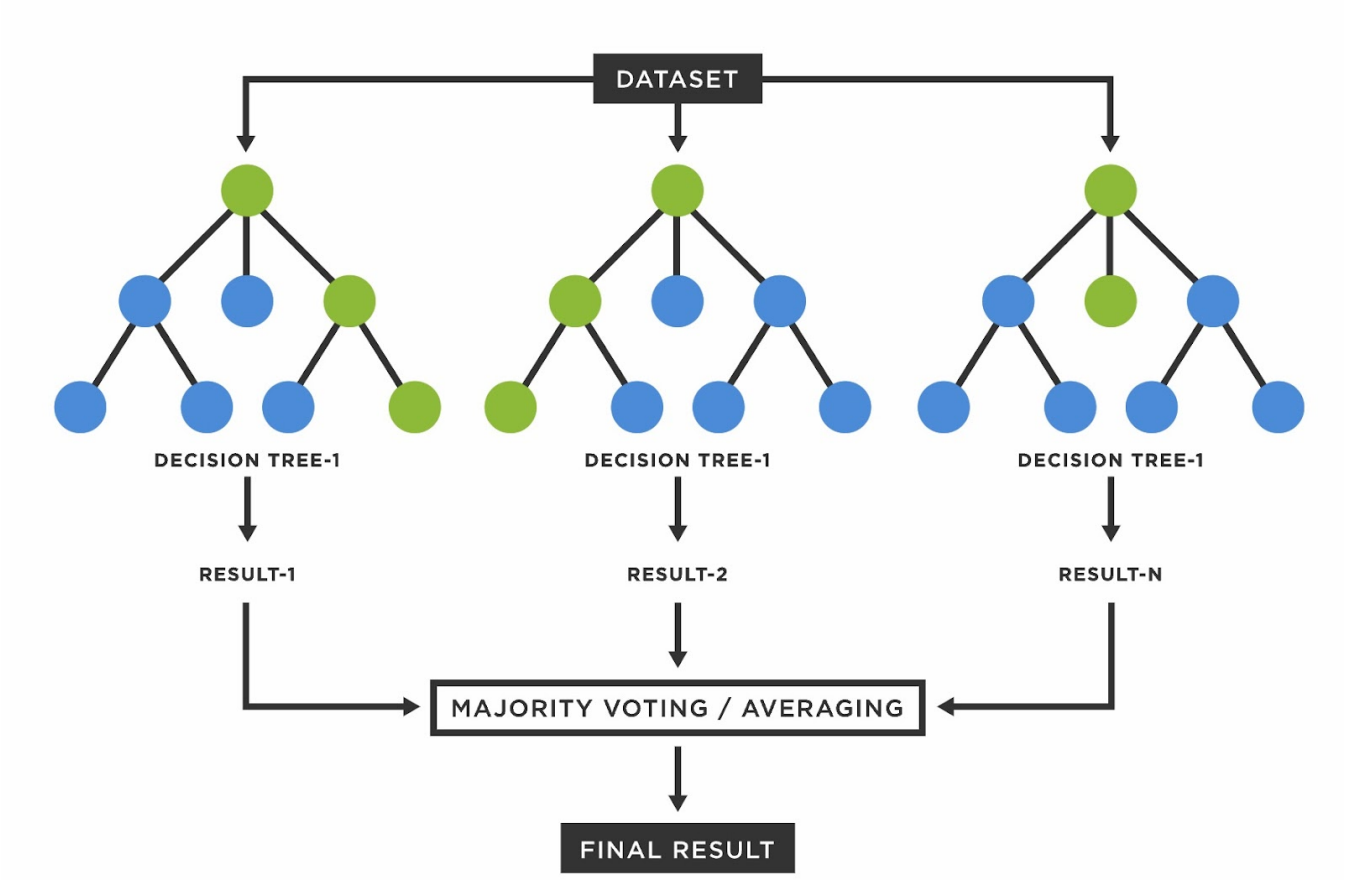


FIGURE 2. Random forest organism

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

FIGURE 3. Relationship equation

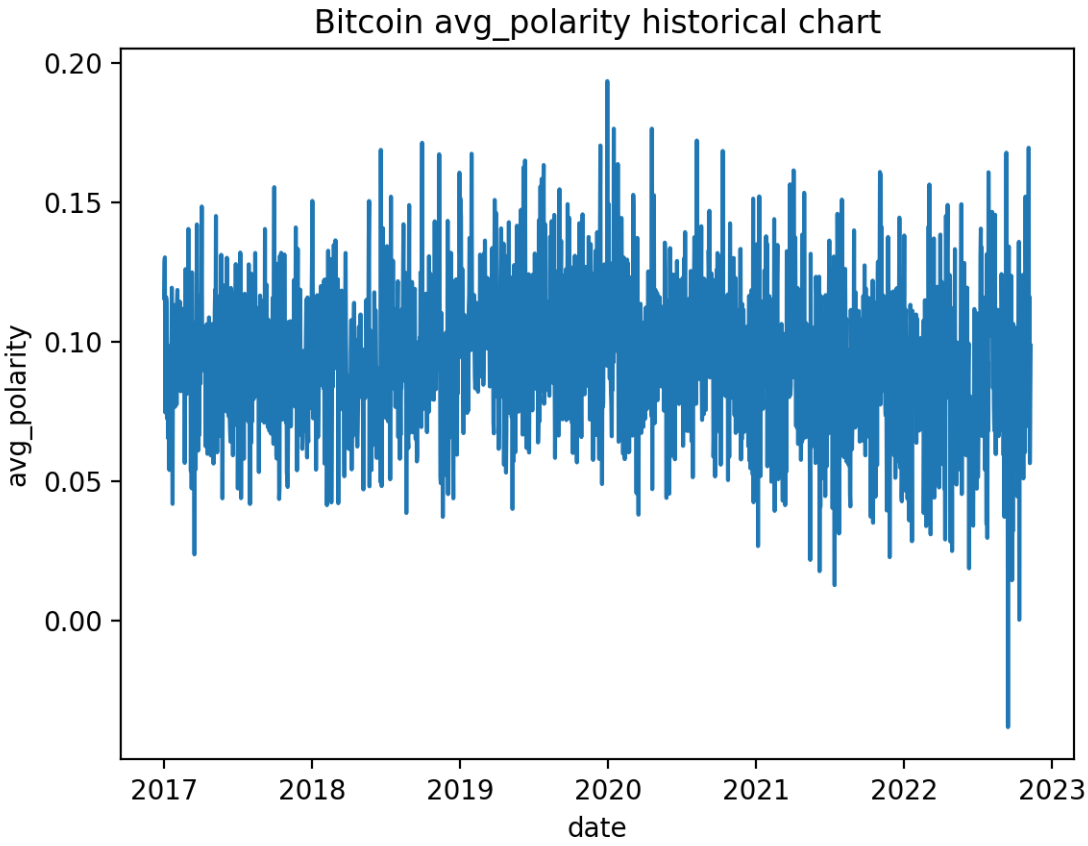


FIGURE 4. Time-avg_polarity historical chart

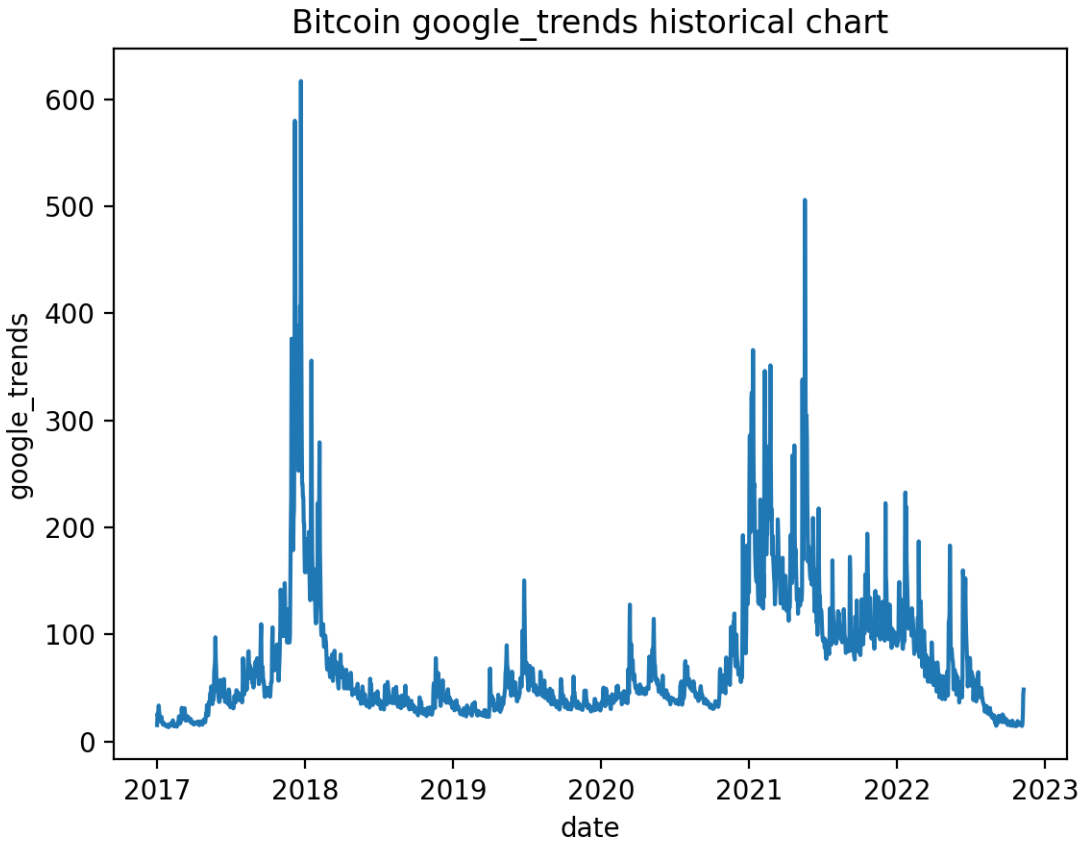


FIGURE 5. Time-google_polarity historical chart

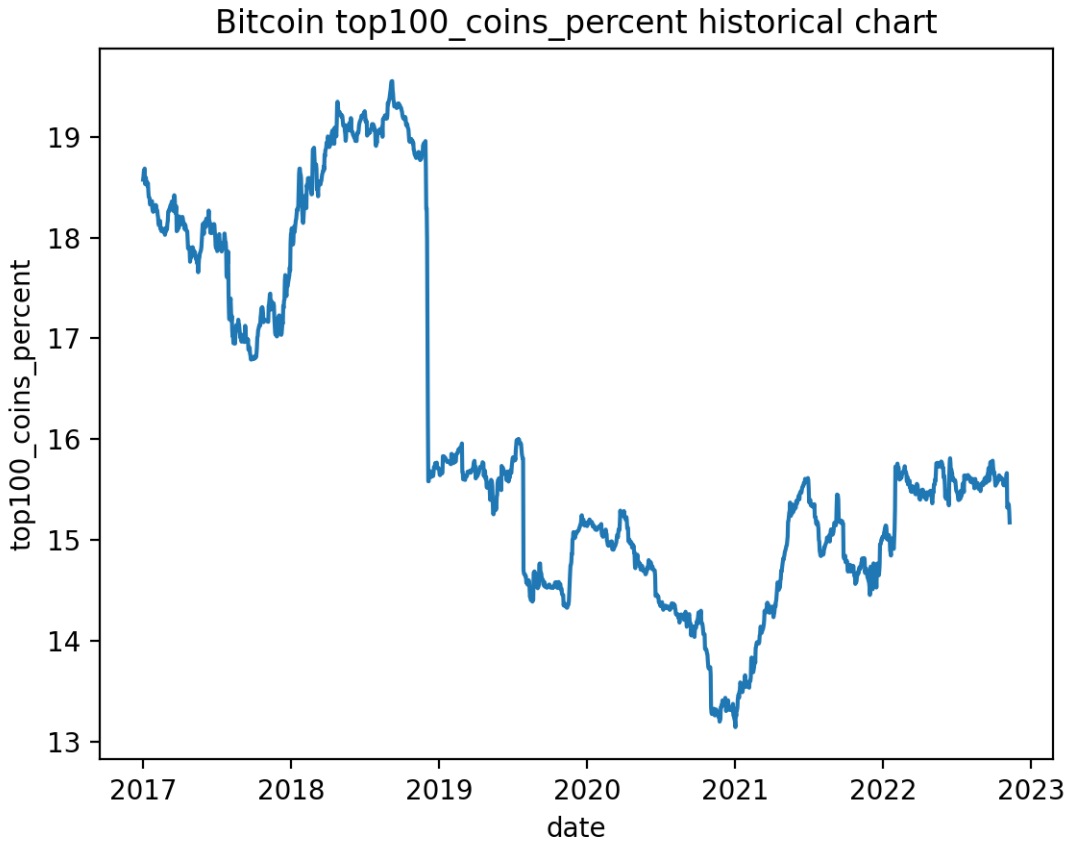


FIGURE 6. Time-top100_coins_percent historical chart

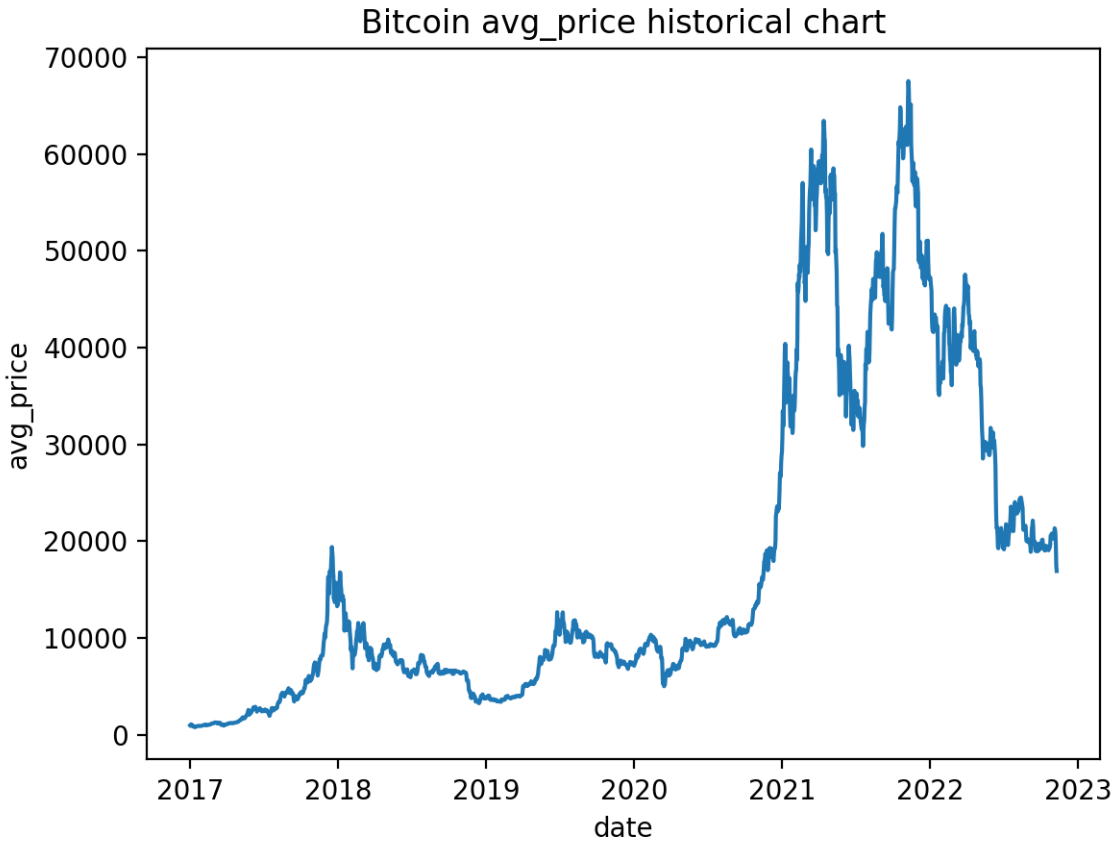


FIGURE 7. Time-avg_price historical chart

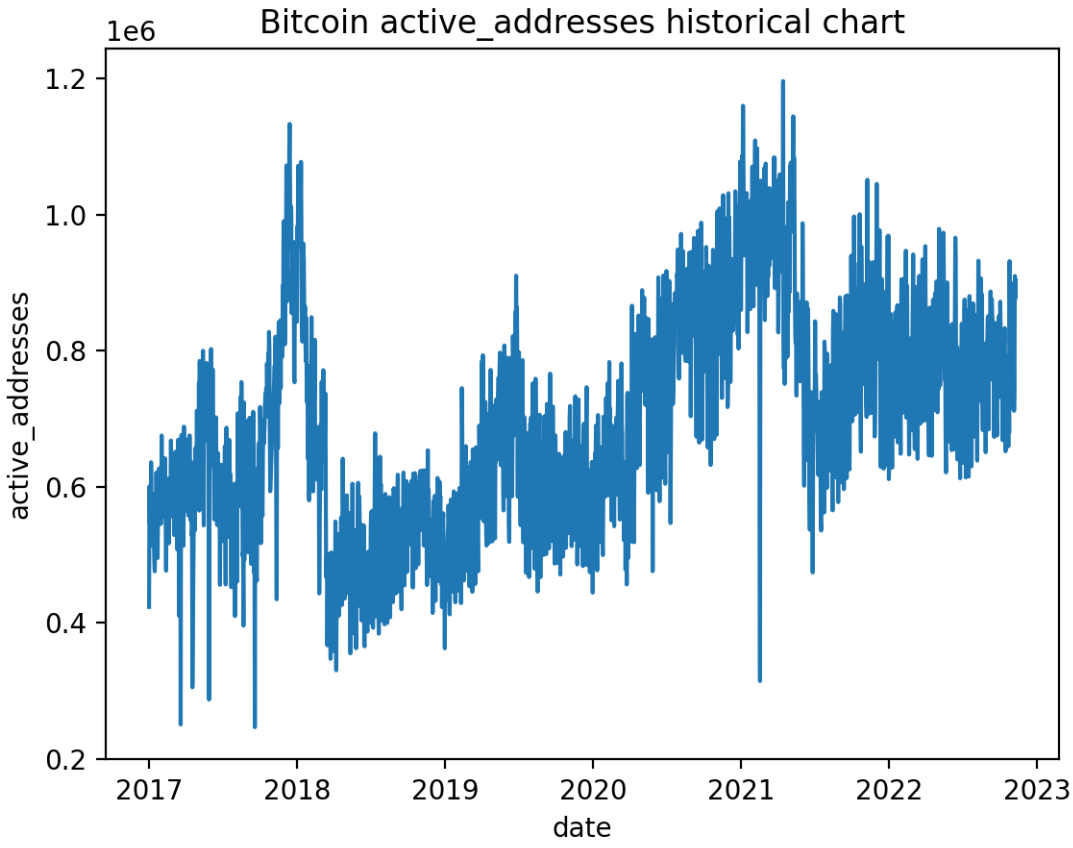
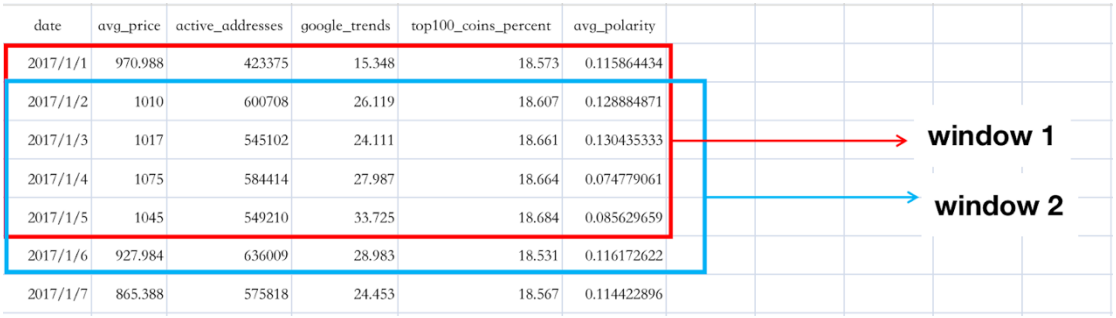


FIGURE 8. Time-active_addresses historical chart

Comparing of Window Size

Window Size	Model	Accuracy	RMSE
5	Random Forest	95.43%	1794.22
	Gradient Boosting Regression	95.51%	2245.84
	Voting Model	95.73%	1910.49
7	Random Forest	95.13%	1887.03
	Gradient Boosting Regression	94.70%	2459.72
	Voting Model	95.15%	2079.71
30	Random Forest	90.98%	2919.00
	Gradient Boosting Regression	89.01%	4006.69
	Voting Model	90.08%	3392.40

FIGURE 9. Comparison of window size



date	avg_price	active_addresses	google_trends	top100_coins_percent	avg_polarity
2017/1/1	970.988	423375	15.348	18.573	0.115864434
2017/1/2	1010	600708	26.119	18.607	0.128884871
2017/1/3	1017	545102	24.111	18.661	0.130435333
2017/1/4	1075	584414	27.987	18.664	0.074779061
2017/1/5	1045	549210	33.725	18.684	0.085629659
2017/1/6	927.984	636009	28.983	18.531	0.116172622
2017/1/7	865.388	575818	24.453	18.567	0.114422896

FIGURE 10. An example of sliding window

$$RMSE(X, f) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2}$$

FIGURE 11. RMSE

$$\text{accuracy} = \frac{|\text{ actual average price } - \text{ predicted average price }|}{\text{actual average price}} * 100$$

FIGURE 12. Accuracy

Table of Model Evaluation (window size = 5)

Model	Accuracy	RMSE
Random Forest	95.43%	1794.22
Gradient Boosting Regression	95.51%	2245.84
Voting Model	95.73%	1910.49

FIGURE 13. Model evaluation

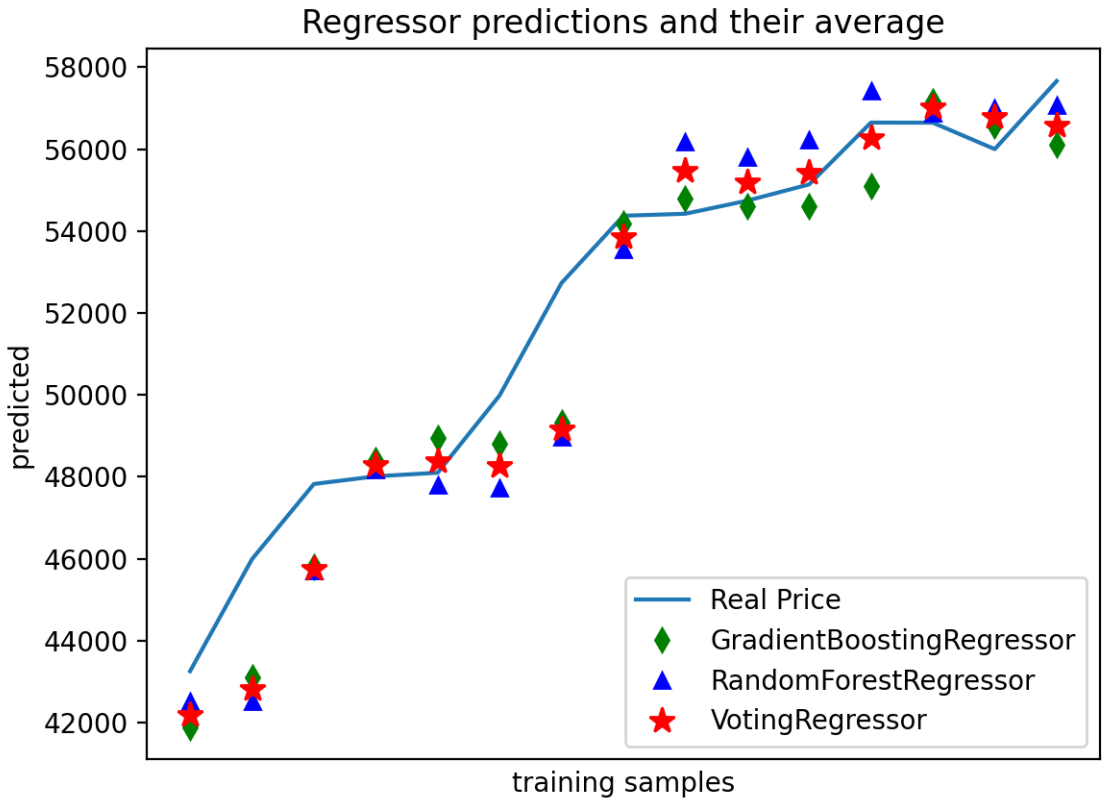


FIGURE 14. Model comparing