Forecasting the Price of Cryptocurrencies and Validating Using Arima

Ruchi Mittal[a], Rashmi Gehi[b], M.P.S Bhatia[a]

[a]*Division of Computer Engineering, Netaji Subhas Institute of Technology, New Delhi, India*
[b]*Division of Electronics and Communication Engineering, Netaji Subhas Institute of Technology, New Delhi, India*

Abstract: With the increase in popularity of cryptocurrencies, it is becoming extremely crucial to predict what the prices of the currencies are going to be in the future. This paper uses a dataset that consists of over 1500 cryptocurrencies with their prices starting from their initiation till May, 2018. A lot of the effort went into getting the data set ready before predicting the future prices of all the cryptocurrencies, i.e., making sure that the cryptocurrencies were stationary time-series. Beginning with learning about the ARIMA model and the conditions to run the model successfully, first validation of the model is done. An average accuracy of 86.424 is observed for 95% of the currencies are observed. After this validation, forecasting is performed on these cryptocurrencies and the percentage change of the price is calculated.

1. Introduction

Cryptocurrency, also popularly known as digital currency, is a medium of exchange that uses cryptography to secure financial transactions. It is a decentralized form of currency as compared to any central banking system. Cryptocurrencies use the blockchain technology that works on a very simple logic, which is that of having multiple copies of the same information. Or in a sense, a 'ledger' that is public information and can be accessed by anyone. Several individuals keep a personal copy of the ledger, which stops from any manipulation of data as whenever a change in information is noticed, the fraud is caught immediately. While 'miners' are using technology to verify the transactions by decoding complex puzzles, data scientists are using technology to predict the future prices of the cryptocurrencies to know how the market will change. Most use various machine-learning algorithms to figure this out. Machine Learning, a subset of artificial intelligence, uses statistical techniques that allow the computer to 'learn' from the data provided to them without explicitly programming them. Learning simply refers to the machine improving how to perform a certain task based on past experience. It has recently become an extremely popular practice of letting computers figure out patterns from data and learn from it as well as humans can. Another aim of machine learning apart from learning is abstract learning and presenting the knowledge the machine learns. Machine learning is popular in classification and prediction. Classification is the property of the machine where the machine can recognize and categorize things based on the data that is fed into it. Prediction is predicting future data by learning from past data. Thus, to simply to put it into words it is the extraction of knowledge from data. Multiple algorithms such as random forest regression have been used in the past to do forecasting. However in this paper we have used the ARIMA model to forecast values. The ARIMA model is a statistical analysis model that aims as understanding the data set better and producing results such as future prospects of the data from an input data that is strictly in time series. In short, ARIMA model works on three components of auto-regression, integration and moving average to predict future possibilities of a financial market by examining the differences in the values in series instead of the actual values.

ARIMA is a model that uses different model parameters according to the characteristics of the data. ARIMA requires the data to be stationary which will further be discussed in this paper and generated a future prediction of the value that is required. The model depending on its previous predictions can make multiple predictions. This model is majorly dependent on the previous data and the patters that the data has observed.

2. Illustrations

Bitcoin had dominated the decentralized banking market space from the beginning of its initialization in 2009. However, the exponential popularity of Bitcoin has also paved the way for thousands of other cryptocurrencies to come up and create a name for them. Extensive energy and hard work have been put into researching about Bitcoin and other highly popular cryptocurrencies like Ethereum, Litecoin, Ripple, etc. Grinberg R. in [1] talks of how Bitcoin is a great alternative digital currency. Nakamoto S. in [2] explains the peer-to-peer network. Herrera-Joancomartí in [3] extensively concentrates on the blockchain technology of Bitcoin and how the mining process works. It also puts in a lot of importance on the Bitcoin network and the anonymity of Bitcoin. Reid F. et al. in [4] also speak about the anonymity of Bitcoin. There are other papers too that have a broader view of Bitcoin such as Garcia D. et al. in [5] quantify different socio-economic factors of Bitcoin and Yermack D. in [6] talks on using Bitcoin as an actual currency.

The ARIMA model has been widely used in the past for forecasting, for example, McNally S. in [7] focuses on predicting the price of Bitcoin exclusively, using the ARIMA model and long short-term memory model of deep learning and compared the results found by both the models. Leopoldo Catania et al. in [8] focuses on using uni-variate and multi-variate models and combinations of both for forecasting the price of the above mentioned four most capitalized cryptocurrencies. Zhang G.P. in [9] focuses on time-series forecasting by creating a hybrid model of ARIMA and neural networks. Bakar, N. A. et al. in [10] has concentrated on using the ARIMA model to forecast the price of Bitcoin in a high volatility environment.

However, very little has been done to evaluate the performance of the thousands of cryptocurrencies that have come up over the years. Most of the concentration has been on Bitcoin, Litecoin, Ethereum, and Ripple.

This paper will concentrate on the performance of all the cryptocurrencies.

Paper Outline: In section 2, we discuss the related work in the area of cryptocurrency and arima model. Section 3 of this paper will concentrate on getting the data ready to use as per requirements. This will involve figuring out the variables that will dominate the others in predicting future values and also treating the missing values. Section 4 of this paper will deal with employing ARIMA to forecast the required values. Finally, the focus will be put on validating the applied model to ensure that the predicted values by the model can be relied upon. Figure 1 shows the process flow for the duration of this paper.

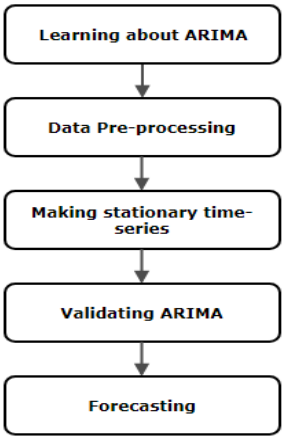

Fig. 1 – Process Flow

3. Data Preprocessing

The data that has been used in this paper is an extensive collection of cryptocurrencies that have existed as early as 2013 till the 19th of May, 2018. There are a total of 1592 cryptocurrencies in this dataset, and the numbers of observations of each cryptocurrency vary from over 1800 for currencies like Bitcoin, Litecoin, Namecoin to less than 15 observations for currencies like Edu.coin, Bank.coin, Zippie, etc.

The dataset contains 13 variables some of which are the currency name, its symbol, the date, the highest and lowest price of the currency of the day, the opening and the closing price of the currency of the day, its market capitalization, etc. In this paper, only the highest rate of the currency is used, and the same has been forecasted. Using multiple variables to predict the price of the currency is beyond the capacity of this paper.

To predict the future values of the highest price, also known as forecasting, ARIMA model has been used. ARIMA, which is an acronym for Auto Regression Integrated Moving Average, is widely used primarily when the data provided is seasonal, is a class of a model that requires the data to be in time series. This needed to convert the dataset, which was formerly of the class data.frame to time-series. One thousand five hundred ninety-two time-series were created, one for each cryptocurrency, to allow the calculation of the highest price individually.

A time-series can be checked if it is stationary only when the data is continuous, i.e., there should be no date for which the price of the cryptocurrency is missing. However, while going through the data, it was found that around one-third of the cryptocurrencies had dates for which the highest price had not been recorded. This would have caused the order of the ARIMA model to be inaccurate because of the missing values. To deal with these missing values and to also maintain the sanctity of the prices, the missing values were replaced by the latest, not missing values. For example, if we consider the cryptocurrency entcash, we observe that we do not have the information about the prices of entcash from the 11th of January, 2018 to the 27th of January, 2018. Hence, the price of entcash for the 10th of January, 0.157429, is used as a reference and made equal to the rates of the missing dates. The same procedure is followed for all the cryptocurrencies that have any dates for which their price is not available. This ensured that the data was complete and could be modeled as a stationary time-series.

Forecasting has also been performed only on those currencies that have more than 15 observations as the predictions of the highest price of such currencies would prove to be highly inaccurate and because there is much less information known on these currencies.

4. Cryptocurrency Price Prediction Approach

ARIMA is mainly a function of 3 variables:

1. AR(p)

This refers to the auto regressive component of ARIMA. It specifies the number of lags used in the model and is used to account the use of the history of the series. For AR(2) the equation would be

$$y_t = c + \Phi_1 y_1 + \Phi_2 y_{t-2} + e_t \quad (1)$$

Where Φ_1, Φ_2 are the parameters of the model and e_t is the error.

2. I(d)

This refers to the integrated component of ARIMA. It specifies the degree of differencing which is simply the subtraction of current values with the previous values. This parameter is used when the given series is not stationary.

3. MA(q)

This refers to the error as a combination of previous error terms.

$$y_t = c + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q} + e_t \quad (2)$$

These three parameters when considered together can be written as a linear equation.

$$y_t = c + \Phi_1 y_{dt-1} + \Phi_2 y_{dt-p} + \cdots + \theta_1 e_{t-1} + \theta_q e_{t-q} + e_t \quad (3)$$

Where y_d is the series differenced d times and c is a constant.

The ARIMA model works efficiently only if the time-series it is worked on is stationary. The time series has to satisfy the following three conditions to be stationary:

- The mean should not be a function of time.
- The variance of the time series should not be a function of time
- The covariance of the i^{th} and the $(i+m)^{th}$ term should not be a function of time.

A famous and reliable test that can be used to see whether the three mentioned conditions are met or not for a time-series to be stationary are verified by running the Dickey-Fuller test. The Dickey-Fuller test calculates the different parameters (p, d, and q) of the series and derives its result from these parameters. For a series to be stationary the p value should be as low as possible (around 0.01). If a series is found to not be stationary, with p value equal to 0.4 for example, it can be simply made stationary by differencing the series, which helps remove any trends in the series that might be present. As I (d) is increased, the order of the model is changed, which results in the reduction of the p value, hence making the series stationary.

In figure 2, the plot of the highest price distribution vs. the number of days for Bitcoin is shown. Bitcoin has been used as an example in this paper to explain a few of the concepts that are used in this paper. However similar analysis has been performed for all 1592 cryptocurrencies that we have available with us in the data. In this figure, we can see that the values of the highest prices gradually increase as the number of days increase. This is not stationary as the highest price is a function of time. ARIMA cannot be employed in such a situation. The stationary series of Bitcoin can be seen in the graphs as the difference is increased to 1, 5 and 10 respectively.

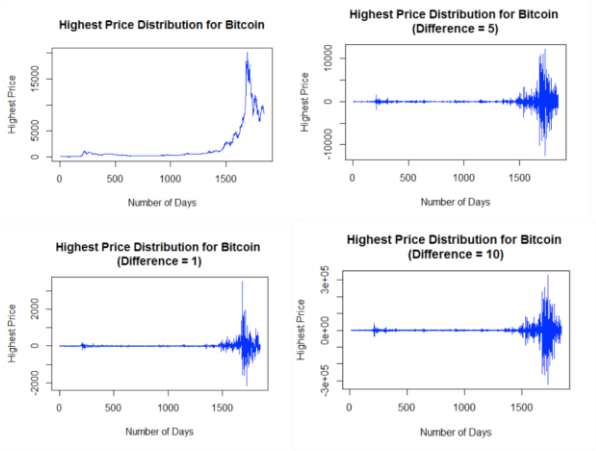

Fig. 2 - Highest price distribution of Bitcoin is shown in the above graphs. The first figure is not stationary but as the 'difference' parameter is changed, the distribution becomes more stationary, i.e., does not increase as the number of days increase.

5. Experimental Results

Because there is no method for us to validate the values predicted by our model for 30 days into the future, an alternative approach has been used for validating the model. To calculate the accuracy of the employed model, first, the last three observations of all the cryptocurrencies were eliminated from the dataset. Then, the model that was used for forecasting is applied to the new dataset with the excluded observations. The model was then used to predict the next three values, i.e., the same observations were eliminated from the dataset. The percentage of accuracy was calculated for the three predicted values and the three values that were removed. The percentage of accuracy for 15 cryptocurrencies has been shown in table 1. Here the values predicted by the model are incredibly close to the actual values and that the percentage accuracy is extremely high. This is because of the number of observations for each cryptocurrency being 1848 or lesser than that. In figure 3, the accuracy plot shows the distribution of the number of cryptocurrencies vs. the range in which their accuracy lies.

Table 1 - table for validating the ARIMA model implementation. The percentage of accuracy is calculated for each of the cryptocurrency for the last three iterations and then the average percentage of accuracy

Slug	Last Three Values			Three Predicted Values			Percentage Accuracy o three Values			Average Percentage
bitcoin	8445.54	8274.12	8372.06	8455.0753	8553.645	8631.2749	99.887096	96.621695	96.903809	97.8042003
ethereum	718.83	695.03	715.58	703.28534	703.28534	703.28534	97.837504	98.812233	98.281860	98.3105327
ripple	0.71461	0.68562	0.69045	0.6985825	0.7160301	0.7230633	97.756080	95.565347	95.276522	96.1993168
bitcoin.cash	1335.41	1218	1208.99	1359.36	1359.36	1359.36	98.206543	88.394088	87.562345	91.3876591
eos	13.94	13.18	13.35	12.542309	12.861255	13.341507	89.973524	97.581603	99.936384	95.8305043

vechain	nem	dash	neo	tron	iota	stellar	cardano	litecoin		
4.73	0.33028	417.8	63.22	0.07344	1.96	0.35048	0.26311	141.17		
4.57	0.31433	396.43	60.28	0.06953	1.78	0.31952	0.24680	137.06		
4.47	0.31516	402.83	60.64	0.07056	1.81	0.32910	0.24835	137.5		
4.7027678	0.3360713	426.28762	62.018954	0.0713599	1.9433493	0.345974	0.2579372	139.55842		
4.6911855	0.352448	427.90003	62.796745	0.0717417	1.9403726	0.3487873	0.2637822	144.90803		
4.7020755	0.3577248	426.26018	63.967031	0.0715646	1.9502889	0.3448667	0.2702494	149.60917		
99.424266	98.247774	97.968497	98.100212	97.165029	99.150477	98.712081	98.033593	98.858413		
97.348238	87.875380	92.061641	95.824908	96.819028	90.990302	90.840905	93.122955	94.274020		
94.808154	86.494234	94.183606	94.513472	98.581952	92.249238	95.210729	91.183802	91.193331		
97.193553	90.872463	94.7379152	96.146197	97.522003	94.130004	94.921238	94.113450	94.7752551		

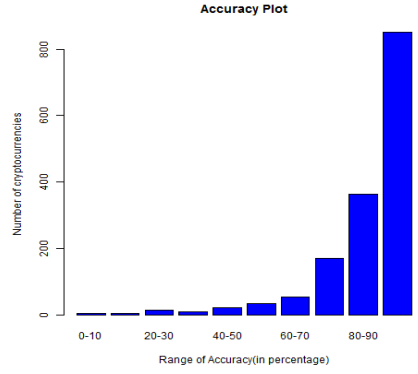

Fig. 3 – Accuracy plot to show the distribution of accuracy of the cryptocurrencies.

In figure 4, the predicted values of Bitcoin are shown. It is observed that the predicted values are very close to the latest iteration that we have in our dataset. This is because of the extreme range of the price of Bitcoin. From the beginning of Bitcoin's life till date, its price has risen exponentially, i.e., from a few cents to thousands of dollars, and the price has never changed in repetitive cycles (i.e., there has been no seasonal change in the price of Bitcoin). This was also observed in other cryptocurrencies. The prices of cryptocurrencies change very unpredictably just like stock prices and are not seasonal. Hence, the seasonality parameter of the ARIMA model was set to 'false' while forecasting. Table 2 shows the forecast of a few cryptocurrencies, the percentage change of the price has been calculated as per the last iteration of each cryptocurrency that we have with us. The same calculation has been performed on the rest of the cryptocurrencies too.

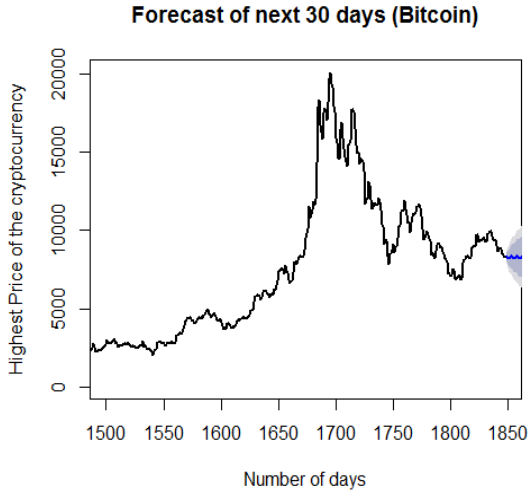

Fig. 4 –Forecast of Bitcoin for the next 30 days

Table 2 – profit table for the mentioned cryptocurrencies would have after 30 days. The negative sign indicates a drop in the price of the cryptocurrency

Slug	Last Iteration	Prediction for After 30 days	Profit
bitcoin	8372.06	8338.687407	-0.398618657
ethereum	715.58	721.9125381	0.884951797
ripple	0.69045	0.709797459	2.802152142
bitcoin.cash	1208.99	1208.99	0
eos	13.35	12.97083398	-2.840194878
litecoin	137.5	141.3503651	2.800265535
cardano	0.248354	0.245089545	-1.314436282
stellar	0.329105	0.320822596	-2.516644915

iota	1.81	1.801263891	-0.482657977
tron	0.070564	0.0722599	2.403350616
neo	60.64	60.3948552	-0.40426254
dash	402.83	404.8887935	0.511082462
nem	0.31516	0.315670903	0.162109037
vechain	4.47	4.434708389	-0.789521499

To validate our results, we compare the results with the approach proposed by Mittal R. in [11] for cryptocurrency price prediction using multi-variant linear regression. In table 3, we made the comparison of the ARIMA model approach with the multivariate approach for bitcoin cryptocurrency. From our comparisons, we found that the percentage of accuracy is a bit high with the ARIMA model compare to the multivariate regression model which concludes that for the price predictions of the cryptocurrencies, ARIMA model works better and provides many accurate results.

Table 2 - table for comparing the ARIMA model implementation with Multi-vartaite linear regression model. The percentage of accuracy is calculated for both the model for bitcion cryptocurrency

	Last Three Values		**Three Predicted Values**				**Average % Accuracy**
ARIMA MODEL (bitcoin Price)	8445.54	8274.12	8372.06	8455.07	8553.64	8631.27	97.8
Multi-Variate Lnear Regression Model (Bitcoin Price)	8445.54	8274.12	8423.10	8475.10	8573.55	8731.49	96.9

6. Conclusion

In this age, where cryptocurrencies are slowly creeping into the stock markets and making a name for themselves, it is becoming crucially important for us to figure out what the prices of cryptocurrencies are going to be in the coming days. With the high instability of cryptocurrencies and highly competitive central bank currencies, the majority of the population is very hesitant when it comes to investing in cryptocurrencies and skeptic of what the profits or losses are going to be. Starting with Bitcoin, in 2009, there have been over 1500 cryptocurrencies that have come up in the past 7 or so years. The exponential rise in the popularity and the price of Bitcoin, has paved the way for other currencies to come up and intrigue the customers.

This paper concentrated on forecasting the values of all the cryptocurrencies and also validating it. A lot of the focus was put into making sure that the data was ready to be programmed on. It was essential for the cryptocurrencies to be stationary time-series and this was achieved by changing the order of the parameters of the ARIMA model and only after this forecasting could be performed. The result mainly depended on the number of observations that we had of each cryptocurrency. And because the world of cryptocurrency is relatively new, the minimal views of a little over 1800 was a small restriction to achieve the result desired. Results could have improved if the number of observations was more.

Also, this paper concentrated on only the price of the cryptocurrencies, however in the future, it would be great to work on other parameters that control the change in the rates of the cryptocurrencies, for example, the relation between the market capitalization and the amount of the currency that is currently in the circulation of the market.

REFERENCES

Grinberg, R. (2012). Bitcoin: An innovative alternative digital currency. Hastings Sci. & Tech. LJ, 4, 159.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

Herrera-Joancomartí, J. (2014). Research and challenges on bitcoin anonymity. In Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance (pp. 3-16). Springer, Cham.

Reid, F., & Harrigan, M. (2013). An analysis of anonymity in the bitcoin system. In Security and privacy in social networks (pp. 197-223). Springer, New York, NY.

Garcia, D., Tessone, C. J., Mavrodiev, P., & Perony, N. (2014). The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy. Journal of the Royal Society Interface, 11(99), 20140623.

McNally, S., Roche, J., & Caton, S. (2018, March). Predicting the price of Bitcoin using Machine Learning. In Parallel, Distributed and Network-based Processing (PDP), 2018 26th Euromicro International Conference on (pp. 339-343). IEEE.

Yermack, D. (2015). Is Bitcoin a real currency? An economic appraisal. In Handbook of digital currency (pp. 31-43).

Catania, L., Grassi, S., & Ravazzolo, F. (2018). Forecasting Cryptocurrencies Financial Time Series.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.

Bakar, N. A., & Rosbi, S. (2017). "Autoregressive Integrated Moving Average (ARIMA) Model for Forecasting Cryptocurrency Exchange Rate in High Volatili.

Mittal, R., Arora, S., & Bhatia, M. P. S. (2018), "AUTOMATED CRYPTOCURRENCIES PRICES PREDICTION USING MACHINE LEARNING". ICTACT

Aparicio, T., & Villanua, I. (2003). Multivariate Linear Regression Model. In Computer-Aided Introduction to Econometrics (pp. 45-120). Springer, Berlin, Heidelberg.