Enhancing Cryptocurrency Price Forecasting: a Comparative Analysis of Feature Quantity and Forecasting Models

Mustabshera Fatima, Mir Sajjad Hussain Talpur, Zeeshan Ahmed Nizamani, Toufique Ahmed Nizamani, Muhammad Yaqoob Koondhar and Zulfikar Ahmed Maher

November 26, 2023

Enhancing Cryptocurrency Price Forecasting: A Comparative Analysis of Feature Quantity and Forecasting Models

Mustabshera Fatima, Mir Sajjad Hussain Talpur, Zeeshan Nizamani, Toufique Ahmed Nizamani,
Muhammad Yaqoob Koondhar, Zulfikar Ahmed Maher
Information Technology Center, Sindh Agriculture University Tandojam
mustabsherafatima@gmail.com

***Abstract*. Cryptocurrencies such as Bitcoin have become a popular digital currency in recent years, attracting investors and traders worldwide. However, the volatile nature of Bitcoin prices poses a challenge for predicting future prices accurately. To this end, this articles examines how different forecasting models for Bitcoin prices are affected by the number of features and which models are more effective at forecasting. We apply different ARIMA, SVM, LSTM, models on the Bitcoin historical dataset from Kaggle to predict the prices. With one-minute intervals, the dataset spans from 2012-01-01 to 2021-03-31. The data is separated into subsets for training (70%) and testing (30%) to get performance results of the models. The results show that the LSTM model performs better only when the appropriate features are selected. The study emphasizes the importance of selecting appropriate features to improve forecasting accuracy, as it can significantly impact the prediction of Bitcoin prices.**

Keywords: Cryptocurrency price forecasting, feature selection, machine learning

I. INTRODUTION

Cryptocurrencies are decentralized digital assets that can be used and transferred without the involvement of a third party, such as a bank, and which are not controlled by a single central. They are created through the mining process and managed by decentralized open-source programs. The exchange of cryptocurrencies occurs on peer-to-peer (P2P) networks, which enable direct contact between any two individuals.

As cryptocurrencies can be used for exchange and have the benefits of decentralization, the removal of middlemen, and trade freedom, their use has grown. With the launch of Bitcoin, developed by a programmer or group of programmers who prefer to remain anonymous (Nakamoto 2009), these currencies began to acquire popularity. There have been many new cryptocurrencies created, their position as a financial fund has grown in addition to their role as a means of exchange. Cryptocurrency prices have risen sharply in recent years as a result of huge investments. Making accurate predictions about their pricing may generate large financial gains. It's interesting to link this to the development of new cryptocurrencies like Bitcoin, Ethereum, Litecoin, and Dogecoin because it is a time series prediction problem in a market that is still in its indeterminate stage. As a result, the market is highly unpredictable, which provides an opportunity for making predictions. Traditional time collection prediction techniques include some machine learning models. Kwon (2019) classified a time series because the more advanced algorithms are timed in nature, recurrent neural networks (RNN) and long short-term memories (LSTM) are preferred over the more conventional multilayer perceptron (MLP). The research is aimed to determine the accuracy of machine learning techniques used to forecast the price of cryptocurrencies like Bitcoin.

The process of cryptocurrency price forecasting involves collecting and analyzing vast amounts of data, including historical price data, trading volume, market capitalization, news, and social media sentiment. One of the primary benefits of using machine learning for cryptocurrency price forecasting is the ability to identify patterns and trends that are not easily visible to human traders. Machine learning algorithms can identify correlations between various factors that may affect the price of cryptocurrencies, including economic indicators, news events, and social media activity. By using machine learning to analyze this data, traders can make informed decisions about when to buy or sell cryptocurrencies, based on their predictions of the market trends, thus improving overall profitability in the cryptocurrency market.

Bitcoin has grown to be the most valuable cryptocurrency in the world. As a result of the massive amount of cryptocurrency transactions, many new currencies were introduced into the cryptography community. It is challenging for researchers to anticipate cryptocurrency rankings.

The popularity of cryptocurrencies, market movements, and the technical aspects of the block chain are just a few of the many elements that affect their price. Due to the ongoing and unpredictable variations caused by these factors, price prediction is extremely difficult. Many studies have focused on the variables that impact the price of cryptocurrencies and the techniques that can accurately predict them. This work focusses on predicting systems using historical Bitcoin dataset.

The major contributions of this paper include the following:

- Investigate the numerous techniques used in forecasting of cryptocurrency prices.
- Apply and Evaluate machine learning techniques to forecast the cryptocurrency prices

Rest of the article is organized as follows: Section II elaborates the related work. Section III describes the dataset acquisition and Methodology. Implementation results are presented in Section IV. Section V is about discussions and Section VI is about conclusion and future directions.

II. RELATED WORK

2.1 Statistical methods

Singh and Malani (2018) utilized crypto attributes including period, open, close, high, low, volume and market cap to predict cryptocurrency market trends. They concluded that Ethereum's price is more predictable than Bitcoin. They concluded that the ARIMA model is more flexible Autoregressive Moving Average (ARMA) model. Bakar and Rosbi (2017) used the ARIMA (2,1,2), and discovered that the strongest significant correlation had two time lags. Coindesk's monthly Bitcoin exchange rates from early 2013 to late 2017 made up the dataset. The ARIMA model performed well with an r-squared of 44.44%. Their study concluded that Bitcoin is rising in value and will do so in the coming months of 2017, making this model an excellent indicator for investors.

Walter and Klein (2018) conclude that the most helpful indicator for predicting value was global real economic activity (GREA). This index categorizes bitcoin market risk. Other exogenous variables, however, might produce superior outcomes. The researchers forecast the price movement of five cryptocurrency markets and its CRIX index, which serves as a benchmark, using the GARCH-MIDAS (Mixed Data Sampling) methodology.

Yang and Kim (2015) published a study that applied network theory and verified several difficulty procedures to examine the relationship among the cost of Bitcoin and entanglement of crypto transactions. Roy and Nanjiba (2018) used historical Bitcoin data from Kaggle and Coindesc, including the date, open, high, low, close, volume, and market capitalization. They used logistic regression, linear regression, autoregressive integrated moving average models (ARIMA), moving average models (MA), and autoregressive models. For the purpose of rendering the model more accurate predictor, the study included Bitcoin with other currencies. Price was included as an additional regressor after the data had been preprocessed and divided into validation, training, and test sets, after choosing features sets. Wang and Chen (2020) approaches based on linear statistical models evaluate linear relationships between prices and explanatory variables. Econometric approaches apply the synthesis of statistical and economic theories to estimate and predict the values of various economic variables. In some cases, statistical model-based techniques can provide good models quickly If you have multiple explanatory variables, you can use multiple linear models to model the linear relationship between the explanatory (independent) and response (dependent) variables.

When studying cryptocurrency price movements using econometrics, researchers typically use statistical models for time series data. Among these models, the most widely used are Generalized Autoregressive Conditional Heteroscedasticity Models (GARCH), Multivariate Linear Regression, Multivariate Vector Autoregressive Models and Extended Vector Autoregressive Models

2.2 Machine learning methods

Machine Learning and Deep learning techniques are frequently utilized in Cryptocurrency price prediction. These networks have been taught to eliminate their defective inputs. Liu et al. (2021) researched sacked denoising auto encoders (SDAE), a new deep learning technique. SDAE outperforms BPNN, PCA-SVR, and (SVR), using the metrics of mean total proportion fault (MAPE), root mean squared fault (RMSE), and turning accuracy (DA). Jiang (2019) predicted the price of Bitcoin using a timestamp and weighted price from a Kaggle dataset. In his research, he evaluated the multi-layer perceptron (MLP), (RNN), (LSTM), and (GRU). He developed the two Layer MLP, three Layer MLP, two Layer LSTM, two Layer LSTM + 1FC, three Layer LSTM, and two Layer GRU. He concluded that while all models performed similarly, LSTM had the best forecast.

Phaladisailoed and Numnonda (2018) used two regression models and two deep learning techniques to estimate the price using two predictive models and two deep learning techniques, the price of Bitcoin is predicted using Kaggle transaction data. To examine model performance Theil-Sen regression & Huber regression, LSTM and GRU (Gated Recurrent Units) models were used. Deep learning models performed better in prediction, with GRU achieving an R2 of 99.2%. Uras et al. (2020) used including SLR, MLR, MLP, LSTM, and Predictive Accuracy (MAPE error), for the price prediction. The paper concludes that the Bitcoin collection of data offers comparable results (0.007 vs. 0.011 MAPE error). The logistic regression model and LSTM network performed better.

Pabuc et al. (2020) used machine learning algorithms to identify the most accurate predictor possible. Two continuous and discontinuous data sets with final, tall, and short values served as foundation for their research. Among continuous data sets, random forestry gave highest efficiency and most accurate.

Guschler (2017) proposed a complex feed forward neural network with Sigmoid and ReLU computations, input and output layers, and three hidden layers. Researchers used time series data and the historical cryptocurrency dataset from Kaggle. AdaGrad (algorithm for gradient-based optimization), an optimizer, demonstrated the highest accuracy, with rmse values of 11.75 and 11.33 for 100 and 200 epochs, respectively.

Samadar (2021) studied KNN, CNN, RNN and Random Forests for the price prediction. Loss function, accuracy, and prediction result were calculated for the neural network after 5 to 10 epochs. The mean squared error (MSE) was the loss function, and the Adam optimizer was also employed. Due to its great accuracy and low loss value, CNN was the most accurate algorithm.

McNally (2018) utilized a number of deep learning algorithms to forecast Bitcoin's price. Early on, feature engineering was used to extract functional patterns from the data. The findings showed that (RNN) had lowest accuracy of 5.45% and the long short-term memory (LSTM) model had the best accuracy of 52.78%. However, LSTM requires 3.1 times more time to train. Along with the open, close, low and high values that make up the Bitcoin Cost Directory, they also utilized the blockchain’s complexity and hash rate. Azari (2019) used Autoregressive Integrated Moving Average (ARIMA) technique to forecast the price of Bitcoin. The results of this investigation revealed a relatively low level of prediction error when the model's performance was analyzed by looking at the mean-squared error.

Maiti et al. (2020) discovered that non-linear neural network models were superior to linear models because they produced more accurate predictions and had lower error rates. The study also discovered that several aspects of the historical data for the coin, such as volume, were useless for forecasting. Statistical and econometric models are frequently used in traditional cryptocurrency price predicting techniques (Brooks 2019).

ML models are applied for the classification and prediction problems. Researchers are working on applying these new techniques to financial markets (Dixon et al., 2020) (El-Bannany et al., 2020) (Galeshchuk et al., 2017) (Ghahfarrokhi et al., 2020) (Nikou, et al., 2020) & (Bagherzadeh, 2019) (Sarlin et al., 2011) (Hatefi Ghahfarrokhi et al., 2020) examined the impact of social media data on predicting variables on the Tehran Stock Exchange. (Galeshchuk et al.,

2017) investigated the ability of deep convolutional neural networks (NNs) to predict the direction of changes in foreign exchange rates. Nikou et al. (2019) evaluated the predictive power of ML models in the stock market. The financial industry has seen an increase in the application and use of ML algorithms for cryptocurrency price prediction.

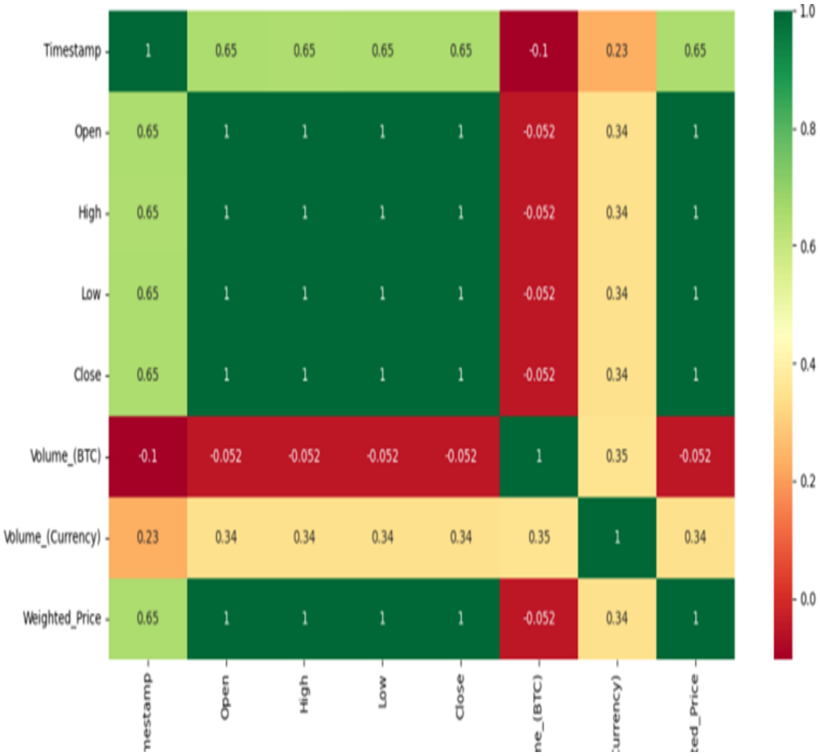

Fig. 1: Heat map with Pearson correlation

III. METHODOLOGY

We construct a database containing eight key features on 238 cryptocurrencies exchanges and use machine learning to predict if a market for the cryptocurrency will continue active or go out of business. We evaluate the feasibility of various models.

3.1 Data Collection

The original dataset was collected from Kaggle in a Comma Separated Values file (.csv). This data collection contains values at intervals of one-minute beginning in early 2012 (early January) and ending on the last day of March 2021. The collection specifically contains the following features: Timestamp, Low(current lowest price), High(highest at the time, Open(period's starting price), Close(price at which the chosen time period ended, Volume(Bitcoin), Volume (Currency): (the volume in US dollars), Weighted price(the cost of bitcoin in US dollars).

Fig. 1 shows correlations between characteristics or goal variables in the dataset. A heat map can be used to determine which features have the highest correlation with the target variable by visualizing the associated features. The Open, High, Low, and Weighted prices have a strong correlation with the Close value. Volume (Currency) has a lower correlation with close, but Volume (BTC) has the lowest correlation.

3.2 Prediction Models

ARIMA (Autoregressive Integrated Moving Average) model, SVM (Support vector machine) model, and LSTSM (Long short-term memory) model are evaluated in this research work for the cyrptocurrency price prediction.

Raw information, data curated from the Kaggle website (Cryptocurrency Historical Data) is processed using scripts. Descriptive statistics for preprocessed data include: Mean data, Median data, standard deviation, First quadrant of data (25%), Second data quadrant (50%), Third data quadrant (75%). We have used Python statistical and machine learning models, for the prediction of Bitcoin prices. The ARIMA model is used for time series by rendering them stationary to aid in analysis.

LSTM approach is used to learn long-term dependencies. The data is scaled or normalized using Min-MaxScaler, with 30% of the dataset used for testing and 70% for training. Adam optimizer and Keras.Sequential() create an LSTM network design.

The SVM model uses a linear model to implement non-linear class borders and searches for a hyperplane in N-dimensional space to classify observations. The dataset is scaled using the StandardScaler object, with 30% of it being used for testing and 70% for training. The sklearn.svm class's SVR algorithm is used for training the SVM model. The training dataset, which make up 0.7 of the overall data set, are used to fit the model and the final 0.3 components are used with a validation set.

IV. RESULTS

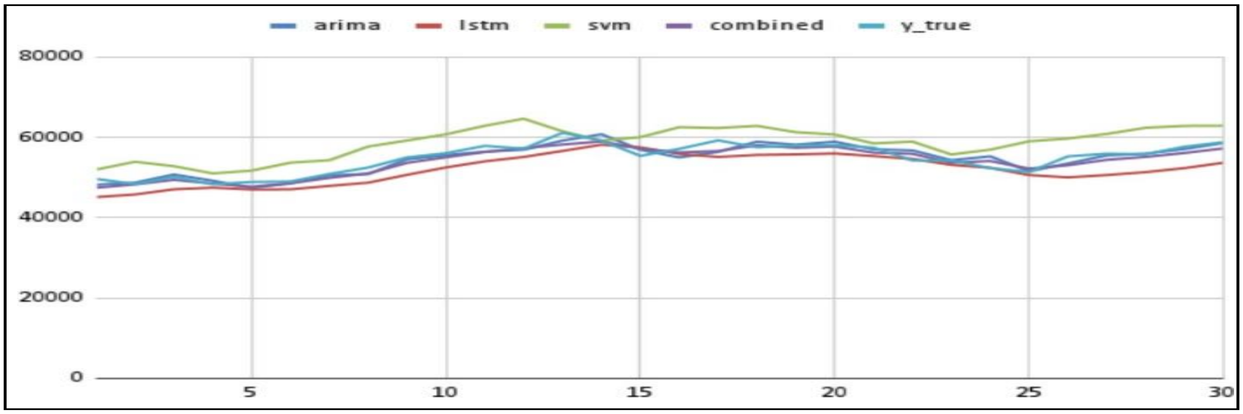

Fig. 2: Performance comparison of the models

For statistical modeling, we have used ARIMA model and an auto-ARIMA model that relies on autofitting, has parameters that were chosen after trying out several models and making mistakes. In machine learning models, we have used SVM and LSTM models. For each model, three criteria will be noted. These are: HQIC (Hannan-Quinn information criterion), BIC (Bayesian information criterion), AIC (Akaike Information Criterion), Log-likelihood. Both the model's complexity and its ability to accurately predict the data are taken into account. The AIC criterion will be the primary factor we use to evaluate each model's quality.

The ARIMA (1,1,1) model is employed in the prediction. This particular form is an ARIMA model with the method ARIMA (1,1,1), which is one the best typical ARIMA model type. In machine learning algorithms, SVM with the RBF kernel, C=1e2, Polynomial-Kernel, Degree=1 is used to develop the model. The new column contains no values for the final 30 rows.

LSTM with activation='relu', one flatten, and one dense are used to train the model. Optimizer=Adam, loss=mean squared error, 64-batch size, and 100 periods are used to build the model. An additional layer with 10 units and an activation relu is added to the LSTM model. The optimizer chosen for the model is "adam," the loss is "mean squared error," whereas the fitting parameters are "100 periods, 32 batches". Training and testing data sets are created from the dataset, accordingly. Predictions for each dataset are displayed in the diagrams below made with initial and mean functions for the Closest value over 1000 epochs.

The fig. 2 compares the statistical and machine learning algorithms. Arima's projected pricing over the previous 30 days are shown by the dark blue line. The LSTM model is shown in red, the SVM model is shown in green, the actual cost value is shown in blue, and the outcome of our forecast model is shown in purple in fig. 2.

V. DISCUSSIONS

Massive volumes of data, including as historical price data, trade volume, market capitalization, news, and social media sentiment, must be gathered and analyzed in order to predict cryptocurrency prices. Machine learning algorithms can find connections among a variety of variables, including as economic data, current affairs, and social media activity, that may have an impact on the value of cryptocurrencies. By utilizing price prediction techniques, traders may forecast future market movements and use those forecasts to decide whether to purchase or sell cryptocurrencies. Due to the growth in the exchange and investment of cryptocurrencies, as well as their decentralization, elimination of agents, and trade freedom, there is a growing interest in forecasting their prices.

The Kaggle bitcoin history dataset is one of the datasets that is most frequently utilized for these investigations. For prediction, we have used features, including Timestamp high, low, open, close, trading volume, and weighted price, standardized using StandardScaler. Training dataset make up 70% of the entire dataset, which is used to fit the model. The remaining 30% is used as test data set.

This paper uses statistical and machine learning models to forecast cryptocurrency price. Three distinct predictive algorithms, including ARIMA, SVM, and

LSTM are applied for prediction. The viability of these models is assessed, and statistics is provided for the preprocessed data.

A time series is made stationary by the ARIMA (Autoregressive Integrated Moving Average) model, which facilitates analysis. LSTM with the Adam optimizer, is effective for time series analysis and is applied in this work.

VI. CONCLUSION AND DISCUSSIONS

6.1 Conclusion

The performance and overall goodness of fit of ARIMA, SVM, LSTM models are evaluated in this study to predict Bitcoin prices. More than a few features have a significant impact on forecasting quality, as demonstrated by the trials.

Our research is constrained by limitations of sample size which has to be lowered because the machine learning models, specially LSTM take an unacceptably long time to handle large datasets. Consequently, the dataset was minimized in a ratio of 1:1400 to all datasets. There was just one row chosen for each day.

Two datasets were produced. The two examples consist of set of data by the mean value for every column besides an information by the initial value for every column. Depending on the number of features employed in the training process, the trials were divided into two different groups: multivariate and univariate approaches. In order to select the ARIMA model with the highest level of predicting accuracy, BIC, HQIC, log-likelihood, metrics are used.

On the 100th and 1000th epochs, univariate LSTM was applied. With mape=0.033, 1000 epochs performed better at forecasting close values. The performance of the Univariate SVM is the worst with mape=0.055.

6.2 Recommendations

There are many machine learning and deep learning applications that may be used with cryptocurrencies in particular and Bitcoin in general. There is a wealth of literature on time series forecasting, it appears that there will be a lot more to be done in the future. Such time series can be subjected to many advanced replacement models for prediction (such as random forests), and artificial neural networks of every kind to assess their predictive potential. A lot of the relevant research yields contradictory findings, making it difficult to determine the relative predicting merit of each one of these methodologies. Future LSTM models, RNN, ANN, decision trees, random forests, and other models can be used with more strategies and more layers.

Given the positive findings, further blockchain traits and factors connected to the solid coin market may help forecast the price rise or fall of cryptocurrency by including features for forecasting that take into account economic occurrences. The economic developments will offer some fresh viewpoints and distinct forecasting vantage points. Future research may also examine how to predict Bitcoin prices using timescales other than 24-hour time series, such as hourly or minute resolution data, which could produce more accurate predictions.

REFERENCES

Abu Bakar, N., & Rosbi, S. (2017) Autoregressive integrated moving average (ARIMA) model for forecasting cryptocurrency exchange rate in high volatility environment: A new insight of bitcoin transaction. International Journal of Advanced Engineering Research and Science, 4(11), 130-137.

Azari, (2019) A. Bitcoin Price Prediction: An ARIMA Approach. arXiv 2019, arXiv:1904.05315.

Brooks, C. (2019) Introductory econometrics for finance. Cambridge, UK: Cambridge University Press.

Dixon, M. F., Halperin, I., Bilokon, P., Dixon, M. F., Halperin, I., & Bilokon, P. (2020) Frontiers of machine learning and finance. Machine learning in finance: From theory to practice, 519-541.

Galeshchuk, S., & Mukherjee, S. (2017) Deep networks for predicting direction of change in foreign exchange rates. Intelligent Systems in Accounting, Finance and Management.

Jiang, X. (2019) Bitcoin price prediction based on deep learning methods. Journal of Mathematical Finance 10, 1, 132–139.

Kwon, D.H., Kim, J.B., Heo, J.S., Kim, C.M., & Han, Y.H. (2019) Time Series Classification of Cryptocurrency Price Trend Based on a Recurrent LSTM Neural Network. J. Inf. Process. Syst. 2019, 15, 694–706.

Liu, M., Li, G., Li, J., Zhu, X., & Yao, Y. (2021) Forecasting the price of Bitcoin using deep learning. Finance research letters, 40, 101755.

Maiti, M., Vyklyuk, Y., & Vukovic, D. (2020) Cryptocurrencies chaotic co-movement forecasting with neural networks. Internet Technol. Lett., 3, e157.

Nakamoto, Satoshi. (2009) "Bitcoin: A peer-to-peer electronic cash system." Decentralized business review: 21260.

Phaladisailoed, T., & Numnonda, T. (2018, July) Machine learning models comparison for bitcoin price prediction. In 10th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 506-511). IEEE.

Singh, A.P. & Malani, S., (2018) Understanding and Predicting Trends in Cryptocurrency Prices Using Data Mining Techniques. IIIT Hyderabad, pp.1-7.

Uras, N., Marchesi, L., Marchesi, M., & Tonelli, R. (2020) Forecasting Bitcoin closing price series using linear regression and neural networks models. Peer Computer Science, 6, e279.

Wang, Y., & Chen, R. (2020) Cryptocurrency price prediction based on multiple market sentiment. In Proceedings of the 53rd Hawaii International Conference on System Sciences.

Yang, S. Y., & Kim, J. (2015, December). Bitcoin market return and volatility forecasting using transaction network flow properties. In 2015 IEEE Symposium Series on Computational Intelligence (pp. 1778-1785). IEEE.