Capstone Paper

Crypto Prediction Pattern Recognition Techniques for Market Analysis

Sergey Tovmasyan, Gurgen Hovakimyan

BS in Data Science

American University of Armenia

May 9, 2024

Contents

Submission Date: May 9, 2024

ABSTRACT

The appearance of cryptocurrencies in financial markets has changed how people think about digital assets and potential variants for investment. This paper aims to explore the implementation of machine learning models to predict cryptocurrency price movements. The BTCUSDT pair has been chosen for the case study. The research methodology includes a review of several papers about machine learning models being used for recognizing patterns in financial markets, the collection of trading data, the preparation of the data for analysis, and the application of ML models that capture dependencies and patterns in price movements. This work aims to contribute to the growing field of financial technology by offering a comprehensive analysis of ML models in cryptocurrency prediction and reliable forecasting tools in the digital finance landscape.

Keywords: cryptocurrency - digital currency, BTCUSDT - the pair of bitcoin and USD tokens, ML - machine learning

GitHub Repository Link

Google Drive Link

1 INTRODUCTION

In recent years cryptocurrencies have incomparably evolved as financial assets. The increasing significance of cryptocurrencies in financial markets makes people think about how to wisely invest in them. The unique characteristics of cryptocurrencies such as decentralization, transparency, and cryptographic security make them an attractive asset class for investments and speculations. These are the main differences between cryptocurrencies and traditional currencies. The peculiarities of cryptocurrencies entail their inherent volatility and inexplicable price movements in financial markets. The highly volatile market becomes a tough challenge for traders to make profitable investments. Understanding the factors influencing the crypto market is essential to avoid risks and capitalize on market trends. Therefore, accurately predicting price movements can help investors and traders to decide upon buying, selling, or holding assets. A well-developed trading strategy can result in accurate forecasts and maximized profits. The role of machine learning has grown in financial market analysis and price prediction. Machine learning consists of the development of algorithms that learn and predict based on a given dataset. ML techniques have gained popularity in the finance industry due to their ability to analyze large datasets and identify complex patterns. The main advantage of advanced ML models is the ability to adapt to changing market conditions and capture nonlinear relationships.

2 PROBLEM STATEMENT

Cryptocurrency markets have unpredictable and dynamic nature, which becomes a challenge to make informed trading decisions. Traditional methods of financial analysis may struggle to capture relationships and patterns in cryptocurrency prices. Machine learning offers a better approach to learning from historical data and identifying patterns compared to analytics by humans. The goal of this project is the exploration and analysis of various ML models for recognizing patterns and predicting cryptocurrency price movements. For this study, the pair of BTC and USDT has been chosen. It represents the trading pair between Bitcoin and Tether. Bitcoin is the first and the best-known cryptocurrency. It is the most worthy coin among all digital coins and has a high volatility. Tether is the token version of the US dollar in the crypto world. It is a stablecoin pegged to the value of the USD and is mostly used by traders as a safe coin during market volatility. The BTCUSDT pair has sufficient liquidity and trading volume, making it a good fit for analysis and price movements. By focusing on this BTCUSDT pair, the study provides useful insights into the patterns of price movements. Figure 1 represents the price of BTCUSDT over time used for the research. Figure 2 represents the traded volume of the pair in the time frame between January 2023 and February 2024.

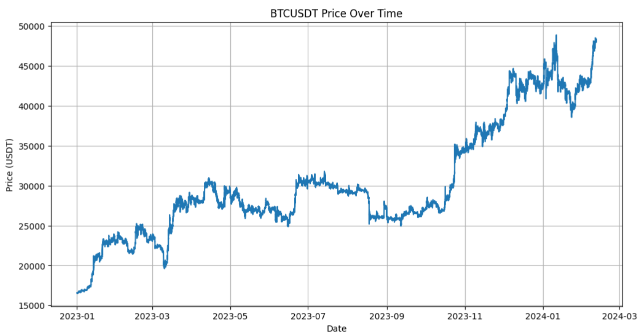


Fig. 1. BTCUSDT pair price since January 2023.

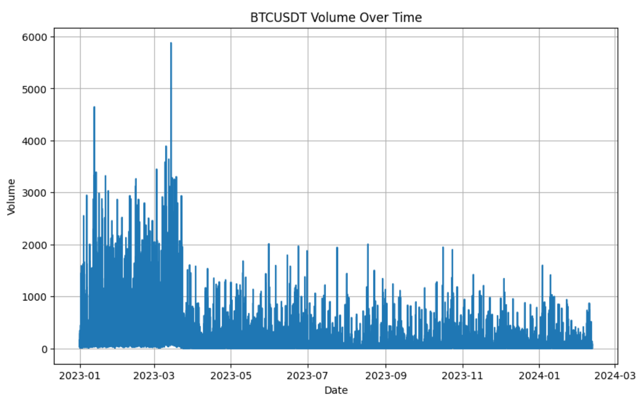


Fig. 2. BTCUSDT traded volume over time

3 LITERATURE REVIEW

Before starting the experiments a systematic literature review has been conducted. Several research papers have been studied about neural networks, machine learning, and statistical models being used to predict prices and directions of the financial market. The review was mostly based on relatively new studies. It helped to better understand the ongoing market, price fluctuations, and expected outcomes. Papers included studies conducted on various models, such as LSTM, CNN, ARIMA, HMM, linear regression, hybrid ML models, etc. For this research study, three models have been used: LSTM, CNN-LSTM hybrid model, and ARIMA. All of the models used for this study were taken from existing articles. Some changes have been implemented on them.

4 DATA COLLECTION

The data used in this study was obtained from the largest cryptocurrency exchange platform Binance via API. The dataset consists of one-minute price data of the BTCUSDT pair from January 1st, 2023 until February 12th, 2024. The one-minute level resolution of the data allows us to identify potential patterns and examine time-related trends of the chosen pair.

5 DATA PREPROCESSING

After obtaining the dataset some modifications occurred. Initially, the dataset contained five columns: open, high, low, and close prices of the pair and the traded volume in one minute. One of the modifications of the dataset is the addition of the columns. The first new column represents the change in price compared to the previous price one minute ago. The second column shows the trend: whether the price change was negative or positive based on the values of the previous column. Then, the mean of the ‘close’ price and the standard deviation were calculated. Those values were used for normalizing the data. The dataset was normalized, making the mean equal to 0 and the standard deviation to 1. To find out how much was the loss, the square root of the value of the loss needs to be multiplied by the standard deviation. Normalization helped to work with the dataset more comfortably and faster. Initially, the dataset was divided into three parts: training, validation, and testing datasets. The model was trained on 70% of the dataset, while validation and testing phases were implemented on 15%. In this study, the training, validation, and testing of the models were implemented with consecutive and random data. As the price of bitcoin is extremely volatile, running the models on random data gave better results for the majority of cases. To predict future prices, we trained the models using only ‘close’ values of each minute of the dataset. The expected output of the models is the prices going either up or down. Based on the outputs(up/down) of the model the accuracy is calculated. During the training, validation, and testing process, the loss is calculated using mean squared error(MSE). It measures the amount of error for each data point and then takes the mean value of errors.

6 METHODS

6.1 Long Short-Term Memory (LSTM)

LSTM is a type of Recurrent Neural Network (RNN) that can detain long-term dependencies in sequential data. The LSTM network consists of a series of cells, each of which has a set of gates: input, output, and forget gates. They control the flow of information into and out of the cell. The gates forget or retain information from the previous time steps, allowing the LSTM to maintain long-term dependencies in the input data [2].

6.2 CNN-LSTM hybrid model

This is the hybrid version of Convolutional Neural Networks (CNN) and LSTM models. It combines both advantages of CNN and LSTM autoencoder in feature extraction and the ability to learn patterns in data over long sequences, where CNN network layers extract the main spatial features of the time series window, and then the LSTM layers learn the time series gradient and the dependencies of long-range in time series [8].

6.3 ARIMA

An autoregressive integrated moving average (ARIMA) is a statistical analysis model that uses time series data to either better understand the dataset or to predict future trends [6]. An ARIMA consists of:

- Autoregression (AR): a model that shows a changing variable that regresses on its own lagged or prior values.
- Integrated (I): distinguishes raw observations to allow the time series to become stationary.
- Moving average (MA): incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

7 RESULTS

7.1 LSTM

The testing results of the LSTM model on consecutive data are shown in Table 1 below:

Loss	Accuracy (in %)
0.27933576093478635	49.04012570595737
0.2621366649188779	50.95987429404263
0.2500248937444253	50.95987429404263

Table 1. LSTM results on a consecutive data

It is visible in the table that no matter how much the loss was decreasing the accuracy stayed the same after the second row. The highest accuracy is around 51%. The interpretation in the table is visualized in Figure 3 below. It shows training, validation, and testing accuracy and loss values over epochs.

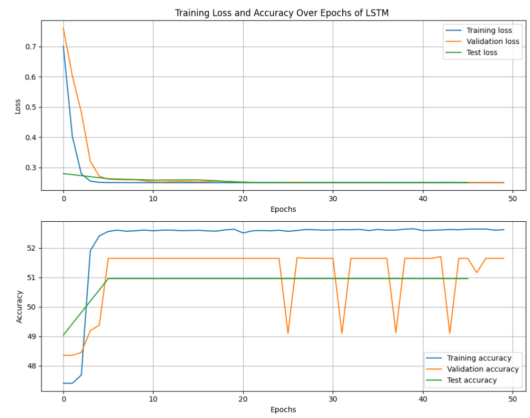


Fig. 3. Loss and accuracy on consecutive data over epochs

The LSTM results for random data are represented in Table 2 and Figure 4 below:

Loss	Accuracy (in %)
0.25017068535089493	52.10762433958826
0.24992500397969375	52.10762433958826
0.24954524179073898	52.10762433958826

Table 2. LSTM results on a random data

In this case, accuracy is higher, about 52.1% for all the cases. Unlike the changing numbers of loss, the accuracy remains the same. It is seen from the graph that the accuracy and the loss do not change after the 20th epoch.

7.2 CNN-LSTM

The testing results of the CNN-LSTM model on consecutive data are represented in Table 3 and Figure 5 below:

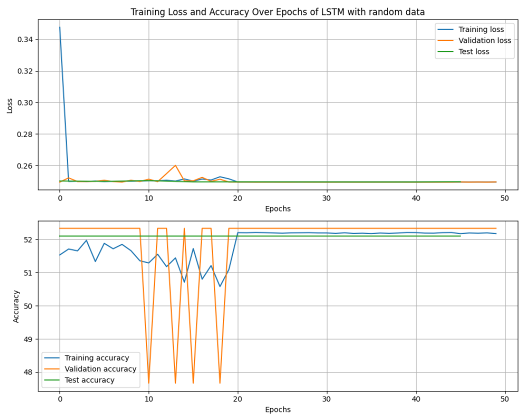


Fig. 4. Loss and accuracy on random data over epochs

Loss	Accuracy (in %)
0.375005300749431951	50.95987429404263
0.250893300737847	49.04012570595737
0.2925010167739608	50.95987429404263

Table 3. CNN-LSTM results on consecutive data

The highest accuracy for this case was around 50,9%. After a certain epoch, the results did not change.

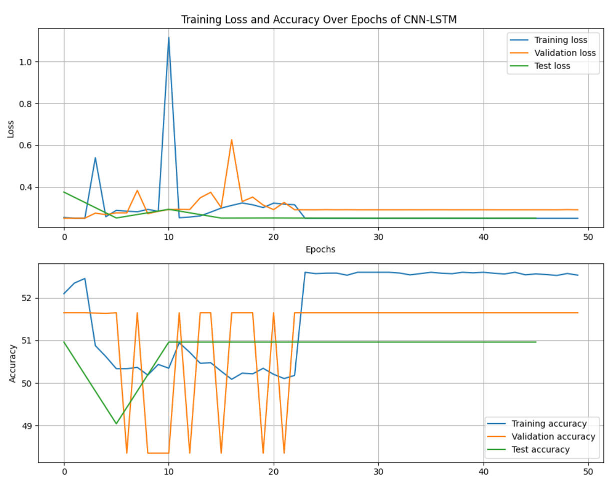


Fig. 5. Loss and accuracy on consecutive data over epochs

The CNN-LSTM results on random data are represented in the above Table 4 and Figure 6:

This model also gave an accuracy of around 52.1%. The test loss and accuracy did not change at all after the 15th epoch.

Loss	Accuracy (in %)
0.28579187427054753	52.10762433958826
0.2782266034998677	52.10762433958826
0.28414344211870973	47.89237566041174

Table 4. CNN-LSTM results on a random data

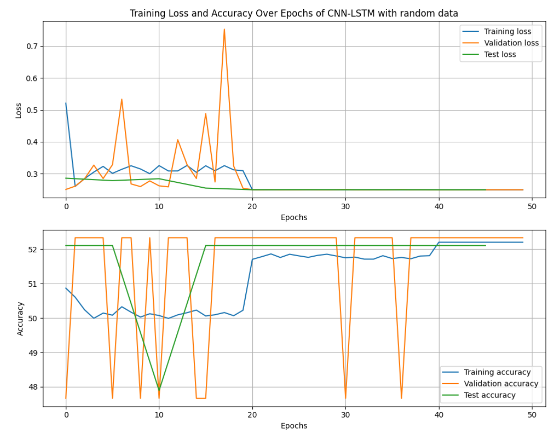


Fig. 6. Loss and accuracy on random data over epochs

7.3 *ARIMA*

The autocorrelation function (ACF) shown in Figure 11 and partial autocorrelation function (PACF) shown in Figure 8 help find the values for p and q parameters. ACF measures the linear relationship between an observation and its previous observations at different lags. PACF measures the direct linear relationship between an observation and its previous observations at a specific lag.

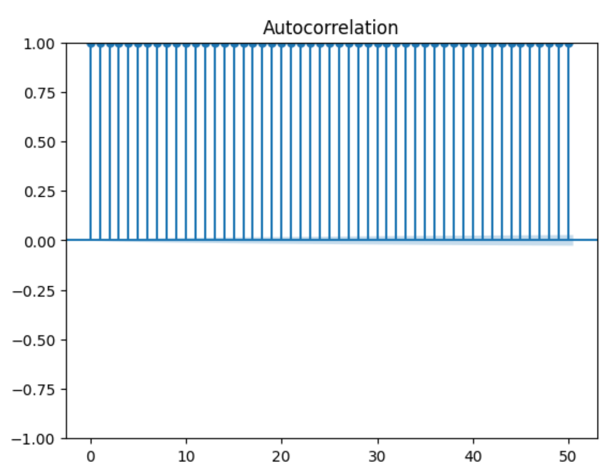


Fig. 7. ACF

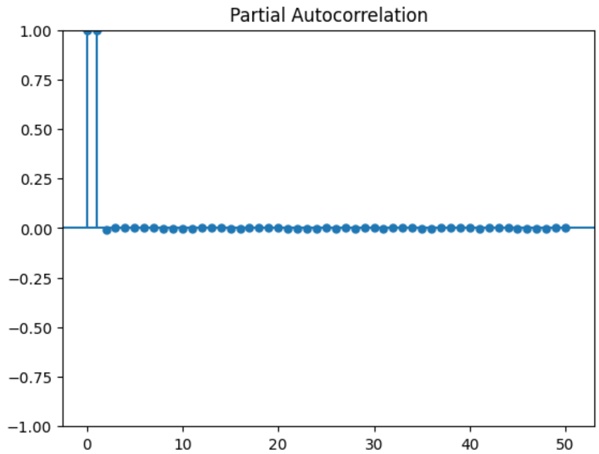


Fig. 8. PACF

The training and testing predictions of the ARIMA model are represented in two graphs. The first graph shown in Figure 9 represents the training and actual data. Figure 10 represents the prediction vs actual data. From the (p,d,q) parameters of the ARIMA model, (1,0,0) gave the best result.

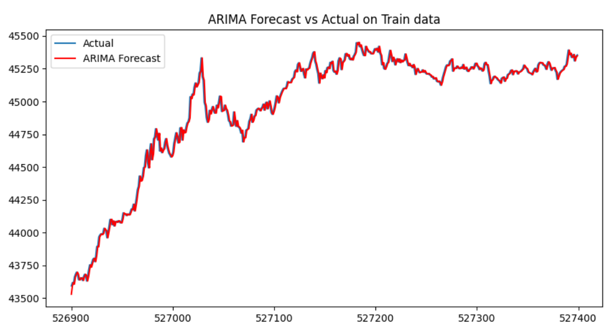


Fig. 9. Training & actual data

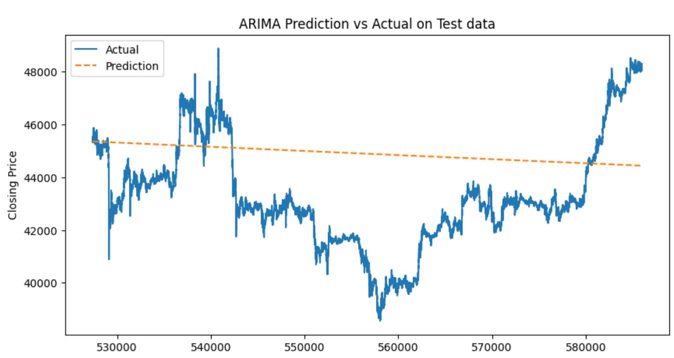
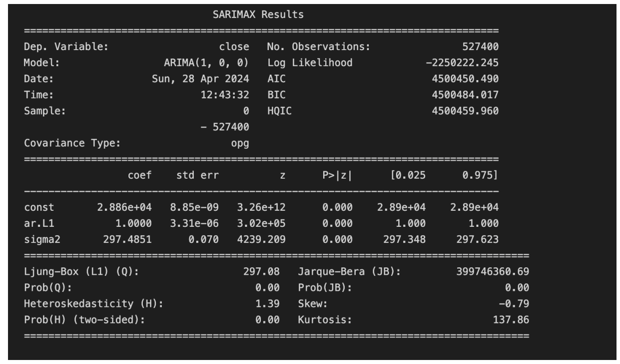


Fig. 10. Prediction

The results of the ARIMA model are shown in Figure 11 below:



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  close   No. Observations:               527400
Model:                 ARIMA(1, 0, 0)   Log Likelihood            -2250222.245
Date:                Sun, 28 Apr 2024   AIC                        4500450.490
Time:                        12:43:32   BIC                        4500484.017
Sample:                             0   HQIC                       4500459.960
                             - 527400
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       2.886e+04   8.85e-09   3.26e+12      0.000    2.89e+04    2.89e+04
ar.L1          1.0000   3.31e-06   3.02e+05      0.000       1.000       1.000
sigma2       297.4851      0.070   4239.209      0.000     297.348     297.623
===================================================================================
Ljung-Box (L1) (Q):                 297.08   Jarque-Bera (JB):         399746360.69
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.39   Skew:                            -0.79
Prob(H) (two-sided):                  0.00   Kurtosis:                       137.86
===================================================================================
```

Fig. 11. ARIMA results

8 CONCLUSION AND FUTURE WORK

In conclusion, the study shows the implementation and the results of LSTM, CNN-LSTM, and ARIMA models on the BTCUSDT pair. Each model performed differently, giving several results. The dataset is extremely noisy and predicting the trend is quite a challenge. LSTM gave the best results with the lowest loss and the highest accuracy. CNN-LSTM model had almost the same accuracy as LSTM, but the loss was higher. Both models performed better on a random data. ARIMA model did not perform well as it was overfitting every time.

There are some possible improvements to make such as;

- Enlarging the dataset to see more fluctuations and recognize complex patterns
- Use Python libraries to get live data and make predictions in a real environment
- Implement more models and experiment with various parameters.

REFERENCES

[1] Saúl Alonso-Monsalve, Andrés L Suárez-Cetrulo, Alejandro Cervantes, and David Quintana. Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators. *Expert Systems with Applications*, 149:113250, 2020.

[2] Mayank Banoula. Introduction to long short-term memory(lstm). *Simplilearn*, 2023.

[3] Sion Chakrabarti. Cryptocurrency price prediction using arima model. *Analytics Vidhya*, 2022.

[4] Prajjwal Chauhan. Stock prediction and forecasting using lstm(long-short-term-memory). *Medium*, 2023.

[5] Yuqi Ding. Enhancing stock price prediction method based on cnn-lstm hybrid model. *Highlights in Business, Economics and Management*, 21:774–781, 2023.

[6] Adam Hayes. Autoregressive integrated moving average (arima) prediction model. *Investopedia*, 2024.

[7] Patrick Jaquart, Sven Köpke, and Christof Weinhardt. Machine learning for cryptocurrency market prediction and trading. *The Journal of Finance and Data Science*, 8:331–352, 2022.

[8] Jixiang Lu, Qipei Zhang, Zhihong Yang, and Mengfu Tu. A hybrid model based on convolutional neural network and long short-term memory for short-term load forecasting. In *2019 IEEE power & energy society general meeting (PESGM)*, pages 1–5. IEEE, 2019.

[9] Benjamin J Mckinnon. I modelled bitcoin prices using arima, so you don't have to. *Medium*, 2022.

[10] Meet Nagadia. Bitcoin price prediction using lstm. *Kaggle*, 2022.

[11] Azeez A Oyedele, Anuoluwapo O Ajayi, Lukumon O Oyedele, Sururah A Bello, and Kudirat O Jimoh. Performance evaluation of deep learning and boosted trees for cryptocurrency closing price prediction. *Expert Systems with Applications*, 213:119233, 2023.

[12] Mimmo Parente, Luca Rizzuti, and Mario Trerotola. A profitable trading algorithm for cryptocurrencies using a neural network model. *Expert Systems with Applications*, 238:121806, 2024.

[13] Navid Parvini, Mahsa Abdollahi, Sattar Seifollahi, and Davood Ahmadian. Forecasting bitcoin returns with long short-term memory networks and wavelet decomposition: A comparison of several market determinants. *Applied Soft Computing*, 121:108707, 2022.

[14] Helder Sebastião and Pedro Godinho. Forecasting and trading cryptocurrencies with machine learning under changing market conditions. *Financial Innovation*, 7:1–30, 2021.

[15] Priya Singh, Manoj Jha, Mohamed Sharaf, Mohammed A Elmeligy, and Thippa Reddy Gadekallu. Harnessing a hybrid cnn-lstm model for portfolio performance: A case study on stock selection and optimization. *IEEE Access*, 2023.

[16] Yijun Wang, Galina Andreeva, and Belen Martin-Barragan. Machine learning approaches to forecasting cryptocurrency volatility: Considering internal and external determinants. *International Review of Financial Analysis*, 90:102914, 2023.

[17] Yingjian Zhang. Prediction of financial time series with hidden markov models. 2004.