System for Analysis and Prediction of Trends in Cryptocurrency Market

Shaad Iqbal Ansari[1], Vani H Y[2],*

[1]Department of Information Science and Engg JSSSTU, Mysore, India
[2]Assistant Professor Department of Information Science and Engg JSSST University Mysuru, India

Abstract

In this article forecasting of daily closing price series of Bitcoin, Ripple, Dash, Litecoin and Ethereum crypto currencies, using data on prices (open, low, high), market capital and volumes using prior days is focused. The value conduct of cryptographic forms of money remains to a great extent neglected, giving new chances to scientists and business analysts to feature the likenesses and contrasts with standard monetary costs. Hence the paper is focused on this area. he results are compared with various benchmarks. Predictions are done using statistical techniques and machine learning algorithms. A simple linear regression (SLR) model that uses only a single-variable sequence of closing prices for forecasting, and a multiple linear regression (MLR) model that uses a multivariate sequence of prices and quantities at the same time. The simple linear regression (SLR) model for univariate serial forecasting uses only closing prices. Mean Absolute Percentage Error (MAPE) and relative Root Mean Square Error (relative RMSE) performance measures are considered. The accuracy achieved by the ARIMA model on our dataset is the highest, followed by Multivariable Linear Regression and LSTM.

---

* Corresponding author. Email: vanihy@sjce.ac.in

1. Introduction

Bit coin is a decentralized digital currency, without a central bank or single administrator that can be sent from user to user on the peer-to-peer bitcoin network without the need for intermediaries. Such digital currency allows decentralized peer- to-peer network or online transactions carried out by 'miners' within a network trading. As a result, there is no centralized authority or any third-party financial institutions that has control of the Bit coin network. All transactions on the Bitcoin network are embedded in blocks to the open ledger that is known as the block chain to be verified by the miners using cryptographic proof-of-work. Therefore, Bitcoin is an entry in currency markets, though it is officially considered as a commodity rather than a currency, and the value conduct of it remains to a great extent neglected, giving new chances to scientists and business analysts to feature the likenesses and contrasts with standard monetary costs. Regardless of whether the quantity of

Bitcoin cost determining contemplates is expanding, it actually stays restricted. In this work, the figure of every day shutting value series of the Bitcoin cryptographic money utilizing information on costs and volumes of earlier days is examined. To predict currency rate different approaches like statistical techniques and machine learning algorithms are used on datasets. The dataset consists of the closing price, volume, market capital, open price, high price and low.

2. Literature Survey

Throughout the long-term numerous calculations have been created for gauging time series in financial exchanges. Over the years researchers have proposed many techniques for forecasting time series in stock markets. Most of the concept is adopted from [1]. A genetic and neural approach[2] is proposed to predict daily prices using technical analysis factor and daily prices. [3] Proposed a hybrid model using fuzzy

clustering with Fuzzy InferemceI System to predict an index level forecast. A combination of firefly and SVR is proposed by [4] to predict stock market prices. A very few papers focused on Bitcoin price prediction. Recently, many authors proposed [5,6,7] machine learning algorithms to predict Bitcoin price.[8] Authors used Fast Wavelet Transform to check the changes in Bitcoin prices. Univariate and multivariate models are proposed by [9] to predict Bitcoin, Litecoin,Ripple and Ethereum. These market conditions are shown in figure below (Fig.1: bull-market condition; Fig. 2: bear-market condition).
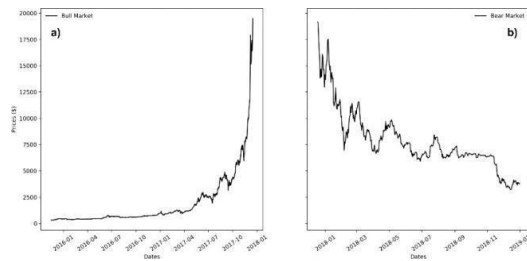
Figure 1. bull-market condition
Figure 2: bear-market condition

Our examination ranges over a time of over 4 years, described by different value elements. Accordingly, we had the option to prepare and test our models, remembering for each stage both bull-and bear-economic situations. Therefore, our investigation advances the best in class, as it is the most refreshed and manages the greatest and more complete dataset frame selection procedure.

3. Methods AND Implementations dataset

In this section notions on time series analysis which helps to take the operational decisions about the algorithms are discussed. Then the dataset, pre-processing analysis of the data is explored. Finally, proposed algorithms and the statistical tools adopted are discussed.
Time series analysis [10, 11, 12]: - Three systematic components are used to analyze the data in time series. They are (i) base level, (ii) trend and (iii).seasonality, non-systematic component called 'noise'.
Base level: The average value in the series.
Trend: It is an area showing an increasing or decreasing slope in the time series.
Seasonality: For a repeated pattern between regular intervals, due to seasonal factors.
Noise: Random variations in the series. Time series is a combination of these four components, where base level and noise always occur whereas trend and seasonality are optional. Depending on the nature of the trend and

seasonality, a time series can be described as an additive or multiplicative model.

Additive model:

y(t) = BaseLevel + Trend + Seasonality + Noise

Multiplicative model [13]: is also called classical decomposition and it is calculated using equation (2):

$$y(t) = BaseLevel*Trend*Seasonality*Noise \quad (2)$$

Statistical measures*: The statistical measures like mean (µ), the standard deviation (σ) and the trimmed mean (ū) obtained discarding a portion of data from both tails of the distribution. The trimmed mean gives an estimate of central tendency and is useful for time series with high volatility.

4. Dataset

The data set consists of 1320 types of Crypto-currencies with 650000 observations of 10 variables.
Ten variables are Date, Open, High, Low, Close, Volume, Market Cap, coin and Delta.

Date	: Date of observation
Open	: Opening price on the given day
High	: Highest price on the given day
Low	: Lowest price on the given day
Close	: Closing price on the given day
Coin	: Name of the crypto-currency
Volume	: Volume of transactions on the given day

Market Cap: Market capitalization in USD and calculated using
Market Cap = Price X Circulating Supply
Where
Circulating Supply : Number of coins available in market.
Data pre-processing:
The following steps are used to pre-process the data .
1.csv file from Kaggle with various crypto-currencies information
2. Dataset is split into different .csv file depending on coin type.
3.The data type of Date is changed date format from string
4. Market Cap and Volume data type changed to float from string
5. Null values are removed.
Then, forecasted future value of Y (Close Price) on the basis of the past values of X. For multiple
Linear regressions, X is chosen as Open, High, Low, and Market Cap and tries to predict the future close price. In LSTM and ARIMA models, have used previous Close Price of BTC as X to predict.

5. Techniques Used

Uni-variate and multivariate forecasting *[14]*
uni-variate forecast: - It predicts time series made by observations belonging to a single feature record. Here the closing price of the series is considered. In multivariate several features are used for prediction. Here 'Open', 'High', 'Low', 'Volume', 'Market Cap' with 'X' as independent variable and Y as dependent variable,

Statistical analysis: As a first step, statistical analysis is carried out using Dickey- Fuller test and autocorrelation in order to check for non- stationary in the time series and its mathematical definition is shown in Eq (3).

$$p_{value} = P(t_{observed} | H_0) \quad (3)$$

p-value: Lower the p- value, the higher the significance.

Linear regression [15]: is a linear approach for modeling the relationship between a dependent variable and one independent variable, represented by the equation is:

$$y = b_0 + \vec{b_1}.\vec{x_1} \quad (4)$$

y and $\vec{x_1}$: Dependent and the independent variable respectively

where

b_0: Intercept

$\vec{b_1}$: vector of slope coefficients

In our case the

$\vec{x_1}$: Values of the closing prices of the previous days.

y: closing price to be predicted.

Multiple Linear Regression model (MLR) [16].

Multiple linear regressions refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables and it is represented by the following equation.

$$y = b_0 + \vec{b_1}.\vec{x_1} + ... + \vec{b_n}.\vec{x_n} = b_0 + \sum_{i=1}^{n} \vec{b_i}.\vec{x_i} \quad (5)$$

where the index x_i refers to a particular independent variable and n is the dimension of the independent variables

ARIMA (Auto Regressive Integrated Moving average)[17]

An ARIMA model is a class of statistical models for analyzing and forecasting time series data A pure Auto Regressive (AR only) model is one where Yt depends only on its own lags. That is, Yt is a function of the 'lags of Yt'

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \varepsilon_1$$
(6)

where, Yt+1 is the lag 1 of the series, βt is the coefficient of lag 1 that the model estimates and α is the intercept term, also estimated by the model. Likewise, a pure Moving Average (MA only) model is one where Yt depends only on the lagged forecast errors.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p}\varepsilon_t + \Phi_1\varepsilon_{t-1} + \Phi_2\varepsilon_{t-2} + .. + \Phi_q\varepsilon_{t-q}$$
(7)

where the error terms are the errors of the autoregressive models of the respective lags. The errors Y_t and $Y_{(t-1)}$ are the errors from the following equations:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_0 Y_0 + \varepsilon_t$$
$$Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} + ... + \beta_0 Y_0 + \varepsilon_{t-1}$$
(8)

For AR and MA models respectively.

An ARIMA model is one where the time series was different at least once to make it stationary by combining the AR and the MA terms. So, the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p}\varepsilon_t + \Phi_1\varepsilon_{t-1} + \Phi_2\varepsilon_{t-2} + .. + \Phi_q\varepsilon_{t-q}$$
(9)

Predicted Yt = Constant + Linear combination Lags of Y (up to p lags) + Linear Combination of Lagged forecast errors (up to q lags)

6. Long Short Term Memory (LSTM)[18]

Long short-term memory or LSTM network is a popular deep learning method in time series forecasting that was first introduced by Hochreiter and Schmidhuber in 1997 [HS97]. The LSTM has a Recurrent Neural Network (RNN) architecture, which uses a loop to pass information from one step of the network to the next. The LSTM is a variation of the RNN with a similar recurrent or chain-like structure but more layers. The key of the LSTM is the cell Ct that stores memory and three gates (input gate, output gate, and forget gate) that control the information added to or removed from the cell. Two activation functions play an essential role in the LSTM network. The first one is a sigmoid function, also known as a logistic function. It regulates how much information to be passed through in these gates, which is defined as:

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x} \quad (10)$$

The outputs of this function are the numbers within range 0 to 1.If the result is close to 0, less information is allowed to be let through. In contrast, if the result is close to 1, more information can be involved in the next step. The second one is a hyperbolic tangent or a tanh function that outputs the values within range -1 to 1,

which is defined as in eq(11). A LSTM unit involves four parts, illustrated in Figure below.

The first part is called a forget gate layer that decides which previous information is discarded from the cell. It passes the hidden vector from the previous step h_{t-1} and new input X_t into the sigmoid function and produce a forget gate vector f_t from the function:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

(11)

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

$$h_t = o_t \text{ o } \tanh(C_t)$$
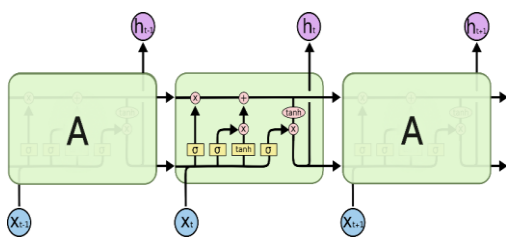

Figure 3. LSTM

7. Results and Discussions

Model	RESULTS				
	No Of Observations		MAE	MSE	RMSE
	Train	Test			
Multivariate Linear Regression	1327 (90%)	148 (10%)	10827.48	20.11	104.06
	885 (60%)	590 (40%)	6551.43	19.94	80.84
LSTM	1106 (75%)	369 (25%)	44680334	5210.64	6684
	883(60%)	590 (40%)	20988897.51	3057.66	4581.36
	1327 (90%)	148 (10%)	187490364.34	13309.83	13692.71
ARIMA	1106 (75%)	369 (25%)	1.70	1.02	1.31
	883 (60%)	590 (40%)	2.78	1.29	1.67
	1327 (90%)	148 (10%)	0.13	0.27	0.37

As we can clearly see from the table, ARIMA model is giving the least error for our dataset. Then Multivariable Linear Regression is giving the minimum error. LSTM is the worst performing model among these three.

8. Improvement as per reviewer comments

This system can be enhanced into an interactive web-based system where the user can know about the details on the go, can fetch the price data in real time and choose the crypto-currencies, they want to invest in. Also, the portal can be developed in such a way that once the user chooses which crypto currency to invest in, he/she can get a hyperlink for the same and they can directly reach the site from where they can invest in it. Also, we can integrate the other time series forecasting models on which the research is currently going on, in the future.

References

[1] J. Grossman and R. J. Shiller(1980), The Determinants of the Variability of Stock Market Prices, National Bureau of Economic Research, Working Paper 564.

[2] Giuliano Armano, Michele Marches,Andrea Murru (2005), Hybrid genetic-neural architecture for stock indexes forecasting Information Sciences 170(1):3-33

[3] Erol Egrioglu ,Cagdas Hakan Aladag,Ufuk Yolcu,(2013), Fuzzy time series forecasting with a novel hybrid approach combining fuzzy c-means and neural networks, Expert Systems with Applications 40(3):854–857.

[4] Kazem, A., Sharifi, E., Hussain, F.K., Saberi, M., & Hussain, O.K. (2013). Support vector regression with chaos-based firefly algorithm for stock market price forecasting. Appl. Soft Comput., 13, 947-958.

[5] 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), 2018, pp. 339-343, doi: 10.1109/PDP2018.2018.00060

[6] Alvin Ho1, Ramesh Vatambeti, Sathish Kumar Ravichandran1 Bitcoin Price Prediction Using Machine Learning and Artificial Neural Network Model, Indian Journal of Science and Technology, Year: 2021, Volume: 14, Issue: 27, Pages: 2300-2308, DOI 10.17485/IJST/v14i27.878

[7] Mrs Vaidehi M , Alivia Pandit et al., Bitcoin price prediction using machine learning. International Journal of Engineering Technologies and Management Research, 8(5), 20. doi: 10.29121/ijetmr.v8.i5.2021.953

[8] Marco Stocchi Michele Marchesi(2017) Fast wavelet transform assisted predictors of streaming time series, DOI: 10.1016/j.softx.2017.09.006

[9] Miller, Dante, and Jong-Min Kim. 2021. Univariate and Multivariate Machine Learning Forecasting Models on the

Price Returns of Cryptocurrencies. Journal of Risk and Financial Management 14: 486. https://doi.org/10.3390/jrfm14100486.

[10] S. Roy, S. Nanjiba and A. Chakrabarty, "Bitcoin Price Forecasting Using Time Series Analysis," 2018 21st International Conference of Computer and Information Technology (ICCIT), 2018, pp. 1-5, doi: 10.1109/ICCITECHN.2018.8631923.

[11] Leopoldo Catania, Stefano Grassi, Francesco Ravazzolo , Predicting the Volatility of Cryptocurrency Time-Series: MAF 2018, Mathematical and Statistical Methods for Actuarial Sciences and Finance

[12] S. Lauren and S. D. Harlili, Stock trend prediction using simple moving average supported by news classification," in Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014 International Conference of. IEEE, 2014, pp. 135-139.

[13] Rathindra Sarathy, Krish Muralidhar,Perturbation Methods for Protecting Numerical Data: Evolution and Evaluation,Editor(s): Ranajit Chakraborty, *C.R. Rao, Pranab* Sen,Handbook of Statistics, Elsevier, Volume 28, 2012, Pages 513-531,ISSN 0169-7161,ISBN 9780444518750

[14] A. Iwok, A. S. Okpe, A Comparative Study between Univariate and Multivariate Linear Stationary Time Series Models, American Journal of Mathematics and Statistics p-ISSN: 2162-948X.e-ISSN :162-8475,2016;6(5): 203-212, doi:10.5923/j.ajms.20160605.02.

[15] Nicola Uras et al., Forecasting Bitcoin closing price series using linear regression and neural networks models,Uras et al. (2020), PeerJ Comput. Sci., DOI: 10.7717/peerj-cs.279

[16] Uras, N., Marchesi, L., Marchesi, M., & Tonelli, R. (2020). Forecasting Bitcoin closing price series using linear regression and neural networks models. PeerJ. Computer science, 6, e279. https://doi.org/10.7717/peerj-cs.279.

[17] A. A. Ariyo, A. O. Adewumi and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, 2014, pp. 106-112, doi: 10.1109/UKSim.2014.67.

[18] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques.Morgan Kaufmann, 2005.

[19] F. A. Gers, D. Eck, and J. Schmidhuber, Applying LSTM to time series predictable through time-window approaches," pp. 669{676, 2001

[20] S. Hochreiter and J. Schmidhuber, LSTM can solve hard long time lag problems Advances in neural information processing systems, pp. 473-479, 1997.