

Cryptocurrency Price Tendency Analysis Using Conventional Statistical Model and Machine Learning Approach

Chenyang Liao^{1†}, Kai Lu^{2†}, Jiamiao Zhang^{3 *†}
Liaoc8@uci.edu, 2052378@tongji.edu.cn, * zhangjm0618@mails.jlu.edu.cn

¹School of Social Science, University of California, Irvine, Irvine, the U.S.

²School of Economics and Management, Tongji University, Shanghai, China

³School of Economics, Jilin University, Changchun, China

[†]These authors contributed equally.

Abstract—The market for cryptocurrency has thrived for more than 10 years and has experienced a drastic change. The success of cryptocurrencies was concerned and analyzed worldwide. This research discusses the way to build machine learning and statistical models to predict the future price of the cryptocurrency based on the current price. This article collects and analyzes closed prices of Bitcoins, Ethereum, BNB coin, Avalanche, and Solana from as early as 2019 to the most recent data. Based on prediction models, anticipating the future price through the current market shows a moderate available result with a high related fitness within the regression model. Through market prices, it is possible to predict the future trend of a cryptocurrency. This research further discusses possible features of the cryptocurrency market and looks for the correlation between cryptocurrencies. According to the analysis, models show a strong relationship between Bitcoins and other cryptocurrency types in long-run market performance. Besides, we also argue that it is possible to trace price changes and predict the market based on the present price. These results shed light on guiding further exploration of stimulating market through externality.

Keywords-Machine Learning; Cryptocurrency; Prediction Model.

1 INTRODUCTION

The cryptocurrency was the idea initialized in 1990 for a currency that could be untraceable and not regulated by centralized entities (e.g., banks). The story of this new market began when Satoshi Nakamoto described what was the Bitcoin blockchain network and bought bitcoin.org in 2008 and mined the first block of the bitcoin network in 2009. Bitcoin is thus the most popular and marked type of cryptocurrency, and it is increasingly treated as a legitimate means of exchange while it becomes a well-accepted currency as a payment method for many companies. In 2017, the value of one Bitcoin has experienced 20 times growth from around \$800 in January to more than \$17,000 within a year; two months later, however, its value was cut in half to around \$8,000 per Bitcoin. In the recent decade, several new cryptocurrencies have occurred and have taken up the market, like Ethereum launched in 2015 as the second most valuable type after

Bitcoin and Dogecoin, quickly developed within a few months. Types and features of the cryptocurrency vary and the correlation within them and with externality is vague, causing the sharp change in prices and thus worthwhile for researchers and investors to analyze as well as predict the market.

The idea of this paper was innovated by and connected to previous research in price prediction and the cryptocurrency market. Adebisi, Adewumi, and Charles predicted stock prices using the ARIMA models, showing a solid example of utilizing machine learning in predicting prices [1]. Ayush with his group members predicted housing prices through machine learning and Neural Networks, which intrigued us to make use of the LGTM, a type of recurrent neural network working in prediction problems [2]. Similar research on airfare price prediction with excellent examples of methods to begin a price prediction model construction by using machine learning tools is also motivating our work [3].

Price prediction research targeting the cryptocurrency brings us more effective literature and academic guidance on our models. Zheshi, Chunhong, and Wenjun employed not only machine learning, but, for the low-frequency data of Bitcoins with different dimensions, they included statistical methods to extend the applicability of their approach [4]. Jethin, Daniel, John, and Juan predicted cryptocurrency prices based on Tweet Volumes and sentiment analysis because they believed that Tweets as a media behaved a great influence on the price of cryptocurrencies, motivating us about the externality impact of this fickle market [5].

Outside impacts intrigue us with the consideration of external factors such as media as well as the economic background of cryptocurrency. Sabrina, Marina, and David glanced at initial coin offerings (ICOs) that are deeply correlated with the popular cryptocurrency, showing the emergence of cryptocurrency and its applicability in increasing employment and avoiding enterprise failure [6]. Paul discussed and analyzed the long-run performance of ICOs with external effects like size and disintermediation of the market [7]. Both researches prompted us to focus on the features of the cryptocurrency market, and come up with possible competitive traits. To better learn features of this powerful market with potential, we tend to build several prediction models using popular machine learning and supervised models to test the correlation between different cryptocurrencies. Besides, we believed that Bitcoin obtains huge market power and impact on the whole cryptocurrency market, prompting us to check the influence of Bitcoin on other currencies. Based on models, this research also compares each method horizontally to test its fitness on various cryptocurrencies. With these models, it is plain to search and discuss the market traits of cryptocurrency.

The rest of the paper is organized as follows: The Sec. II will cover the methodology we used, which includes data collection, models, and metrics. The Sec. III will illustrate the results of the models and discuss the limitations of this research.

2 METHODOLOGY

2.1 Data

Five datasets from the Yahoo Finance transaction database of cryptocurrencies are used in this paper. They are daily transaction records for Bitcoin (BTC), Ethereum (ETH), Binance Coin (BNB), Avalanche (AVAX) and Solana (SOL). For each cryptocurrency, there are ten columns:

date, open price, highest price, lowest price, close price, volume, quote asset volume, number of trades, take buy base volume, and take buy quote volume. In the prediction models, date, close price and volume are mainly used as independent variables, along with the highest price and lowest price of the day.

The dataset records transactions of BTC, ETH and BNB from 2019/7/30 to 2022/5/7, which is 1011 records for each. Although ending on the same day, AVAX records started at 2020/9/22, SOL started at 2020/8/11, which are 593 and 635 records for them. This makes 4261 records in total. The first eighty percent records of each coin will be set as train sets, leaving the later twenty as test sets.

The reason these five cryptocurrencies were chosen is based on a natural delamination within the market. Bitcoin is definitely the unshakable first class of cryptocurrencies. Used to represent 41.5% of the entire crypto-market capitalization, Bitcoin has grown the percentage to 44.76% (May, 22nd) since the significant market recession from May 5th. Although ETH has been struggling to transform to ETH 2.0, it is expected to update proof-on-stake in Q3 2022. Capturing 19.37% of cryptocurrency market share, leading 13% from the one behind, ETH holds a stable tier two place. BNB, SOL, and AVAX are issued by transaction platforms and blockchain service platforms, and could represent a solid tier three coin and approximately 6.2% of the cryptocurrency market share, which the authors believe is the last confidential tier of the market. These five coins show different characteristics and are able to represent their own tiers.

2.2 Models

There are many different ways for people to perform price prediction. Traditional statistical ways are used widely, and machine learning has become increasingly popular with scholars recently. Especially in the field of stock price, people are racking their brains to come up with all kinds of predictions, including different statistical or mathematical ways and machine learning methods (e.g., Moving Average, Time Sequence Analysis, and PCA) and try to anticipate a better result. The cryptocurrency market has some similarities to the stock market. Scholars have done quite a bit of research on cryptocurrency prices, and interestingly a lot of it is about correlation analysis between sentiment and bitcoin. There are many articles on cryptocurrency prediction, using both traditional and machine learning methods to study them.

A number of scholars have examined the correlation between cryptocurrency prices and media sentiment. Traditional methodologies are widely used to analyze the cryptocurrency market and make predictions. A popular research direction is to use traditional statistical methods to predict cryptocurrency prices through news and social media sentiment [8]. Contemporarily, more scholars have experimented with deep learning methods to build models and improve accuracy [9]. Machine learning methods have also been used, such as using NLTK sentiment analysis using pre-trained VADER models to determine the relationship between cryptocurrency price and Tweet Volumes and Sentiment [5].

Due to the peculiarity of the cryptocurrency market, people, including but not limited to data scientists, prefer to use machine learning to predict its price movements. We have searched for information and found that the current cryptocurrency price forecasts are mainly based on the following methods.

Most Data Scientists choose to use LSTM to perform the prediction of cryptocurrency prices. The LSTM model is so popular because of its powerful capabilities. It is always able to win all kinds of competitions in Kaggle. As a widely-used method, it has its advantages:

- Long time memory function
- Good performance in sequence modeling

LSTM has been criticized because it has the following disadvantages:

- Need a considerable number of data to train a valid model
- Poor adaptation to sudden increases and decreases in data

The basic logic of the LSTM model is shown in Figure 1.

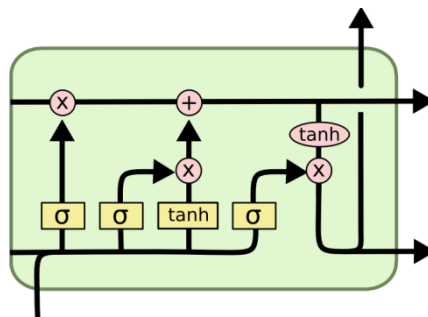


Figure 1. A sketch of LSTM [<https://colah.github.io/images/post-covers/lstm.png>]

ARIMA models, as illustrated in Figure 2, are denoted with the notation ARIMA(p, d, q). These parameters account for seasonality, trend, and noise in datasets. Here, p is the number of lag observations to include in the model, or lag order (AR); d is the number of times that the raw observations are different, or the degree of differencing (I); q is the size of the moving average window, also called the order of moving average (MA). The ARIMA model can be upgraded to a SARIMAX model. Compared with the basic ARIMA model, SARIMAX has two distinct advantages [10]:

- A seasonal component is introduced to cope with the weekly or monthly effects of price fluctuations.
- Exogenous variables can be taken into account.

Such time series models are widely used in many situations, such as predicting stock prices and prices of physical goods.

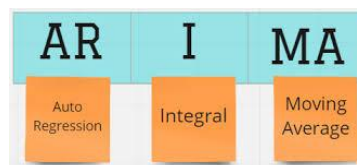


Figure 2. ARIMA Components [https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcT8wLkR1_apd2SXXZrCocsJCpemk2c1-tKN_g&usqp=CAU]

XGBoost is a decision tree model that is applicable to both classification and regression problems, suitable for application to a variety of prediction models. The architecture of XGBoost is demonstrated in Figure 3. It's widely used because of its high precision and good accuracy. Nevertheless, the XGBoost model is sometimes inferior to the LGBM model in terms of accuracy and running speed. Hence, there are also some people using the LGBM method to build a cryptocurrency price prediction model. The LGBM(LightGBM) model, the structure of which is shown in Figure 4, is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages:

- Faster training speed and higher efficiency than XGBoost.
- lower memory usage.
- Better accuracy.
- Support of parallel, distributed, and GPU learning.
- Capable of handling large-scale data.
- However, there are still some problems while using LGBM methods, though:
- A small number of data sets are prone to overfitting.
- lightGBM is a bias-based algorithm, so it is sensitive to anomalies.

In this case, cryptocurrency markets can generate anomalies (e.g., spikes and drops) due to the lack of restrictions, which can greatly affect model efficiency. Both XGBoost and LightGBM use the decision tree approach, and the machine learning principles are basically the same. On balance, we decided to use the first three methods for prediction and model comparison.

Underneath are simple diagrams of these four methods.

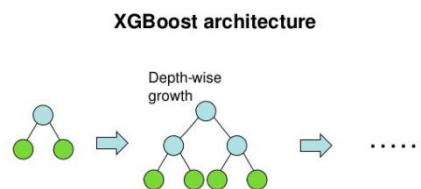


Figure 3. XGBoost Architecture [https://miro.medium.com/max/1294/1*ozf-ftCx-jy2jII4cEv9YA.png]

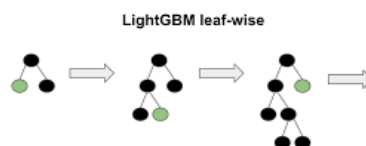


Figure 4. The structure of LGBM [<https://datascience.eu/machine-learning/1-what-is-light-gbm/>]

2.3 Metrics

Goodness of fit and root mean square error are used to measure performance of model predictions. We use adjusted r-square to estimate the relationship between movements of model predicted cryptocurrency prices based on the actual close prices' movements.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2} \quad (1)$$

$$Adj R^2 = 1 - \left[\frac{(1-R^2)(n-1)}{(n-k-1)} \right] \quad (2)$$

Where \hat{y}_i is the predict price, \bar{y} is the mean value of the actual price. R^2 has the range of [0,1]. We use Root Mean Square Error (RMSE) to measure the standard deviation of the prediction errors. This would tell us the goodness of how concentrated the data is around the line of the best fit.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

3 RESULTS & DISCUSSION

3.1 Regression analysis

Overall, LSTM has relatively the best results on metrics among the three models. The adjusted R^2 and RMSE results of LSTM, XGBOOST and ARIMA are listed in Table 1. Between LSTM and XGBOOST, the former one has larger Adj R^2 and smaller RMSEs, which means the array of prediction prices fits better to actual prices. LSTM still has the characteristic of time series forecasting models that predict prices that are apparently affected by the tendency from previous time spans. XGBOOST has more radical predicted prices than other models and even the actual prices. While only giving monthly predictions, ARIMA has a fiercely beyond-actual result, due to the volatile markets of cryptocurrency. Based on the measuring metrics set up, LSTM shows better performance on predicting cryptocurrencies.

Table 1 Metrics Results for test sets (Values in each cell are following the order of LSTM results, XGBOOST results, and ARIMA results.)

	BTC	ETH	BNB	SOL	AVAX
Adj R-Square	0.8954	0.8921	0.8924	0.8046	0.6070
	0.8649	0.7622	0.8915	0.6762	0.6024
	0.8696	0.8660	0.7955	0.7993	0.7993
RMSE	2143.9	183.65	22.716	9.7614	5.0779
	2205.6	198.6	28.38	5.853	7.7682
	7847.4	605.14	121.45	39.295	39.295

Another significant trait of different models' results is that the RMSEs from the three models are all higher for the cryptocurrency with larger market value, while the Adj R-squares are close. From this perspective, all models perform worse on top tier cryptocurrencies.

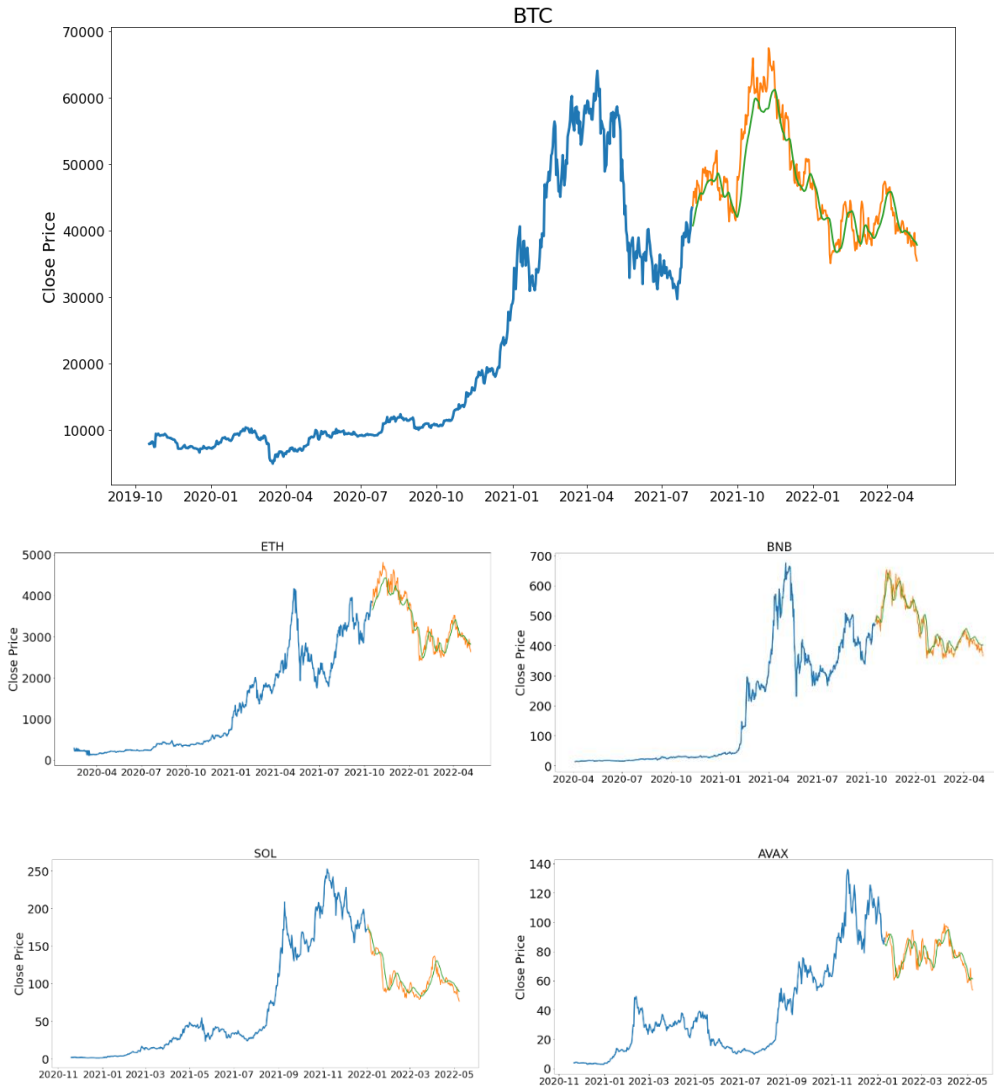


Figure 5. LSTM Predict Prices of Test Sets and Train Sets of BTC, ETH, BNB, SOL and AVAX.

3.2 Correlation analysis

There is a double hook system exists between BTC and ETH, and among ETH with other cryptocurrencies. Figure 6 shows the correlation matrix among five coins from September 22nd, 2020 to May 7th, 2022, the correlation between the two main coins is high on price and volume. Besides BTC and ETH, third tier coins have a stronger connection with ETH, rather than with BTC. Although third tier coins have a high correlation on price with ETH, at minimum 0.8, their trading tendencies are less connected, especially the last two coins, SOL and AVAX. While none of them have such a strong connection with BTC, it shows like tier three coins are anchoring to

ETH, other than BTC. Meanwhile, ETH has the highest correlation on volume with BTC, referring to the cryptocurrency system as a double hook monetary system.

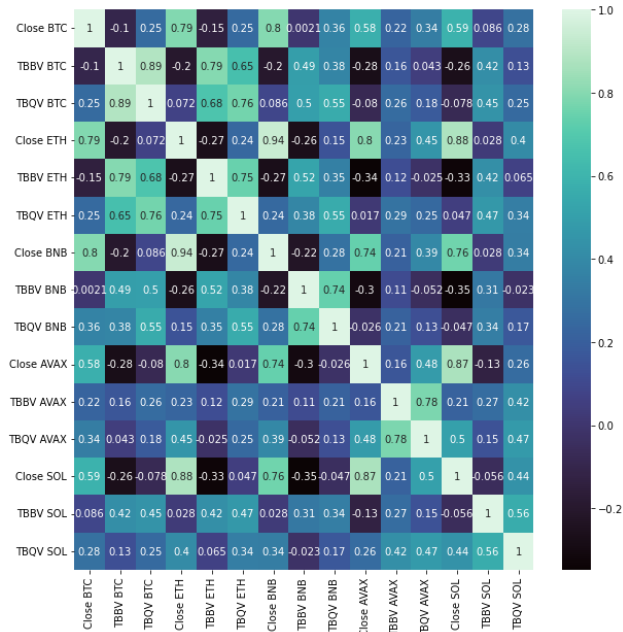


Figure 6. Correlation Matrix Among BTC, ETH, BNB, AVAX, and SOL. Here, TBBV stands for Take Buy Base Volume, and TBQV stands for Take Buy Quote Volume

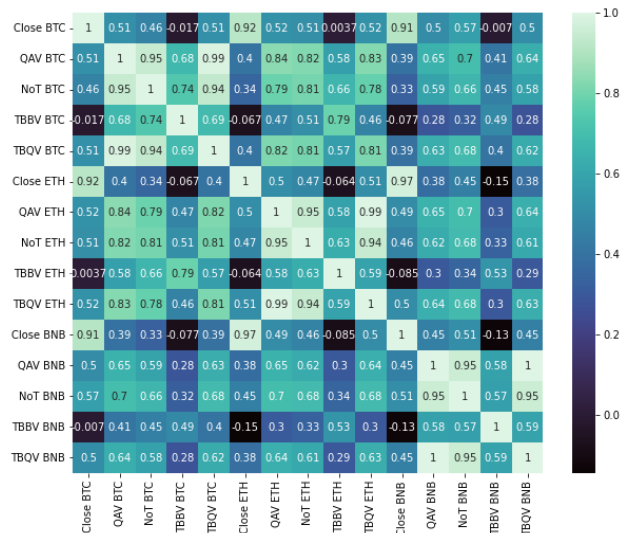


Figure 7. Correlation Matrix Among BTC, ETH and BNB. Here, QAV stands for Quote Asset Volume, NoT stands for Number of Trades, TBBV stands for Take Buy Base Volume, and TBQV stands for Take Buy Quote Volume

By narrowing down the kind of coins range to BTC, ETH and BNB, while extending the start time to July 7th, 2019, Figure 7 shows the correlations among the three main coins are all becoming higher. In this condition, BNB still shows a higher correlation with ETH than with BTC, while the difference is no longer significant than before.

In the long run, cryptocurrencies show a general similar tendency, both in their prices and volumes. Especially, the take-buy transactions show there are common beliefs that exist in the cryptocurrency community. By arguing the value of cryptocurrencies, the authors believe that as the community has naturally grown a spontaneous order, the similar tendencies of different cryptocurrencies are the proof of the existing common belief of the entire cryptocurrency concept is the utility value recognized by current crypto market entities. It is a common belief, the foundation of cryptocurrencies prices.

3.3 Explanation and discussion

The global cryptocurrency markets have a well-developed mechanism. Regulations around the world are very limited compared to financial and physical markets. As a representative trend of regulation, although United States authorities. In this case, The Financial Crimes Enforcement Network and The Internal Revenue Service, do not recognize cryptocurrencies as legal tenders, but define it as a digital representation of value that functions as a medium of exchange, a unit of account, and/or a store of value [11]. Besides, cryptocurrency exchanges only need to follow the Bank Secrecy Act. Additionally, without the restriction of Peel Regulations and central banking system, cryptocurrency markets are approaching a fully competitive market condition. Therefore, having been shown to have competitive traits of pre-American banking systems, cryptocurrencies could be considered a challenger for a modern sovereign-controlled currency system. This would make cryptocurrency issuers value more of the credit of their coins.

As mentioned above, BTC has been the unparalleled first-class cryptocurrency on the market. Which shows the same characteristic of having confidential R-squares while relatively huge RMSEs in prediction models. While combining this characteristic with the continuous deviation between predicted prices and actual prices, we believe BTC faces the initial market impacts. Then, according to common belief, the impact wave would affect second tier currencies (e.g., ETH), and transmit it to the entire cryptocurrency market. After every time a recession or growth of the entire market, top tier cryptocurrencies would take more benefit due to the Mathew effect that exists on competition of currency area.

3.4 Limitation

Nevertheless, this research has some drawbacks and shortcomings. Primarily, the real reasons that affect the actual price are beyond the models, which are external variables. While the entire crypto-market is more easily affected by external variables, compared to the stock market due to the common belief. Therefore, price predicting models are less precise when they are applied to cryptocurrencies. Right after the last day of the dataset, the cryptocurrency market suffered a huge recession on May 7th. The issuer of Luna used more than three thousand bitcoins to save the Luna market but failed, while led shocks to the entire crypto-market. It is these kinds of incidents that impact the common belief that we are unable to predict. Therefore, finding ways to measure the common belief of the crypto-market and determining factors that have impacts on it would be opportunities in this area.

4 CONCLUSION

In summary, this paper investigates cryptocurrency prices based on machine learning and statistical methods. To better understand the market, we make a cross-comparison between variables and models. Specifically, five typical cryptocurrencies including Bitcoin, Ethereum, Binance Coin, Avalanche, and Solana were selected as research targets and three models including LGTM, ARIMA, and XGBoost are utilized for comparison. According to the analysis, the more market power a cryptocurrency has, the bigger the R-square and RMSE will be, which means all models have generally good fitness yet are hard to predict the price precisely when discussing a powerful cryptocurrency. Despite that, LSTM model has still shown the best performance on predicting short run cryptocurrency prices. On the other hand, the paper found the Bitcoin-leading structure of crypto-market from the perspective of prediction models, and gave the authors' explanation. Moreover, the market has a spontaneous order-common belief that decides its price and change; it behaves similar to the credit system that will be manipulated by public opinions. In addition, any wave that causes price change will impact BTC first then transfer to currencies with less market power.

In the future, models should consider methods outside of machine learning and simple supervised models, while improving their accuracy on main coins. At the same time, there lies a demand to find new models for Bitcoin since its potential and market impact on smaller cryptocurrency types. Overall, these results offer a guideline for predicting smaller cryptocurrency prices based on their current values and their correlation with BTC.

REFERENCES

- [1] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," 2014 UKSim-AMSS 16th International Conference on Computer Modeling and Simulation, 2014, pp. 106-112.
- [2] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1936-1939.
- [3] K. Tziridis, T. Kalampokas, G. A. Papakostas and K. I. Diamantaras, "Airfare prices prediction using machine learning techniques," 2017 25th European Signal Processing Conference (EUSIPCO), 2017, pp. 1071-1074
- [4] Z. Chen, C. Li, W. Sun, "Bitcoin price prediction using machine learning: An approach to sample dimension engineering," Journal of Computational and Applied Mathematics, Vol. 365, 2020, 112395, ISSN 0377-0427, pp. 1-13
- [5] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," SMU Data Science Review, Vol. 1, No. 3, Article 1, 2018, pp. 1-21
- [6] S. T. Howell, M. Niessner, and D. Yermack, "Initial Coin Offerings: Financing Growth with Cryptocurrency Token Sales," The Review of Financial Studies, Volume 33, Issue 9, September 2020, pp. 3925-3974.
- [7] P. P. Momtaz, "The Pricing and Performance of Cryptocurrency," The European Journal of Finance, vol. 27:4-5, 2021, pp. 367-380.

- [8] C. Lamon, E. Nielsen, and E. Redondo, "Cryptocurrency price prediction using news and social media sentiment." *SMU Data Sci. Rev.*, vol. 1(3), 2017, pp. 1-22.
- [9] R. Parekh et al., "DL-GuesS: Deep Learning and Sentiment Analysis-Based Cryptocurrency Price Prediction," in *IEEE Access*, vol. 10, pp. 35398-35409, 2022.
- [10] M. Xie, C. Sandels, K. Zhu and L. Nordström, "A seasonal ARIMA model with exogenous variables for elspot electricity prices in Sweden," 2013 10th International Conference on the European Energy Market (EEM), 2013, pp. 1-4.
- [11] Online information, "Cryptocurrency Regulations Around the World." *Comply Advantage*, 2022, available at: complyadvantage.com/insights/cryptocurrency-regulations-around-world