

Chapter 39

Bitcoin Prediction and Time Series Analysis



Krishna Chakravarty, Manjusha Pandey, and Siddharth Routaray

39.1 Introduction

The most important part of any research is data collection and its analysis. The collected data is summarized and interpreted by statistical and logical methods to identify patterns which can predict relationships or trends. Time series prediction has been in use to predict stable financial markets like the stock market and in-depth research is ongoing in this field. Python & R are the main technologies for the daily data processing chores for today's data scientist.

For controlling the creation of additional units, verifying assets transaction and securing all transactions, cryptography is used and Cryptocurrency, which can be referred as digital or virtual asset work as the medium of exchange. The control is decentralized here in comparison with centralized economy like centralized banking systems. A public transaction database through which the decentralized control works for each cryptocurrency is termed as *Blockchain*, which functions as distributed ledger. Some of the transactional properties are: (1) *Irreversible*: transaction cannot be reversed after confirmation, (2) *Pseudonymous*: real-world identities are not connected to accounts or transactions, (3) *Fast and global*: instantaneous transactions, (4) *Secure*: a public key cryptography system is in place where cryptography funds are locked and (5) *Permission-less*: no permission is required to use cryptocurrency. Adoption of bitcoin as a leading cryptocurrency is growing consistently in the world at present. Bitcoin presents an interesting platform in time series prediction problem as it is still in its transient stage. As a result, the market is highly volatile and thus there is an opportunity in terms of prediction. At present, the factors affecting bitcoin price are (1) bitcoin supply and

K. Chakravarty (✉) · M. Pandey · S. Routaray
KIIT Deemed to be University, Patia, Bhubaneswar, Odisha, India
e-mail: manjushafcs@kiit.ac.in; siddharthfcs@kiit.ac.in

increasing/decreasing demand, (2) regulations enforced by governments on bitcoin transactions, (3) bitcoin users and developers influence the rise and fall of price, (4) bitcoin in news, the influence of media on garnering negative and positive publicity and (5) new technological changes to bitcoin. Bitcoin has open nature and it operates on a decentralized, peer-to-peer system which is termed as trust-less as all transactions are irreversible in nature and recorded on an open ledger called blockchain. Transparency of this level is not common in other financial markets. Ethereum (ETH) is a decentralized software platform, which was launched in 2015, helps to create Distributed Applications and smart contracts. These applications can run without downtime and will not be interfered by fraud or any third-party control.

Cryptocurrency Analysis: Fundamental analysis implicates financial health evaluation and the chance of survival of a company, which is an essential input for stock investments. This is mainly done by analysing company's financial statements. Good numbers give us confidence that the company has good fundamentals and we tend to invest. But in case of cryptocurrency fundamental analysis, absence of financial statements has made the situation radically different. There are no financial statements because: (1) Financial statements are applicable for corporations whereas cryptocurrencies are actually representations of assets within a network; they are not related to any corporations. Its value directly depends on the community participants like developers, users and miners and not impacted by any revenue generating system.

Through different applications of Blockchain technology, these decentralized cryptocurrencies are manifested. (2) The present time can be considered as infant stage of cryptocurrency which is mainly the development stage and thus limits the real-world use cases. There is a lack of track record and we need a different methodology for fundamental analysis. The current situation and the complex nature of crypto technology demands more work in research field to evaluate the viability and potential of any cryptocurrency. More understanding and investigation in this area will ensure more informed investment decisions. The in-depth knowledge of a currency's fundamentals will also help us to form our own opinion which is not common in the complex crypto world. Creating our own stand will definitely be of unique phenomenon.

While still in its beginning stages, big data analytics is starting to be used to analyse bitcoin and other cryptocurrencies. While many may decry the potential uses of big data analytics for cryptocurrency like identifying users, saying that such uses undermine the spirit of cryptocurrency itself, there are still ways in which big data analytics can legitimately benefit cryptocurrency, such as by identifying fake or dangerous users, preventing theft, and predicting trends. Cryptocurrency analysis is applied in the predicting trends of cryptocurrency prices. Throughout bitcoin's short history, it has been affected many times by world events and overall bitcoin community sentiment. For example, analysts analysed social media after the shutdown of Mt. Gox, once the largest bitcoin exchange, in early 2014. Social media trends were used to identify community sentiment, key voices, and stakeholders, and then tie this information to things like the currency's price performance, which is similar to other financial assets that can be affected by major events.

39.2 Literature Survey (Table 39.1)

39.3 Proposed Cryptocurrency Prediction Analysis

A series of data points indexed (or listed or graphed) in the order of time is termed as time series. Usually successive equally spaced points in time are recorded to form a time series sequence. Meaningful statistics and other characteristics of time series data are extracted by methods which fall under the umbrella of time series analysis. A model is used to predict future values based on previously recorded values, this process is termed as time series forecasting. The regression analysis which aims at value comparison of a single time series or multiple dependent time series at different points in time cannot be considered as “time series analysis”. For our data we are required to predict the high, low, or close values of the bitcoin. In such

Table 39.1 Details of literature survey

Author, Year	Title	Purpose
Tian Guo and Nino Antulov-Fantulin February 2018	“Predicting short-term Bitcoin price fluctuations from buy and sell orders”	Analysis of short-term fluctuation of bitcoin market mixture model was proposed to capture the dynamic effect of order book features and to provide interpretable results
Jeffrey Chu, Saralees Nadarajah, and Stephen Chan July, 2015	“Statistical analysis of the exchange rate of bitcoin”	The log-returns of the exchange rate of bitcoin as compared to the United States Dollar have been statistically analysed
Evita Stenqvist and Jacob Lonno, 2017	“Predicting bitcoin price fluctuation with twitter sentiment analysis”	Twitter data related to bitcoin is studied and sentiment analysis is performed. This study shows how twitter analysis can help in prediction of probable future variation of bitcoin price
Abhyudit Bisht, Puru Agarwal, 2017	“Analysis of bitcoin using linear regression and data mining techniques”	Cryptocurrency historical data is used with the help of data mining techniques and regression algorithms to predict different attributed like volume, market cap, etc.
Nakamoto, S Google Scholar, 2008	“Bitcoin: a peer-to-peer electronic cash system”	Instead of trust, the need for an cryptographic proof based electronic payment system is investigated. This allows willing parties to transact directly without any trusted third party interference

cases there is no independent variable like in multivariate systems where we have a dependent variable and an independent variable. It's just the values that come up for particular attributes based on the dates and other upcoming dates. Our data is a data properly curated with all the attributes for a particular day. Moreover, our data is consistent and every record is for a particular day which is constant, hence it is a time series based data. So for such kind of problems we either prefer time series based algorithms such as ARIMA or MA or AR or ARMA and deep learning based models such as RNN and LSTM. Regression and tree-based algorithms are perfectly supervised learning where we have an output variable against every record or feature set. But our dataset comes under semi-supervised learning where we particularly don't have any output variable rather we have a certain trend on particular features which is dependent on time. Hence the liner regression model's basic assumption that observations are independent is true in this scenario. Most time series show increasing or decreasing seasonality trends which means for a particular time frame, variations can be observed. For example, if we observe the sales of winter jackets over time, data will show higher sales value in winter seasons because of obvious reasons.

Essentially when we model a time series we decompose the series into four components: **trend, seasonal, cyclical, and random**. The random component is called the residual or error. It is simply the difference between our predicted value(s) and the observed value(s). Serial correlation is when the residuals (errors) of our TS models are correlated with each other. In layman's terms, ignoring autocorrelation means our model predictions will be bunk, and we're likely to draw incorrect conclusions about the impact of the independent variables in our model.

Terms relating to serial correlation:

Expectation: The expected value $E(x)$ of a random variable x is its mean average value in the population. We denote the expectation of x by μ , such that $E(x) = \mu$.

Variance: The variance of a random variable is the expectation of the squared deviations of the variable from the mean, denoted by $\sigma^2(x) = E[(x - \mu)^2]$.

Standard deviation: The standard deviation of a random variable x , $\sigma(x)$, is the square root of the variance of x .

Co-variance tells us how linearly related any two variables are. The co-variance of two random variables x and y , each having respective expectations μ_x and μ_y , is given by $\sigma(x,y) = E[(x - \mu_x)(y - \mu_y)]$. Co-variance tells us how two variables move together.

Correlation—It is a dimensionless measure of how two variables vary together, or “co-vary”. In essence, it is the co-variance of two random variables normalized by their respective spreads. The (population) correlation between two variables is often denoted by

$$\rho(x, y) : \rho(x, y) = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y} = \frac{\sigma(x, y)}{\sigma_x \sigma_y}$$

If a TS's statistical properties such as variance and mean do not vary over time, the TS can be referred as a stationary time series. Very strict parameters are used to define stationarity. To assume a TS to be stationary in practical situations, we can consider some factors like (1) variance and (2) mean are constant over time, and (3) contrivance is independent of time. The dataset can be observed to show overall increasing trend along with some seasonal variation. But it is not always possible to infer from the visual presentations. Hence we will be using the following statistical methods.

Plotting Rolling Statistics: The moving average or moving variance can be plotted and can be observed if it varies with time. 'Moving average/variance' means that we'll take the average/variance of the last year (last 12 months) at any instant time ' t '.

Dickey-Fuller Test: Time series stationarity can be checked by this test. TS is considered as non-stationary as part of the null hypothesis here. A test Statistic and some critical values for difference confidence levels can be furnished as test results. If the 'Critical Value' is greater than the 'Test Statistic', the null hypothesis can be rejected and the time series can be called as stationary.

In real-world situations almost no time series can be considered as stationary. Therefore we can take help of statistics to make a series perfectly stationary. Although this job is almost impossible in reality but we can try to make it as close as possible. Two factors are contributing to the non-stationarity nature of a TS: (1) *Trend*: over a period of time, the mean varies. For example, population in an area is growing over time. (2) *Seasonality*: variations according to specific time frames, e.g. people may buy cars or home appliances in a particular month because of festival deals or pay increment.

First method to estimate and eliminate trend is to reduce trend through transformation. For example, for the significant positive trend a transformation can be applied which penalize higher values more than smaller values. A log, square root, cube root, etc. can help here. To make the TS stationary, different statistical techniques work well where we have to forecast a time series. We can create a model on TS through this most used technique. Once the exercise is completed, it is relatively easy to add noise and seasonality back into predicted residuals in this situation. In these estimation techniques to test trend and seasonality, we can observe two cases: (1) A rare case where the values are independent to each other and we can term this as a strictly stationary series. (2) When the values are significantly dependent on each other. In this case ARIMA, a statistical model, is used to forecast the data.

The full form of ARIMA is Auto-Regressive Integrated Moving Averages. A linear equation (e.g. linear regression) is used to forecast the time series in this model. The parameters of this model are (p, d, q) : (1) Number of AR (Auto-Regressive) terms (p): this is the number of lag observations present in the model, also called the lag order. For example, if p is 6, the predictors for $x(t)$ will be $x(t - 1) \dots x(t - 6)$. (2) Number of MA (Moving Average) terms (q): this is the

size of the moving average window, also called the order of moving average. These are lagged forecast errors in prediction equation. For example if q is 6, the predictors for $x(t)$ will be $e(t-1) \dots e(t-6)$. (3) Number of Differences (d): number of non-seasonal differences are referred here, also called the degree of differencing. We can pass this variable and put $d = 0$ or pass the original one and make $d = 1$. Both the cases yield same results.

Our dataset is split into train and test data beforehand only, i.e. we store our train dataset as `bitcoin_train.csv` file and test dataset as `bitcoin_test.csv` file. The model trains itself then tests itself on the test data (Table 39.2).

Table 39.2 Pseudo-code used in ARIMA prediction

Algorithm: Predictions using ARIMA (Auto-Regressive Integrated Moving Averages) Mod

Preprocessing: `Import(statsmodels.tsa.arima_model)`

Input: `Ds= Data Set (bitcoin_test.csv)`
`Dst= Training Data Set`
`Dss=Testing Data Set`

Step1: *Work on training data, grouping by time series, remove trend and seasonality*
1.1 `dataset['Close'].plot()` #Plotting the Yearwise trend of dataset
1.2 `dataset = dataset['Close']` ##As here we want to predict the Closing Value of the bitcoin for every day
1.3 *#Rounding up of data to a weekly, monthly, yearly, quarterly basis*
`weekly=dataset.resample('W').sum(). weekly.plot()`
`monthly=dataset.resample('M').sum() .monthly.plot()`
`year=dataset.resample('Y').mean(). year.plot()`
`quarter=dataset.resample('Q').mean() . quarter.plot()`
1.4 *#Perform Dickey-Fuller test:*
`dfctest = adfuller(timeseries, autolag='AIC')`
`dfcoutput = pd.Series(dfctest[0:4], index=['Test Statistic', 'p-value', '#Lags Used', 'Number of Observations Used'])`
1.5 *#remove trend and seasonality*
`transform_dataset_log_t = np.log(transform_dataset)`
`seasonality expwighted_avg =`
`transform_dataset_log_t.ewm(halflife=7,min_periods=0,adjust=True,ignore_na=False)`
`.mean()`

Step2: #Fit to ARIMA model
`model = ARIMA(transform_dataset_log_t, order=(8, 1, 18)) results_ARIMA = model.fit(dispatch=-1)`

Step3: *#Prediction and forecast*
`predictions_ARIMA_diff=pd.Series(results_ARIMA.fittedvalues, copy=True)`
`predictions_ARIMA_diff_cumsum = predictions_ARIMA_diff.cumsum()`
`predictions_ARIMA_log = pd.Series(transform_dataset_log_t.iloc[0])`
`forecast = pd.Series(results_ARIMA.forecast(steps=7)[0],dates)`
`forecast = np.exp(forecast)`

Step4: Output
`print(predictions_ARIMA_diff_cumsum.head()).print(forecast)`
#Plotting the Actual and The PRedicted With the RMSE
`plt.plot(forecast, 'Predicted rates')`
`plt.plot(test, 'Observed from test data')`

39.4 Analysis and Result: Time Series Analysis

This section presents the results of time series analysis for bitcoin using python and the result generated for predicted values of the cryptocurrency were very much similar to the original values as depicted by the graphs below. The following Fig. 39.1 represents a graph of weekly, monthly, yearly and quarterly variation of bitcoin values which presents a steep increase between 2017 and 2018. The steep identified symbolizes the acceptance of bitcoin as preferred cryptocurrency by many people throughout the world.

The below code snippet represents logic used for checking the stationary characteristics of data. The characteristics are checked with the help of dickey fuller test that considers rolling mean and rolling standard deviation (Fig. 39.2).

The following figures present the code snippet and the results achieved after removing the trend and seasonalities using log transformations. This is done to make our data set stationary for further experiments (Figs. 39.3 and 39.4).

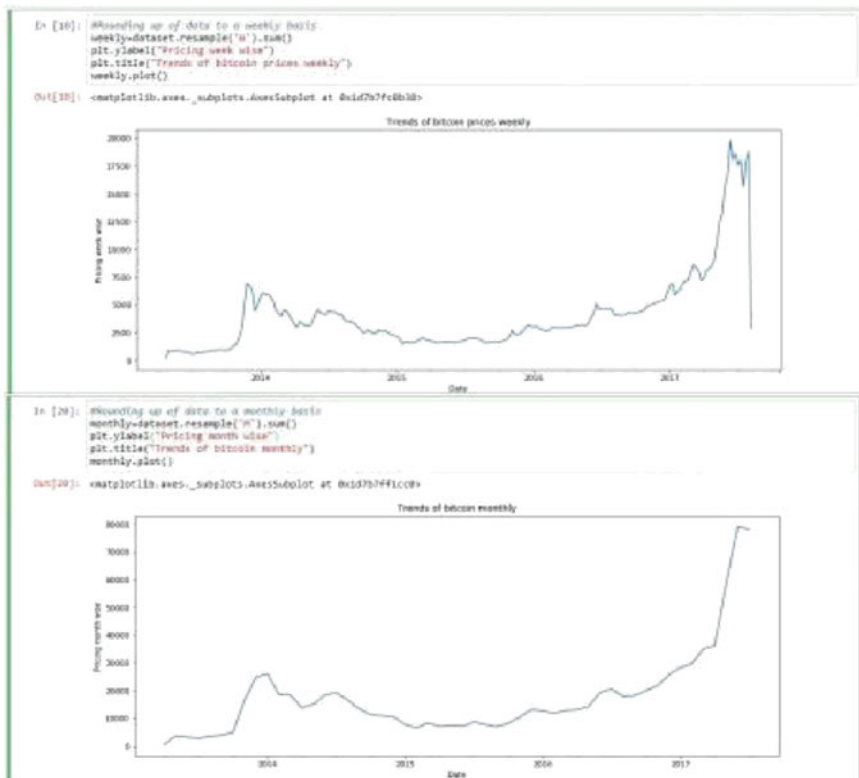


Fig. 39.1 Plotting the graph on a weekly, monthly, yearly and quarterly

```
In [14]: def test_stationarity(timeseries):
#Determining rolling statistics
rolmean = pd.rolling_mean(timeseries, window=12)
rolstd = pd.rolling_std(timeseries, window=12)

#Plot rolling statistics:
orig = plt.plot(timeseries, color='blue',label='Original')
mean = plt.plot(rolmean, color='red', label='Rolling Mean')
std = plt.plot(rolstd, color='black', label = 'Rolling Std')
plt.legend(loc='best')
plt.title('Rolling Mean & Standard Deviation')
plt.show(block=False)

#Perform Dickey-Fuller test:
print ('Results of Dickey-Fuller Test:')
dftest = adfuller(timeseries, autolag='AIC')
doutput = pd.Series(dftest[0:], index=['Test Statistic','p-value','lags used','Number of Observations used'])
for key,value in dftest[0:].items():
    #output['critical value (5%)'] = value
    print (doutput)
```

```
In [10]: #Dickey Fuller Test on The Data
test_stationarity(transform_dataset)
```

Fig. 39.2 Checking stationarity of data by dickey fuller test for rolling mean, standard deviation

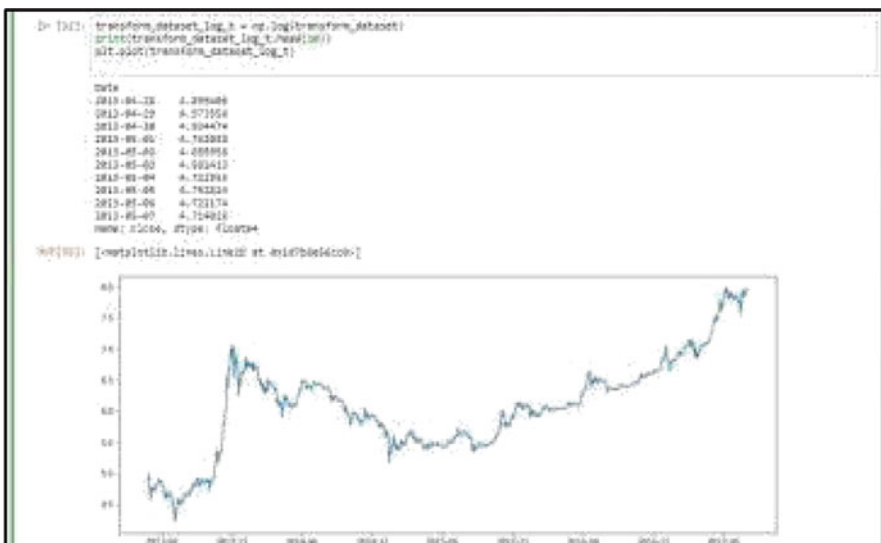


Fig. 39.3 Removing the trends and seasonality by performing log transformation to make our dataset stationary

The following figure is the most important result generated throughout our experiments for Time Series Analysis of change in values for bitcoin using ARIMA model. The ARIMA model is one of the most preferred models for analysis of dynamically changing datasets as of cryptocurrencies. The result thus obtained was used for further forecasting for open and closed values of bitcoin for coming days which was then compared with the actual values. The results thus obtained were encouraging and presented an accuracy of 99% (Figs. 39.5 and 39.6).

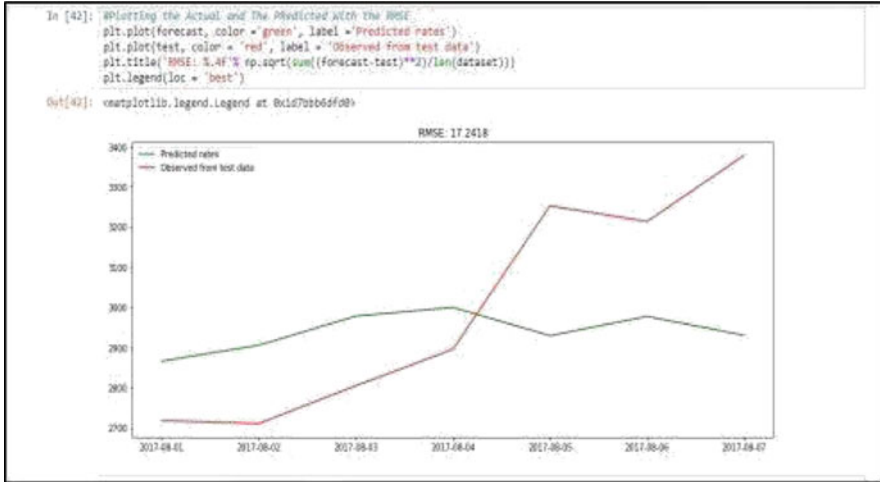


Fig. 39.6 Graph depicting the actual and the predicted

learning algorithms, i.e. Linear Regression and Polynomial Regression. Since the accuracy of Linear Regression was higher than Polynomial Regression we prefer the former model. Using time series analysis, we plotted the variations of close values over weekly, monthly, quarter-yearly and yearly basis.

Through the predictive analytics, graphical visualizations and machine learning, we could estimate the trends of bitcoin.

We would like to implement complex algorithms like RNN and CNN of deep learning. We would also try to look into methods or modify our existing algorithms which can give more accuracy percentage for prediction. We would also try to make a complete graphical user interface for administrator to view analysis and predict more efficiently by entering a specific date.

References

1. Muhammad Amjad, Devavrat Shah, Trading bitcoin and online time series prediction, in *NIPS 2016 Time Series Workshop* (2017)
2. T.G. Andersen, T. Bollerslev, F.X. Diebold, P. Labys, Modeling and forecasting realized volatility. *Econometrica* **71**, 2 (2003)
3. F. Black, M. Scholes, The pricing of options and corporate liabilities. *J. Polit. Econ.* **81**, 3 (1973)
4. W. Bolt, On the value of virtual currencies. *SSRN Electronic J.* (2016)
5. S. Brahim-Belhouari, A. Bermak, Gaussian process for non-stationary time series prediction. *Comput. Stat. Data Anal.* **47**, 4 (2004)
6. Tianqi Chen, Carlos Guestrin, Xgboost: A scalable tree boosting system, in *SIGKDD (ACM, 2016)*

7. D.L.K. Chuen, *Handbook of digital currency: Bitcoin, innovation, financial instruments, and big data* (Academic Press, 2015)
8. J. Civitarese, Volatility and correlation-based systemic risk measures in the US market. *Physica A* **459**, 55–67 (2016)
9. Jonathan Donier, Jean-Philippe Bouchaud, Why do markets crash? Bitcoin data offers unprecedented insights (2015)
10. H.N. Duong, P.S. Kalev, C. Krishnamurti, Order aggressiveness of institutional and individual investors. *Pacific-Basin Finance J.* **17**, 5 (2009)
11. S. Nakamoto, *Bitcoin: A peer-to-peer electronic cash system*, in *Google Scholar*, (2008)
12. F. Mosteller, J.W. Tukey, *Data analysis and regression: A second course in statistics* (Addison-Wesley, Reading)
13. A. Gelman, J. Hill, *Data analysis using regression and multilevel/hierarchical models* (Cambridge University Press, Cambridge)
14. <https://www.coindesk.com/information/understanding-bitcoin-price-charts/>
15. <https://www.cs.waikato.ac.nz/ml/weka/>
16. Kenshi Itaoka, Regression and interpretation low R-squared
17. S.L. Nelson, E.C. Nelson, *Excel-data-analysis-for-dummies* (Wiley, Weinheim)