Price prediction of speculative cryptocurrencies using ARIMAX model

Predrag Glavaš
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
email: pedja.glavas@uns.ac.rs

***Abstract*—With rising popularity of cryptocurrencies and constant increase in number of new investors many speculative and joke-themed currencies rose to prominency. As price of such cryptocurrencies is artificially inflated and deflated many new investors lose money. Good model must be found to adequately predict their prices in short term and help new investors avoid huge losses. Related work suggests that ARIMA with exogenous variables (ARIMAX) shows very good results in both cryptocurrency and stock price predictions. This paper attempts to improve model's accuracy in predicting prices of Dogecoin, Shiba Inu and IDENA by introducing community and developer data as predictors in ARIMAX model. Model hyperparameters were chosen with auto arima function in Python, criterion for model choice being Akaike information criterion (AIC). Further improvements in accuracy were made with correlation analysis using heatmaps. All models were evaluated with mean average percent error (MAPE) and it was determined that best generalization was achieved for prediction of IDENA one day ahead with MAPE of 7.46%. Models for other two cryptocurrencies remained sensitive to sharp price fluctuations and couldn't predict their amplitudes correctly.**

Keywords—cryptocurrencies, ARIMAX, price, prediction, dogecoin, idena, shiba

I. INTRODUCTION

Cryptocurrencies have exploded in price since the worldwide outbreak of Coronavirus in 2020. With the price increase and overall market expansion, a lot of highly speculative cryptocurrencies gained popularity among new investors. It is, therefore, very important to find an adequate model that can predict the price of these cryptocurrencies and lower a potential investment loss.

Dogecoin as one of such currencies, which was created purely as joke in 2013 quickly gained a loyal community following and its price skyrocketed from the end of 2020 to mid-2021 rising 100 times from 0.3 to 30 US dollar cents [1]. Another one in line of such speculative dog-themed currencies is Shiba Inu coin, which was dubbed "Dogecoin killer" by its creators. Driven only by a loyal community following, and an unrelenting influx of new holders, Shiba's price saw an increase of over 2500% since the coin's inception in 2020 [2]. IDENA coin, unlike Shiba and Doge was founded with an intention far more serious than a community joke. Core idea is that every node in the blockchain must prove that it is human by solving human made puzzles within a specified time. Despite having a real technological value, price of IDENA coin didn't grow nearly as much as Shiba or Doge, its price now in 2022 being lower than on launch in 2020 [3].

Goal of this paper is to determine whether the price of these cryptocurrencies can be accurately estimated based upon previous price, community and developer data. Community data was deemed potentially important as Doge and Shiba owe their success to a strong community following. Developer data was deemed important since IDENA has a viable technological use and also because Shiba began creating a serious crypto ecosystem [2] far greater than a community joke. ARIMAX model was chosen based upon literature. It proved to be very efficient for short term price prediction 1-7 days ahead, but it remained sensitive to sharp price fluctuations and couldn't accurately anticipate amplitudes of those fluctuations.

In the second chapter, literature was overviewed along with concrete takeaways from each paper listed. Chapter three explains how datasets were downloaded, aggregated and preprocessed. Fourth chapter explains methodology used for price prediction along with test results. In the final chapter, retrospective of the whole process was done and potential further improvements listed.

II. LITERATURE OVERVIEW

Algorithm and methodology in this paper were chosen based upon relevant literature found on the topic of cryptocurrency price prediction.

In the paper [4] published by Amin Azari, he tested the usefulness of ARIMA model in predicting the price of Bitcoin one day ahead. Data was sampled daily for a period of 3 years, preprocessed to ensure stationarity and tested over different subperiods to check its robustness. It was determined that ARIMA can indeed be used to predict cryptocurrency price, but cannot adequately anticipate sharp increases and decreases. This model could potentially be improved by adding exogenous variables such as community and developer data.

Paper [5] by Scalzotto Giovanni studies impact of Twitter on Ethereum and Dogecoin cryptocurrencies and tests various ARIMAX models to analyze influence of exogenous variables on price prediction. Data was sampled for first 4 months of 2021 by retrieving tweets with various crypto related hashtags.

Results show that Twitter has strong impact on Dogecoin's price as opposed to Ethereum, mainly because it is community driven.

In [6], macroeconomy-related time series data was analyzed, GDP and unemployment rate were forecasted using both ARIMA and ARIMAX models. Model parameters were estimated from PACF and ACF graphs. ARIMA model proved to be slightly more accurate than ARIMAX with MAPE of 1.77% compared to 3.78% of ARIMAX.

Publication [7] tests ARIMAX, XGBoost and Facebook prophet models for the purpose of Bitcoin price prediction. RMSE and R^2 metrics were used for evaluation, and they showed that ARIMAX was slightly better than other two with RMSE of 322.

III. Dataset forming

Dataset for each cryptocurrency was formed by sampling data daily from Yahoo Finance [8], CoinGecko API [9] and Google trends [10]. Start date being the first available date on API for each currency and end date being 4.2.2022.

Yahoo Finance data was downloaded manually for each of the cryptocurrencies, this sample provided financial attributes to the final dataset. Data from CoinGecko was downloaded by writing a Python script which queries CoinGecko API and downloads relevant community and developer data for each day separately, aggregates it, and stores it in a single CSV file for each currency. Google trends data was retrieved in the same manner as CoinGecko, its API was queried for the same period with keywords "shiba inu coin", "idena" and "dogecoin". Geolocation was set to worldwide, providing popularity stats globally.

Attributes in merged dataset are the following :

a) Obtained from Yahoo Finance
- Open (opening price for the day)
- Close (closing price for the day)
- High (highest price for the day)
- Low (lowest price for the day)
- Adj Close (adjusted closing price)
- Volume (sum total of trades for the day)

b) Obtained from CoinGecko
- facebook_likes
- twitter_followers
- reddit_average_posts_48h
- reddit_average_comments_48h
- reddit_subscribers
- reddit_accounts_active_48h
- forks (number of forks on github)
- stars (number of stars on github)
- subscribers
- total_issues (number of total issues on github)
- closed_issues (number of closed issues on github)
- pull_requests_merged
- pull_request_contributors
- commit_count_4_weeks

c) Obtained from Google trends
- search term unscaled (dogecoin, shiba inu, idena, score 0-100)
- search term monthly
- isPartial
- scale
- search term

During the explorative analysis phase all three separate files for each currency were merged into one with inner join, based on the sample date. After that, each of the three datasets were analyzed separately.

It was discovered that time sample was not continual for all 3 cryptocurrencies, most likely because of inner join used to merge files. Problem was resolved by propagating last valid observation forward. Missing values were filled in using linear interpolation if missing data was less than 50% of the attribute sample, otherwise attribute was discarded from the dataset. In case of Shiba Inu sample, first few days were without community and developer data so it was assumed that value of each variable was 0.

Before models were fitted additional columns were created in the dataset to store lagged values of all predictive attributes, with lag windows of 3, 7, 14, 21, 30 and 60 days. This was done for all 3 datasets. In the case of Dogecoin, total number of columns after this change was 138 including the date column, for Shiba 87, and for IDENA 94. After preprocessing, Dogecoin's dataset had 1549 entries, Shiba Inu's had 553 and IDENA's had 543.

IV. Methodology and results

ARIMAX (Autoregressive Integrated Moving Average with explanatory variables) [11] was chosen for cryptocurrency price prediction, target variable being closing price for each day. ARIMAX model hyperparameters *p, q* and *d* were chosen so that they yield lowest AIC (Akaike Information Criterion) [12] score. ARIMAX model's accuracy was evaluated with MAPE (mean average percent error).

Dogecoin data sample was split into 2 subsets one for model fitting and one for testing. Fitting subset consisted of first 90% of the data sample and testing subset was last 10%. This ratio was chosen because price fluctuations from 2017 to 2021 were insignificant compared to 100x rise that happened at the end of 2020. Model was fitted using auto arima function from Python pmdarima package, which attempts to optimize hyperparameter combination to yield lowest AIC score. In total 6 models were fitted in this way, one for each variable lag window. Table I shows test results after first test, including hyperparameter combinations and MAPE. Best results were achieved with 3 day lagged variables and model (2,0,4).

TABLE I. RESULTS FOR DOGECOIN AFTER FIRST TEST

Variable lag (days)	Test results			
	AR (p)	*I (d)*	*MA (q)*	*MAPE*
3	2	0	4	16.88 %
7	5	0	2	24.32 %
14	1	0	1	35.29 %
21	1	0	1	42.70 %
30	5	0	0	41.66 %
60	3	1	0	55.05 %

Shiba Inu sample was, at first, split in 70:30 ratio for training and testing. This however proved inadequate as Shiba Inu had 2 major price jumps, first in May 2021 and second in October of the same year. To catch the second price jump training subset was extended to 90% of the sample. Model was fitted for each lag window with auto arima. Results of the first test for Shiba Inu coin are displayed in Table II.

TABLE II. RESULTS FOR SHIBA INU AFTER FIRST TEST

Variable lag (days)	Test results			
	AR (p)	*I (d)*	*MA (q)*	*MAPE*
3	1	0	0	6.46 %
7	1	0	0	10.50 %
14	1	0	0	16.11 %
21	1	0	1	17.06 %
30	0	1	0	25.79 %
60	0	1	0	32.41 %

IDENA's sample was split in 70:30 ratio, as there are no significant price spikes in the sample. Model was fitted with auto arima as well. Results for IDENA are displayed in Table III.

TABLE III. RESULTS FOR IDENA AFTER FIRST TEST

Variable lag (days)	Test results			
	AR (p)	*I (d)*	*MA (q)*	*MAPE*
3	1	0	2	13.75 %
7	1	0	0	23.67 %
14	1	0	2	21.50 %
21	0	1	0	17.52 %
30	0	1	0	28.95 %
60	1	0	0	104.32 %

First tests performed on all three cryptocurrencies showed that ARIMAX model's accuracy was not good enough on test samples. Based on stats from tables I, II and III it can be determined that least inaccurate model was one fitted on Shiba Inu sample. This doesn't mean that it is the best model though, as training sample had to be increased to 90% because it couldn't accurately predict price spike of October 2021. In order to improve accuracy for all three cryptocurrencies, correlation was analyzed between all dataset variables by drawing heatmaps. Correlation was calculated for lagged values of independent variables and current value of dependent variable, for all lag windows. Independent variables with correlation coefficient between them and dependent variable lower than 0.6 were excluded from the dataset as their predictive value was deemed inadequate. Independent variables with very high correlation coefficient between them, higher that 0.9, were removed as well to avoid multicollinearity issues. In order to further improve accuracy for prediction of IDENA's and Dogecoin's price, instead of lagged variable values, average value and standard deviation were used for lag windows of 3 and 7 days. This means that for each variable four additional columns were created to store their standard deviation and average value for both lag windows. Although this approach significantly improved accuracy for both currencies, it also limited prediction to one day ahead as next day's price is predicted based on calculated averages and standard deviation for the past *n* days.

After this, models were fitted and tested again on the same samples as before, this time only with 3 and 7 day lags as those proved most accurate. Results after this test are shown in Table IV.

TABLE IV. RESULTS FOR ALL CRYPTOCURRENCIES AFTER SECOND TEST

Currency	Test results				
	Lag (days)	*AR (p)*	*I (d)*	*MA (q)*	*MAPE*
Shiba Inu	3	1	0	0	7.15 %
IDENA[1]	3	1	0	0	7.46 %
Dogecoin	3	4	0	2	7.12 %
Shia Inu	7	1	0	0	11.35 %
IDENA[1]	7	2	0	1	12.38 %
Dogecoin	7	4	0	3	8.64 %
Dogecoin[1]	3	1	0	0	6.21 %
Dogecoin[1]	7	2	0	3	12.73 %

[1] Model for prediction one day ahead

Model with best generalization seems to be one for prediction of IDENA's price one day ahead with 3 day lag window and parameters (1,0,0), as it performed very well even with 30% of data sample being in test subset (Fig. 1).

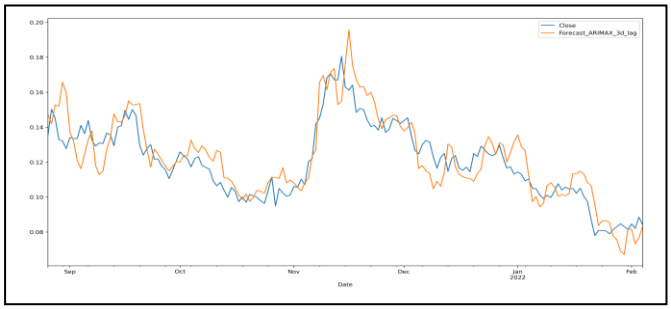

Fig. 1. Predicted and actual price 1 day ahead for IDENA

V. Conclusion

ARIMAX model was used to predict prices of 3 cryptocurrencies: Shiba Inu, Dogecoin and IDENA.

Model proved to be effective in short term predictions 1-7 days ahead, especially if exogenous variables are used with their average value and standard deviation. When correlation is analyzed, lag must be taken into account. ARIMAX's effectiveness is limited when it comes to sharp price increases and decreases, even if such events occurred in the past and are in the training dataset. It also seems that the best generalized model is one for IDENA, as fluctuations in its price are limited which can be explained by the fact that IDENA's technological value is much higher than that of Shiba or Dogecoin, and as such investors tend to be more informed about their decision to invest in it.

Further improvement is possible with ARIMAX model, if ARIMAX model parameter search is narrowed by analyzing possible parameter ranges with Autocorrelation and Partial autocorrelation graphs. Another possibility would be Grid search of model parameter combinations, either with parameter ranges from aforementioned graphs or with some heuristics.

References

[1] Chohan, Usman W. "A history of Dogecoin." *Discussion Series: Notes on the 21st Century* (2021).

[2] https://shibatoken.com/

[3] https://docs.idena.io/docs/wp/summary/

[4] Azari, Amin. "Bitcoin price prediction: An ARIMA approach." *arXiv preprint arXiv:1904.05315* (2019).

[5] Scalzotto, Giovanni. "Social Media Impact on Cryptocurrencies." (2021).

[6] Peter, Ďurka, and Pastoreková Silvia. "ARIMA vs. ARIMAX–which approach is better to analyze and forecast macroeconomic time series." *Proceedings of 30th international conference mathematical methods in economics*. Vol. 2. 2012.

[7] Iqbal, Mahir, et al. "Time-series prediction of cryptocurrency market using machine learning techniques." *EAI Endorsed Transactions on Creative Technologies* (2021): e4.

[8] https://finance.yahoo.com/

[9] https://www.coingecko.com/

[10] https://trends.google.com/

[11] W. Wei. Time Series Analysis. Addison–Wesley, 1994

[12] Sakamoto, Yosiyuki, Makio Ishiguro, and Genshiro Kitagawa. "Akaike information criterion statistics." *Dordrecht, The Netherlands: D. Reidel* 81.10.5555 (1986): 26853.