Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

MACHINE LEARNING-BASED ANALYSIS AND FINANCIAL RISK MANAGEMENT IN CRYPTOCURRENCY MARKET

[#1]GUNDA SAI NIKHIL,

[#2]B.ANVESH KUMAR, ***Assistant Professor,***

[#3]Dr.V.BAPUJI, ***Associate Professor& HOD,***

Department of Master of Computer Applications,

VAAGESWARI COLLEGE OF ENGINEERING, KARIMNAGAR,TELANGANA

ABSTRACT: For daily bitcoin market forecasting and trading, we deploy and analyze a range of machine learning algorithms. The algorithms have been trained to forecast the binary relative daily price movements of the top 100 cryptocurrencies. All of the models we tested produced statistically plausible estimates, with average accuracy values ranging from 52.9% to 54.1% across all cryptocurrencies. Based on the subset of predictions with the 10% greatest model confidences per class and day, these accuracy results range from 57.5% to 59.5%. We find that after transaction costs, a long-short portfolio strategy based on the forecasts of the deployed LSTM and GRU ensemble models yields annualized out-of-sample Sharpe ratios of 3.23 and 3.12, respectively. In comparison, the benchmark buy-and-hold market portfolio strategy has a Sharpe ratio of only 1.33. These findings point to a threat to the efficiency of the bitcoin market, albeit the impact of certain arbitrage constraints cannot be completely ruled out.

Keywords: Financial market prediction; Market efficiency; Statistical arbitrage; Machine learning; GRU; LSTM; Neural network; Random forest; Gradient boosting; Temporal convolutional neural network

1. INTRODUCTION

In 2008, Nakamoto1 officially introduced the Bitcoin peer-to-peer currency system. Since then, numerous other cryptocurrencies have been developed, each with its own set of technological characteristics and possible uses, all of which can trace their origins back to Bitcoin. As the cryptocurrency market has grown exponentially during the past decade, individual digital currencies' prices have fluctuated widely.There isn't enough space in the user's text to rewrite it scholarly. Different market participants have different opinions on whether or not Bitcoin and similar cryptocurrencies are effective.The user entered a three-four-five number sequence. Auto-regressive statistical approaches, which explicitly represent any non-linear interactions, are frequently used in such investigations. Due to their ability to understand the malleable functional relationship between features and targets, machine learning algorithms have been successfully used in the past to forecast the cryptocurrency market, including Bitcoin.Seven, eight, and nine are mentioned. Thus, these methods can detect and capitalize on intricate linkages among several variables in high-dimensional areas, including but not limited to those not specifically discussed in studies of market performance. The purpose of this research was to compare and contrast the performance of various machine learning models for use in financial market prediction. Accordingly, the primary inquiry driving this work is as follows: Can statistical arbitrage be effectively established by machine learning algorithms in the bitcoin market?

We use six machine learning classifiers to forecast the daily relative performance of the top 100 cryptocurrencies by market capitalization in order to answer this research question. In addition, we use each model's out-of-sample estimates as the basis for a long-short trading strategy. After that, we take a look at how every deal turned out. The examination spans a total of 800 days, 400 of which are spent analyzing each of the five time periods. The two most important things that this study adds are:

To begin, it's crucial to emphasize machine learning's predictive capacity in regards to the cryptocurrency market, as all employed models produce estimates that are statistically reliable. As a result, we may infer that recurrent neural networks and tree-based ensembles are the most effective methods for comparing the day-to-day prices of various cryptocurrencies. Even after factoring in transaction fees, the results of the long-short portfolio strategy show that it outperforms the market benchmark. As a result, it appears that the Bitcoin market may offer statistical arbitrage opportunities.

The article's remaining sections will follow this format: Our results are presented in Chapter 4, while our methodology is discussed in Chapter 3, and the corresponding literature is offered in Chapter 2. The ramifications of these findings are explored in Chapter 5, and a final analysis of this investigation is presented

2. RELATED WORK

In order to determine whether or not machine learning forecasts may be used to facilitate statistical arbitrage in the cryptocurrency market, Fischer et al. (2018) analyze data collected from June to September of that year. The study uses a random forest classifier and a logistic regression model to forecast the relative performance of the top 40 cryptocurrencies over the next 120 minutes using the temporal distribution of historical returns seen over the previous day. The scientists' findings may lead to a decrease in the efficiency of the bitcoin market, since an out-of-sample long-short trading strategy employing these model forecasts provided a daily return of 7.1 basis points. Fil and Kristoufek (2010) extend the idea of pairs trading to the cryptocurrency market on the assumption that many cryptocurrency pairs will exhibit long-term stability. The authors of this paper analyze trade activity at 5-minute, hourly, and daily intervals from January 2018 through September 2019. According to Fil and Kristoufek (2018), pairs trading has shown promise in high-frequency trades on the cryptocurrency market. But keep in mind that the outcomes of such a trading strategy are highly sensitive to the exact market factors used, such as the degree of transaction fees.

Betancourt and Chen (2011) evaluated deep reinforcement learning for bitcoin trading using data from August 2017 to November 2020. The suggested system's agents regularly review 20 days' worth of data on a cryptocurrency's price, volume, and market capitalization to make trading decisions within a day's time. Betancourt and Chen's technique (2011, IEEE) has the potential to simplify Bitcoin transactions. McNally et al. (2012) assessed the efficacy of three methods for predicting daily binary Bitcoin market movements. Several other methods were employed, including Elman recurrent neural networks, extended short-term neural networks, and autoregressive integrated moving average. Based on an analysis of data from August 2013 through July 2016, the extended short-term neural network outperforms its competitors with a model accuracy of 52.78 percent. Several neural network algorithms have been developed to predict Bitcoin's daily value, and Dutta et al. (2013) examine these algorithms using a wide variety of technical, blockchain-based, asset-based, and interest-based variables. According to their analysis of data from January 2010 through June 2019, the best outcomes can be achieved by employing a gated recurrent unit with recurrent dropout.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

Chen et al. (2014) employed a combination of linear statistical approaches and machine learning strategies to forecast the 5-minute and daily Bitcoin markets. Data from February 2017 to February 2019 are used for this analysis. In comparison to machine learning approaches, statistical methods have been found to be more effective when making daily forecasts. Alessandretti et al. (2015) use gradient boosting classifiers and extended short-term neural network techniques to forecast the daily returns of 1681 cryptocurrencies. The authors provide empirical evidence from November 2015 through April 2018 to support their argument that portfolio strategies based on these forecasts outperform a baseline approach. Using a neural network trained with long short-term memory and another trained with extended regression approaches, Lahmiri and Bekiros (2016) compared their respective efficacies. The goal of their research was to forecast the price movements of digital currencies like Bitcoin, Digital Cash, and Ripple. The time periods covered by the data sets used by the researchers extended from the distant past to as recently as October of this year. Based on their findings, a neural network that is taught to prioritize short-term memory outperforms one that is taught to prioritize generalized regression.

3. METHODOLOGY

Following the structure established by Fischer et al.8, Fischer Krauss17, the study is broken down into four parts. In the first stage, relevant information is compiled from a number of different sources. To model future coin returns, we first take the original price data to create characteristics and targets. Backtesting can be performed by splitting the full data set into test folds that do not overlap and study folds that do overlap with different market factors. After training and independently tweaking each of the deployed models for each time period, the final step is to simulate trading based on model predictions.

Data

The analysis relied on the daily close price and market capitalization data obtained via the CoinGecko (CG) API. The USD values in this data collection span the time period from February 8, 2018 through May 15, 2022.

Coin market capitalization data

To mitigate survivorship bias, we restrict our analysis of the investing universe to the top 100 cryptoassets by market capitalization on the first trading day of the training set. This technique ensures that each coin gets sufficient training data and mitigates "look-ahead bias" in the coin universe generation process. Since the USD value of stablecoins pegged to the USD or another fiat currency is either predetermined or entirely dependent on exchange rates, they do not qualify. Due to data issues, such as missing data or incorrect values in the data sources, ten additional coins have been removed from the study. The complete list of prohibited coins is provided in Appendix B.1.

The CG API provides daily USD market capitalization statistics for the top 1750 coins as of June 8, 2022. This is calculated using the asset's price and the amount that is known. All cryptocurrencies are rated based on their training set day one market cap during each research period. This is useful for assembling crypto asset portfolios.

Coin price data

The return is calculated using CoinGecko's data on market prices. The CG platform provides prices that are the mean of all possible pairings of cryptocurrencies and fiat currency or cryptocurrencies and cryptocurrencies that are accessible on all monitored exchanges. The volume of transaction is factored into these pricing. Even if the given prices are totals of prices that aren't sold, the value of 19 indicates that these fictitious prices accurately reflect the current situation of the crypto-currency business. According to the author's findings, the efficiency of the liquid cryptocurrency market is not affected by the addition of prices from other exchange sites. Due to the 24-hour nature of bitcoin exchanges, the market price at midnight (UTC) is sometimes utilized to create inflated artificial closing prices. Every day at midnight UTC, the CG API updates with the day's price data. Results for the previous day can be calculated by including an extra day in the time series of daily quotes. Returns are calculated using the market price of coin c at the end of day t, which is denoted as rm,c. The metric is taken on day t+1 at midnight UTC. How much value has been added to Coin C over the previous m days is represented by the sum of its closing price on day t.

The asset's daily earnings at m = 1 can be found using this formula. When m is greater than one, however, rm,c displays the cumulative returns over the previous m days. 20% is the risk-free rate of return.

Excess returns are calculated using the rate on the secondary market for the three-month Treasury Bill (T-bill) issued by the United States Treasury. T-bills are a type of short-term loan guaranteed by the United States Treasury.

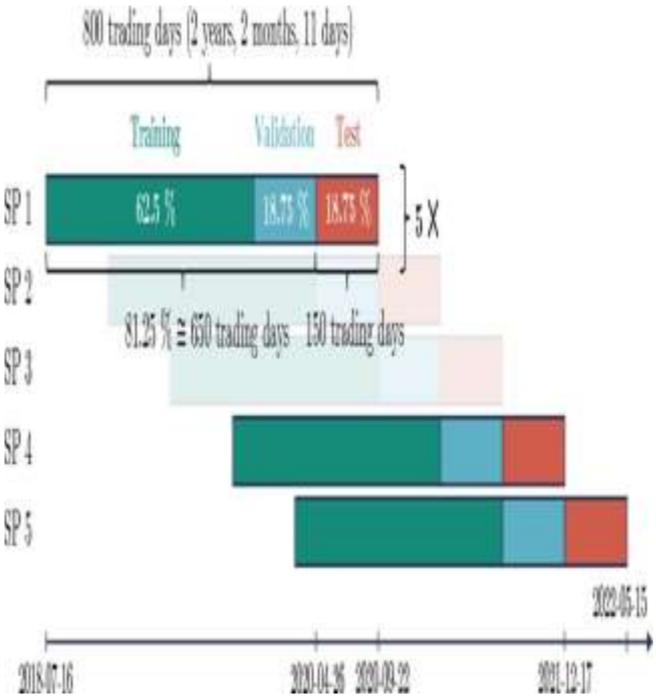

Fig. 1. Study period composition and train-validation-test split.

Three months is the state government's maturity period. It is important to deannualize the annual interest rate into daily returns in order to calculate risk-adjusted return measures such as the Sharpe ratio and Sortino ratio. The risk-free rate, as demonstrated by the aforementioned T-bill rate, is consistently small throughout the investigated timeframe, ranging from 2.4 x 107 to 2.8 x 105 on a daily basis and averaging 3.9 x 106.

Software and hardware

For this project's data collection, processing, and analysis needs, Python 3.9 is used extensively. Pandas22 and numpy21 are used for data processing and feature generation, respectively. When it comes to creating and training traditional machine learning models, Scikit-learn is a popular choice of library. However, deep learning models are generated using Keras and the TensorFlow backend. Models are trained using 2.8 GHz Intel Core i5-8400 processors.

Data split

The five study periods (SPs) of this research project, also known as the prediction targets, have a total of 800 trading days. Each model used in the forecasting process uses data from the prior three months as inputs. Therefore, the SP for each asset is calculated using data from the 90 days prior to the first trading day. Figure 1 depicts the study period as a triangular structure comprised of a training set, a validation set, and an out-of-sample test set. For model training, we use a data collection that covers 500 days. When adjusting hyperparameters, we use the 150-day validation set. Finally, the performance of the model is assessed using an out-of-sample test set that consists of 150 days. The allocation of each study period across the three levels of data is shown in detail in Table 1.

The training and validation stages make up the formation stage. The models are trained during the training phase, and the optimum hyperparameters are determined after validation using the results of the validation procedure. During each study session, the testing department is put to use to do simulations and real-world trade tests. Because there will be five different sets of tests given in quick succession, the length of the testing period will determine the necessary alterations to the study sessions. To account for the concept drift caused by the ebb and flow of the market,

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

researchers can retrain their models at regular intervals over the course of multiple study sessions.

Features

All of the models use the same training data, which is a binary classification issue. The goal is to forecast whether or not a given coin will have a better return than the cross-sectional mean the day after. This conclusion is based entirely on data from the prior ninety days of prices. Therefore, three months before trading, the characteristics for all models are calculated using the returns on each coin.

Table 1

Study periods and the respective date ranges for the training, validation, and test sets.

SP No	Training Set	Validation Set	Test Set
1	2018-07-16 - 2019-11-27	2019-11-28 - 2020-04-25	2020-04-26 - 2020-09-22
2	2018-12-13 - 2020-04-25	2020-04-26 - 2020-09-22	2020-09-23 - 2021-02-19
3	2019-05-12 - 2020-09-22	2020-09-23 - 2021-02-19	2021-02-20 - 2021-07-19
4	2019-10-09 - 2021-02-19	2021-02-20 - 2021-07-19	2021-07-20 - 2021-12-16
5	2020-03-07 - 2021-07-19	2021-07-20 - 2021-12-16	2021-12-17 - 2022-05-15

Classifiers with and without a memory function are used extensively throughout the research. Because of their unique characteristics, these classifiers are developed independently. Three deep learning models with their own memory and the ability to generate 90-character daily return patterns are the LSTM, the GRU, and the TCN. The standardization method involves dividing the daily mean by the training set's standard deviation. Logistic regression (LR) and tree-based classifiers both require historical data as inputs due to memory constraints. The procedure entails repeatedly assembling a set of input sequences and the target labels that best describe them. To achieve this, 90-degree loops are created, which overlap and advance the calendar by a single day. A sample set of input sequences and labels for use with a deep learning approach is provided below.

Fig. 2.

Since memory-free models such as GBC, RF, and LR cannot utilize temporal input data, we address this limitation by generating time-lagged features by averaging over increasingly more distant periods of time. In this research, we combine the findings of the previous studies by Takeuchi and Lee (26) and Krauss et al. (27). We focus on gaps with a step size of 10 days, examining them for values of m from 1 to 90. To do this, we first calculate the first 20 days' worth of daily variations, yielding a total of 27 characteristics for each sample. Results for various time intervals can be calculated using equation (1). In Figure 3, we can see how tree-based techniques and logistic regression are used to generate return features and target labels. There are 50,000 training samples, 15,000 validation samples, and 15,000 test samples generated across all coins and research periods when using both approaches.

Targets

Within a day of constructing a portfolio, the primary objective of the binary forecast challenge is to determine whether or not a single coin will perform better than the cross-sectional median. Daily returns across all coins are sorted into descending order at the end of each trading day. In this system, one is assigned to coins with values less than or equal to the cross-sectional median while zero is assigned to coins with values greater than the median. The groups into which coin c falls at instant t are as follows: The user's text "yc" lacks sufficient detail to be reconstructed as a scholarly

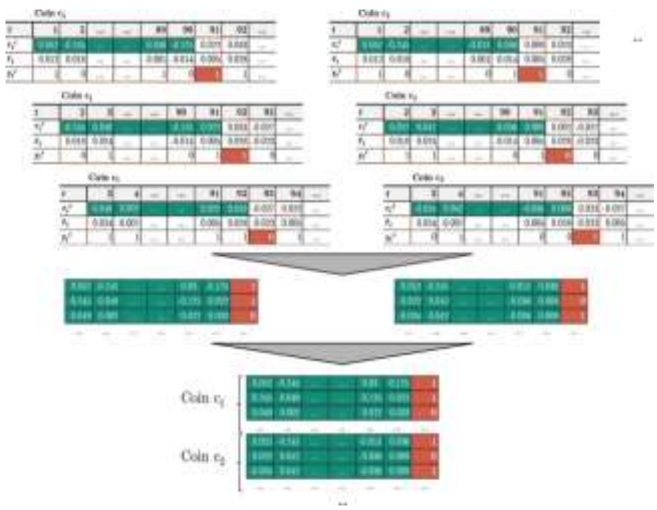

Fig. 2. When a model has storage, it can remember the sequences of features it has created and the labels it has assigned to those sequences.

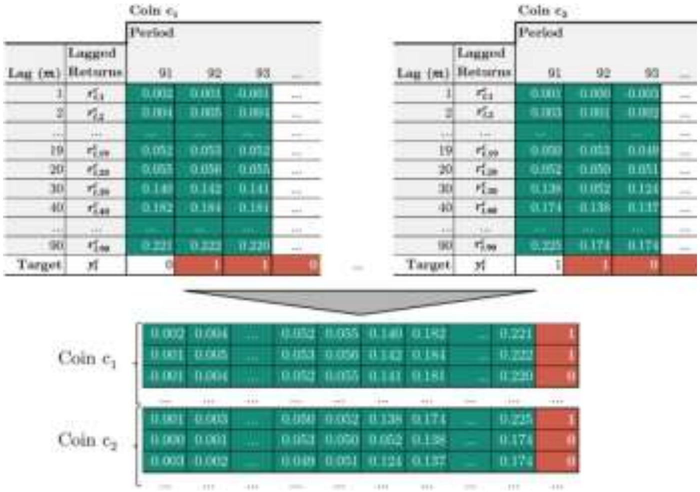

Fig. 3. The development of tree-based and logistic regression models relies heavily on the creation of feature sets and target labels.

4. CONCLUSION

In this research, we use a variety of machine learning algorithms to forecast the day-to-day price changes of the 100 most valuable cryptocurrencies. The statistical reliability of the results predicted by each model used is demonstrated here. Depending on the model used, the average accuracy for all cryptocurrencies varies from 52.9% to 54.1%. The accuracy values, which range from 57.5 percent to 59.5 percent, were derived from a sample of predictions made using the day's top ten percent highest model confidences for each class. When taking into account transaction costs, the long-short portfolio strategy that employs the LSTM and GRU ensemble models produces yearly out-of-sample Sharpe ratios of 3.23 and 3.12, respectively. The Sharpe ratio for a benchmark market portfolio that is bought and held is 1.33. This research suggests that specific constraints connected to arbitrage may have an effect on the emergence of challenges to the efficiency of the weak form in the cryptocurrency market. A final conclusion regarding the impact of these constraints, however, cannot be drawn with certainty.

REFERENCES

1. Nakamoto S. Bitcoin: a peer-to-peer electronic cash system. Working Paper URL: https://bitcoin.org/bitcoin.pdf; 2008.
2. Coinmarketcap. Coinmarketcap; 2022. URL: https://coinmarketcap.com/. Accessed July 21, 2022. https://coinmarketcap.com/.
3. Kristoufek L, Vosvrda M. Cryptocurrencies market efficiency ranking: not so straightforward. Phys Stat Mech Appl. 2019;531, 120853.
4. Le Tran V, Leirvik T. Efficiency in the markets of crypto-currencies. Finance Res Lett. 2020;35, 101382.
5. Kakinaka S, Umeno K. Cryptocurrency market efficiency in short-and long-term horizons during covid-19: an asymmetric multifractal analysis approach. Finance Res Lett. 2022;46, 102319.
6. Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning. Rev Financ Stud. 2020;33:2223–2273. https://doi.org/10.1093/rfs/hhaa009.
7. Huang JZ, Huang W, Ni J. Predicting bitcoin returns using high-dimensional technical indicators. J Finance Data Sci. 2019;5:140–155.
8. Fischer T, Krauss C, Deinert A. Statistical arbitrage in cryptocurrency markets. J Risk Financ Manag. 2019;12:31. https://doi.org/10.3390/
9. jrfm12010031.
10. Jaquart P, Dann D, Weinhardt C. Short-term bitcoin market prediction via machine learning. J Finance Data Sci. 2021;7:45–66.
11. Fil M, Kristoufek L. Pairs trading in cryptocurrency markets. IEEE Access. 2020;8:172644–172651. https://doi.org/10.1109/ ACCESS.2020.3024619.
12. Betancourt C, Chen WH. Reinforcement learning with self-attention networks for cryptocurrency trading. Appl Sci. 2021;11:7377. https:// doi.org/10.3390/app11167377.
13. McNally S, Roche J, Caton S. Predicting the price of bitcoin using machine learning. In:

Proceedings of 2018 Euromicro International Conference on Parallel, Distributed, and Network-Based Processing. 2018:339–343. https://doi.org/10.1109/PDP2018.2018.00060.

14. Dutta A, Kumar S, Basu M. A gated recurrent unit approach to bitcoin price prediction. J Risk Financ Manag. 2020;13. https://doi.org/10.3390/
15. jrfm13020023. Article 23.
16. Chen Z, Li C, Sun W. Bitcoin price prediction using machine learning: an approach to sample dimension engineering. J Comput Appl Math. 2020;365. https://doi.org/10.1016/j.cam.2019.112395. Article 112395.
17. Alessandretti L, ElBahrawy A, Aiello LM, Baronchelli A. Anticipating Cryptocurrency Prices Using Machine Learning Complexity; 2018. https://doi.org/10.1155/2018/8983590.
18. Lahmiri S, Bekiros S. Cryptocurrency forecasting with deep learning chaotic neural networks. Chaos, Solit Fractals. 2019;118:35–40.
19. Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions. Eur J Oper Res. 2018;270:654–669. https://doi.org/10.1016/j.ejor.2017.11.054.
20. CoinGecko. Methodology; 2022. https://www.coingecko.com/en/methodology. Accessed July 21, 2022.
21. Vidal-Toma´s D. Which cryptocurrency data sources should scholars use? Int Rev Financ Anal. 2022;81, 102061.
22. Board of Governors of the Federal Reserve System. 3-Month treasury bill secondary market rate. Discount Basis [DTB3]. URL: https://fred. stlouisfed.org/series/DTB3; 2022.
23. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with numpy. Nature. 2020;585:357–362. https://doi.org/10.1038/s41586- 020-2649-2.
24. McKinney W. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. 2010:56–61. https://doi.org/10.25080/Majora-92bf1922-00a. SciPy. Chollet F. keras https://keras.io/; 2015. Accessed July 21, 2022.
25. Abadi Martín, Agarwal Ashish, Paul Barham, et al. Tensorflow: Large-Scale Machine Learning on Heterogeneous Systems; 2015. https://www. tensorflow.org/. Accessed July 21, 2022. accessed.
26. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.
27. Takeuchi L, Lee YYA. Applying deep learning to enhance momentum trading strategies in stocks. 2013. Working Paper.
28. Krauss C, Do XA, Huck N. Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500. Eur J Oper Res. 2017;259:689–702.