JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Machine Learning Approach in Predicting Bitcoin Prices Using LSTM and 10-Fold Cross Validation

[1]Shreya Kashid, [2]Tejas Machkar, [3]Stephanie Kedari, [4]Harshika Mishra, [5]Dr.Archana Kale

[(1,2,3,4)]Student, [5]Associate Professor

[(1,2,3,4,5)]Department of Computer Engineering,

Modern Education Society's College of Engineering,

19, Late Prin. V.K. Joag Path, Wadia College Campus, Pune - 411001.

Abstract : Bitcoin is a cryptocurrency founded in 2008. It is a new currency that is recognized as a smart and intelligent payment network. Bitcoin uses peer-to-peer technology to operate without central authority or banks; managing transactions and issuing bitcoins is done jointly by the network. The open source code structure of Bitcoin allows it to be uncontrolled and uncontrolled by any organization with bitcoin limited and isolated. It has a central entertainment and tracking system for all transactions and authentication of all payments is protected using public key encryption. Bitcoin prices fluctuate at high prices making it difficult to predict, this is the main reason for this study. The study focuses on the pricing of popular bitcoin currencies using various neural network methods namely Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) as well as ten times cross verification. Analysis of various styles from the bitcoin market is performed and important factors are considered and daily price changes are measured by the neural network model. Live streaming data, as well as the database, is considered a test function from a website called coinmarketcap. Mean Absolute Error (MAE) is considered as a comparative parameter in analyzing the performance of the proposed model with the existing ones. The experiment led to testing hyperparameters to increase the accuracy of the prediction which was significantly lower than the predicted results.

Keywords **:** Bitcoin, Recurrent Neural Network, Long Short Term Memory, K-fold cross validation, Machine learning, Prediction.

I.INTRODUCTION

1.1 A Brief History

In 2008, the first decentralized digital currency or crypto-currency was introduced in a paper written by Satoshi Nakamoto, known as bitcoin [1]. It is seen as the most valuable crypto-currency and is traded across 40 exchanges throughout the world. All transactions involving bitcoin are secured and maintained by a decentralized(not owned by anyone) system called a blockchain, which is like a online ledger of bitcoin transaction created and maintained by the system without any interference or involvement of an individual/organization see [2].It acts like a virtual bank for storing bitcoins. Each individual transaction is represented by a block in the blockchain, and it holds the information persisting to that particular transaction. Every single block in the blockchain is connected to the previous block by holding a data element generated using some data from the previous block. By default, encryption is provided for all the data contained inside a block in the blockchain. It makes use of peer-to-peer technology to provide transactions between any two interested parties securely without the involvement of a third party system. Each transaction contains a public key of the receiver and the owner validates using his own private key.

Advantages of using bitcoin are:

1.Greater Liquidity Relative to Other Cryptocurrencies.
2.Increasingly Wide Acceptance as a Payment Method environment.
3.International Transactions Easier Than Regular Currencies.
4.Generally Lower Transaction Fees.
5.Anonymity and Privacy Relative to Traditional Currencies.
6.Independence from Political Agents and Creators.
7.Built-In Scarcity.

1.2 Motivation

- Currently the financial world is trying to adopt modern blockchain technologies which is trying to create a store of value which will be used in the future, and bitcoin is the first step towards this effort.
- It becomes important then to create models which can help predict the value of bitcoin so that investors today can invest their money wisely.
- They need a reliable source predicting prices of cryptocurrencies which can be used as a base guideline to formulate investment strategies in the CryptoMarket, and this has been the motivation behind this study.

1.3 Requirements

Using open machine learning libraries such as TensorFlow, Scikit-Learn, 10-fold cross-verification and RNN algorithms and LSTM this paper suggests how to predict bitcoin prices using sensory networks. The different types of software and methods used in this project will be discussed in the following sections-

1.3.1 Keras

Keras [3] is an in-depth Python learning API, running on the TensorFlow machine learning platform. Built with a focus on compliance and rapid testing. Being able to move from a point of view as quickly as possible is the key to doing good research. Works on both CPU and GPU. Fast and easy prototyping is possible with this library. Contains the use of layers, enhancements, unlocking functions etc. and supports both convolutional and conventional emotional networks.

1.3.2 Scikit-Learn

Scikit-learn [4] is an open source machine learning library that supports both supervised and supervised learning. It also offers a variety of models for measuring models, pre-data processing, 7 Bitcoin Price Prediction used to select and test a machine learning model, and many other resources written in Python language [5]. random, k-methods etc. It works and combines numpy and explores numerical and scientific libraries.

1.3.3 TensorFlow

TensorFlow [6] is a free and open source software for machine learning. It can be used for a wide range of activities but focuses on training and understanding deep neural networks. TensorFlow is a symbolic mathematical library based on data flow and segmentation. Due to its flexible structure it supports various platforms like CPU, GPU, TPU etc. Developed by the Google Brain team.

1.3.4 K-fold cross validation

It is used to test machine learning models and test model capabilities in invisible data using the re-sampling method. K-cross validation allows training and testing of random data each time and as a result there are variations. This helps to increase the accuracy of the model prediction as the model predicts new data that has not been used in its measurements. Produces a standard result for independent data. Here, K refers to the grouping or folding where the associated data needs to be sorted. In each group we take the Kth group as test data and rest all groups as training data. Then we measure the model in the training database and the test in the test database. This reduces the variability and measurement of steps to provide the most accurate prediction estimate

II. LITERATURE SURVEY

A few prediction models are available as Bitcoin is widely used crypto-currency nowadays. The prediction of Bitcoin process using empirical investigation is carried out in the recent research work [7]. This paper reveals the impact of Bayesian neural networks (BNNs) by analyzing the timeline of the Bitcoin process. This study selects the most relevant from the Blockchain information that is deeply involved in Bitcoin's supply and demand and uses them to train models to improve the prediction performance of the Bitcoin prediction process. It provides meaningful insights by performing Bayesian neural network and the existing non-linear as well as linear benchmark methods.

A study conducted [8], daily Bitcoin closing rates between 2012-2018 to predict Bitcoin prices using Linear Regression (LR) and Support Vector Machine (SVM) in machine learning methods using a time series. The performance of the detected model is measured by statistical indicators such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Pearson Correlation. This study concludes that price prediction performance of the proposed SVM model for the Bitcoin data set is higher than that of LR model.

As cited in [9] , use one minute interval trading data on Bitcoin Exchange from the website named bitstamp in the interval of 2012-2018 to train an efficient and highly accurate model to predict Bitcoin prices. Some different regression models with scikit-learn and Keras libraries have been experimented. The best results showed that the Mean Squared Error (MSE) was as low as 0.00002 and the R-Square (R 2) was as high as 99.24%.

The study in [10], aims to identify and gain insight into the optimal features surrounding Bitcoin price, by understanding and identifying the daily trends in the bitcoin market. The data set recorded over the course of five years, consisting of various features relating to Bitcoin price and payment network. Random Forests and Bayesian regression have been used to predict the estimated changes in the Bitcoin price.

As reported in [11], the Artificial Neural ensemble Genetic Algorithm based Selective Neural Network Ensemble, developed using the Multi-Layered Perceptron as a basic model for each neural network the ensemble, is used to explore the

relation between the features of Bitcoin and the next day change in price of Bitcoin. The consistent accuracy achieved for the classification process was lying in the range from 53% to 63%.

According to the study in [12] Recurrent Neural Network (RNN) along with Long Short Term memory (LSTM) are utilized to achieve the classification accuracy of 52% and Root mean Square Error (RMSE) of 8% .The authors of this work claim that the RNN used with LSTM offers better performance than the traditional RNN and ARIMA models.

A recent research work [7], using blockchain oriented Bayesian neural networks; empirical work has been done for the modeling as well as prediction of Bitcoin price. The authors have presented various Bitcoin log evaluations and prediction results under the rollover. The authors of this work impressively explain the volatility of the Bitcoin price and the related price time series.

The study in [13] addresses the use of machine learning as well ways to predict statistics to detect fraud in the Bitcoin market. Accordingly, the most commonly used price measurement methods, which include a time series forecasts, machine and in-depth reading strategies used to determine weekly / monthly increases and falling Bitcoin price trends. According to the preliminary result, comments made on the popular social media platform we reviewed had a positive effect on SVM. methods. LSTM had no positive or negative side effects with emotional effects. Both the weekly and monthly results of ARIMA had negative results while evaluating prices for emotional outcomes. In addition to the given conclusion, manipulators have played a major role in changing trends throughout the times confusing detection. These roles were seen as inflation, downgrading and price stability. Other than that, SVM gains the best performance when used with the results of emotional analysis and confusing discovery.

As cited in [14], the previously mentioned method of predicting the price of cryptocurrency used in the past price indicators predict future price, which was strongly encouraged by the larger body of work in stock market predictions and in the cryptocurrency market. However, this does not contribute to the dynamic behavior of network companies that may indirectly influence the price, without the previous price indicators. In this article without a common way to use past prices, find value patterns and consider various network features and identify the most closely related determinants. number. Using the data obtained from the analysis of the above factors, this study proposes to develop a machine learning model to predict Bitcoin and Ethereum values, supervised learning method using retrofit, LSTM networks, and conjugate gradient Algorithm is suggested. It is noted that previous network features and prices may be used to accurately predict the value of cryptocurrency. Using those features, machine learning models can be trained and tested. Moreover, the purpose of this article is broad and is intended to provide a first step toward characterization of Blockchain based on Blockchain forecasting. In that regard, it contains a detailed analysis of the top two cryptocurrencies on the market, namely Bitcoin and Ethereum. with forensic research using past prices and trends to predict future prices. The additional key points of the study are as follows:

1. General Styles: The study attempts to analyze the trends in each cryptocurrency database. To do that, data usually uses min-max normalization. After setting the graphs the same as us note
    a. The number of wallets, hash rate, bitcoins value, per transaction cost, difficulty, and miner's profit varies according to price if possible of bitcoin.
    b. In Ethereum databases, features include addresses, hash rating, blocking time, and gas limit by closely following price fluctuations.
2. Purchase and Demand: This Research is trying to study the trends in the need to supply the cryptocurrency market. It recognizes that as the number of wallets and addresses grows differently in Bitcoin and Ethereum, to increase the need for a limited amount of coins. Increasing the level of wallets means more users are joining Bitcoin, leading to an increase in demand for coins. Since the rate of inflation is always small, a new coin provides the system is below demand, which explains the main reason for the increase in the number of wallets.
3. External Features: It has been postulated in the literature that crude oil price may have an impact on the cryptocurrency market. The price of crude oil is affecting electricity prices worldwide, which also has an impact on the performance of mining ponds. Up the price of electricity can force mining ponds to close, and as a result, hash power and throughput of a cryptocurrency usage may decline
4. Hash Rate: Using the block generates new coins in the system, which are given to the miner as a Coinbase reward. Miners earn coins from Coinbase prizes and user fees for the processing process. As the price increases, the corresponding amount of miner's income (in USD) also follows. We noted that Coinbase's earnings and revenue have increased over time. With growth incentive for income, many miners join the ponds in the mines in hopes of making a profit and increasing the monetary reward, which explains why the hash rate increases in price.

According to [15], Price prediction of bitcoin is more widely studied among other topics using AI to address hidden money issues, however there is still room for further research on the use of AI strategies to predict cryptocurrency price. Most of the research focuses on predicting Bitcoin prices in USD. A very few papers cover the price predictions in other fiat currencies.

Conducted research relied mostly on the cryptocurrency price history indicators such as open, high, low and closing prices, while few studies take into consideration different sources of social media, online metrics and other stock markets indicators. These include social media sites and their content such as tweets and their sentiments, Reddit posts, Wikipedia views, and Google Trends data. Most studies rely heavily on cryptocurrency price history indicators such as open, high, low and closing prices, while very few studies also take into account the various social media and other stock market indicators. Social media posts and their feelings, tweets, Reddit posts, Wikipedia views, and Google Trends data are some of the many resources that can be used to predict more accurate results.

A few research efforts considered using BitcoinTalk forum posts. While sources that can shape the opinion of some novice investors and influence users' buying or selling behavior like news sources, such as newspapers and news agencies have not been taken into account at all. These sources are very crucial in predicting bitcoin prices as technical news about security breaches or instability of the crypto market have a huge impact on the market.

The proposed model in [16] is a stochastic neural network model for Cryptocurrency price forecast as it is based on the concept of random movement, which is widely used in financial markets to make a stock price model. There are 3 important data sources, the first of which are market statistics that include day and night and volume. Second, blockchain network information including transaction calculations, transaction fees, etc. Lastly, public sentiment details such as google styles and tweet volume are considered

To that end MLP and LSTM models are used because there is a need to capture the indirect dependence between market factors, blockchain data and public sentiments and neural networks are the most suitable for these tasks. MLP is a very basic type of neural network and LSTM models are widely used in time-dependent data situations, such as market prices for cryptocurrencies

The stochastic neural network module is the stochastic, which is added to the output of the entire layer in the neural network. After performing the matrix multiplication and activating the layer, the output is transferred to a stochastic module. The model takes into consideration the previous step timestep's stochastic activations as well as the current activation as input. The combination of stochasticity and non-linear dependence of the model thus produces improved results.

III. PROPOSED METHODOLOGY

3.1 Data Exploratory Graphs:

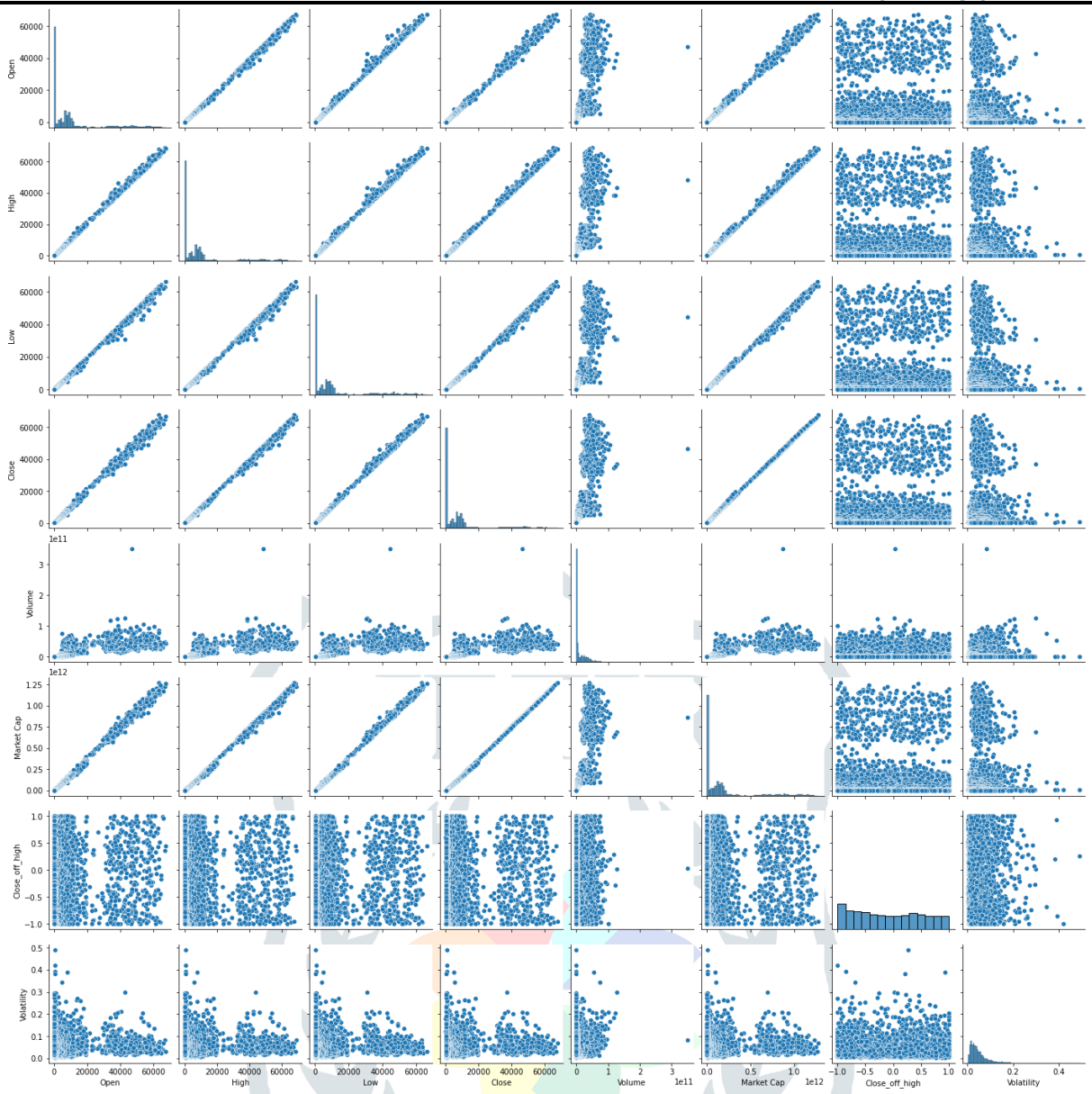
The Data Set used in this project is scraped from the site coinmarketcap using the cryptocmd scraper.The data set is made up of seven features namely:

- Open: The opening price of bitcoin on that data.
- High: The highest price bitcoin was traded at on that particular day.
- Low: The lowest price bitcoin traded at on that particular day.
- Close: The closing price of bitcoin on that day.
- Volume: The total amount of bitcoin traded on that particular day.
- Market Cap: The market cap of bitcoin as of that day.
- Time: DateTime value according to which all the features are recorded.

The following graph shows the bitcoin weighted price throughout the years since it was first created.

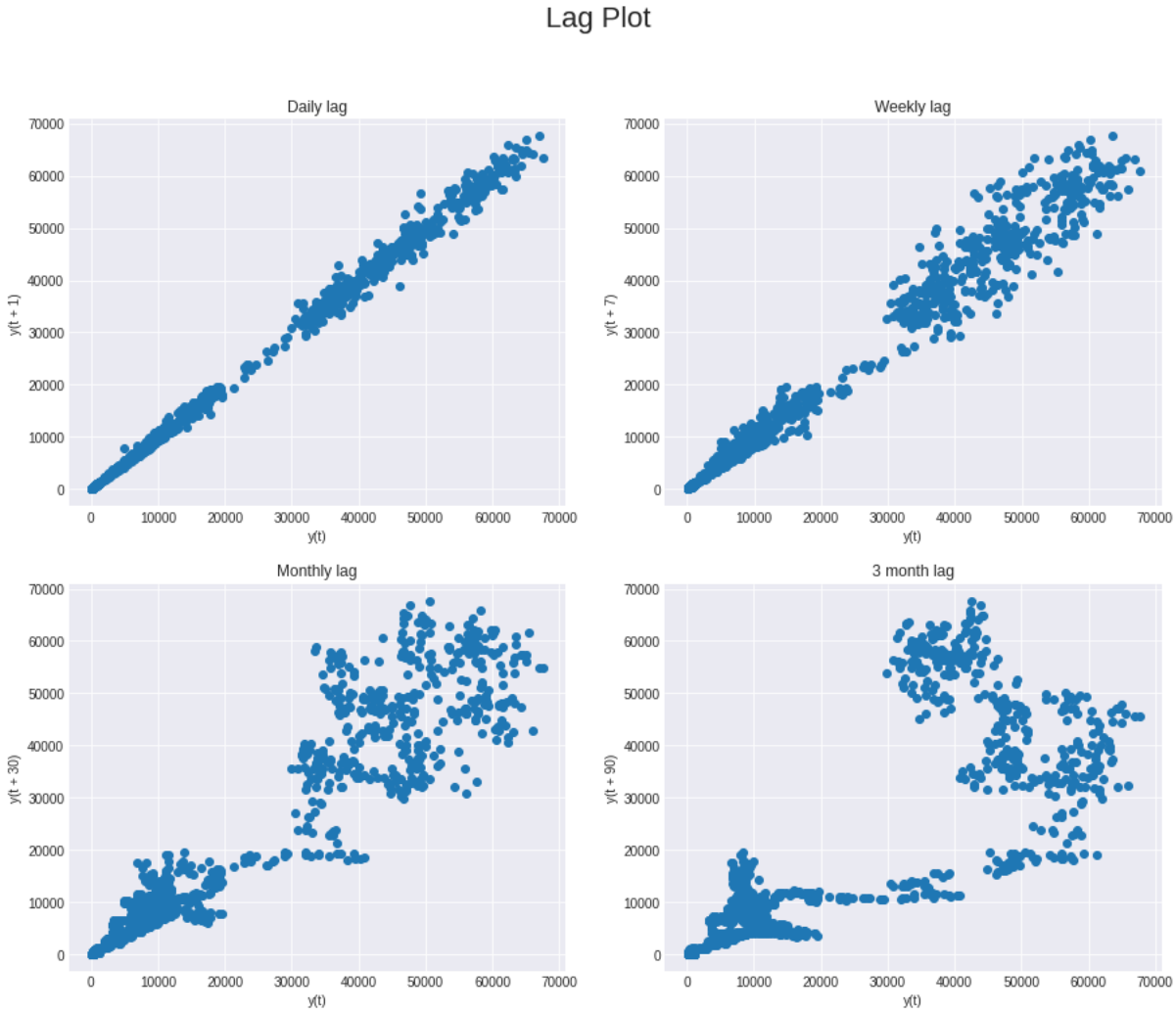
From the graph we can see a general upwards trend in the prices of bitcoin which has seen a significant spike in the last couple of years.

The following is the pair plot of the features of the data set:



Next we will take a look at lag plots of the close price of the data set to check the autocorrelation of the bitcoin closing prices with the time series data in our data set.

- Lag plots are useful in analysis of autocorrelation in time series data. Autocorrelation is the phenomenon or tendency of values belonging to time series data to be correlated to previous values in the data.
- We want to check whether the current values are correlated to values a week before and also to values month before.



From the above plots we conclude the following things:

- The autocorrelation is strongest on the daily basis
- The autocorrelation starts to break down a little when we move on to a weekly basis.
- For monthly and 3 month lag there is almost no correlation which exists between the values.

As our dataset is already grouped on the daily time interval we conclude that there is no further grouping required on the time series data.

3.2 Recurrent Neural Network with LSTM and 10-fold Cross Validation

RNN with LSTM is a widely used neural network algorithm for predicting price as it remembers some important data which is received by the input and helps them to predict price of the next output accurately. It is mostly used in sequential data. Previously in the feed forward network the data flows from input to hidden layer then to output layer for obtaining results and ignores some important information, hence RNN comes into picture. RNN has two inputs one for the present data and another for past data for consideration. Based upon this fact that RNN also contains some crucial historic data which form an essential ingredient for consideration, RNN can do certain things which other algorithms cannot. Gradient in RNN is responsible for obtaining the change in weight with respect to change in the error. Simple RNN faces two problems namely Vanishing Gradient and Exploding Gradient. Exploding Gradient is the problem where the models focus only on weights whereas in the Vanishing Gradient, low values of gradient forms the cause of the problem. Thus, LSTM comes into view. LSTM is basically the extension. LSTM increases the memory of the model for solving the above two problems. Hence so much information is stored which further reads, writes and can be deleted from memory. Gates helps to identify what data is required to store or not over a time. It has three gates namely Input gate, Output Gate and Forget Gate. Here, Sigmoid function is used which ranges from the values 0 to 1.

The LSTM has been utilized in the proposed work. Here, firstly the optimum size of the window is selected for storing information. By analyzing the graph of closing price with date, estimation of the optimal window length for better prediction is done. Close price in the graph would have a lag of up to 10 days so we have used temporal window length to be 10 for RNN. But LSTM shows better results in large window Length = 100 considering two hidden layers. Here, the authors have considered a window length of 100 days and two hidden layers for enhancing the efficiency. In this work, LSTM with 10-fold cross validation is also applied.

First the data is retrieved from the website known as coinmarketcap In addition, data purification and customization is performed when prices are close to volume from -1 to 1.Some new columns are created namely price volatility (simple means difference between high and low price divided by the opening price) and price high for that day which signifies the differences between high and close divided by difference between (high and low)-1.

It is therefore advisable to select model hyperparameters (their value parameters are used to control the learning process) in such a way that the training is conditionally effective in both time and fit (whether the model "knows" the training data too well, or too poor; to constrict any form of overfitting or underfitting).In the model, the window length of the model is randomly chosen. Every LSTM layer should be accompanied by a dropout layer, such a layer helps avoid overfitting in the training phase by bypassing randomly selected neurons, thereby reducing the sensitivity to specific weights of individual neurons. A range of dropout (viz. 15-30%) values was tested against the model and it was found that the best compromise between preventing model overfitting and retain model accuracy is found at dropout value of 25% .A network activation function means that the measured amount of input is converted into output from a location or nodes in the network layer. The choice of activation function has a significant impact on the power and function of the neural network.The choice of activation function has a large impact on the capability and performance of the neural network. Historically, the tanh function became preferred over the sigmoid function as it gave better performance for multi-layer neural networks, and hence "tanh" is the choice of activation function in this model. Epoch is the hyperparameter that sets how many complete iterations of the dataset is to be run. The model was trained over a range of epoch starting from "50" and "300" with an increment of 50 throughout the range. It was found that above the value of 100 epochs the validation accuracy of the model started decreasing even though the training accuracy of the model was increasing, which is an indication of the model overfitting on the learning dataset. Hence, the value of epoch was set to 100 in the final model. The batch size is a hyperparameter that describes the number of samples to be processed before the internal parameters of the model are reviewed. Upon research it was found that a good default value of batch size which is widely accepted is 32. The complex training dynamics of recurrent neural networks are better handled by adaptive optimizers such as "adam" and hence it was chosen as the choice of optimizer. Our target i.e. the bitcoin price, is normally distributed, conditioned to the input and we want our large errors to be significantly penalized more than our small errors, and hence we have used mean squared error as our loss function. Now the close price and volume was fitted in the final model with the hyperparameters tuned. After getting the MAE (Mean Absolute Error) of the training data, a similar process was carried out for the test data. Then the Mean Absolute Error and Mean Square Error for the test data were calculated. The result was verified for every activation function and the one was chosen which shows highest accuracy. In the second model, first a 10-fold cross validation was applied. Then the model was split and trained into 10 folds or groups. After that, the model was executed for each fold. After fitting the model MAE was calculated corresponding to each fold.

IV. RESULTS

The proposed model reflects enhanced accuracy in terms of the mean absolute error (MAE).These MAE values were achieved by implementing the proposed (RNN along with LSTM) model on live streaming data with some selected features which affect the bitcoin prices in significant ways. These features were chosen after the analysis part. In the proposed work, the 10-fold cross validation was applied with RNN and LSTM to increase the continuity and efficiency which resulted in the reduced MAE value. After calculating MAE of each fold, the total MAE (in percentage) was calculated to be 0.36% which was a significant decrease in MAE from 0.56% as per [17]. This was achieved by testing the hyperparameters and by feature selection. The hyperparameters which contributed largely to the decrease of MAE were optimizer, epochs and the batch size which helped us give better accuracy.

V. CONCLUSION

By comparing the proposed model with the two existing ones, it can be concluded that cross validation in RNN with LSTM helps in increasing the efficiency of the model for Bitcoin prediction. It is observed that this is due to the Bitcoin data being unstable and with the usage of cross validation with RNN and LSTM, the results are more enhanced as it estimates the mean of all MAE by taking different test and training data dynamically. The proposed RNN with LSTM using 10-fold cross validation gives MAE of 0.36% which is significantly less than what is offered by the Random Forest and Linear Regression models.

VI. ACKNOWLEDGMENT

We are grateful to the Head of Department, Dr. N. F. Shaikh and entire Computer Engineering Department of Modern Education Society's College of Engineering, Pune for their constant guidance and support.

VII. REFERENCES

[1] S. Nakamoto, "Re: Bitcoin p2p e-cash paper," The Cryptography Mailing List, 2008.

[2] W. Zheng, Z. Zheng, X. Chen, K. Dai, P. Li, and R. Chen, "Nutbaas: A blockchain as-a-service platform," IEEE Access, vol. 7, pp. 1344.

[3] F. Chollet et al., "Keras: The python deep learning library," Astrophysics Source Code Library, pp. ascl–1806, 2018.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in python," the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.

[5] G. Van Rossum and F. L. Drake, The python language reference manual. Network Theory Ltd., 2011.

[6] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., "Tensorflow: A system for large-scale machine learning," in 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp. 265–283, 2016.

[7] H. Jang and J. Lee, "An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information," IEEE Access, vol. 6, pp. 5427–5437, 2018.

[8] S. Karasu, A. Altan, Z. Sarac¸, and R. Hacioglu, "Prediction of bitcoin prices with ˘ machine learning methods using time series data," in 2018 26th Signal Processing and Communications Applications Conference (SIU), pp. 1–4, 2018.

[9] T. Phaladisailoed and T. Numnonda, "Machine learning models comparison for bitcoin price prediction," in 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 506–511, 2018.

[10] S. Velankar, S. Valecha, and S. Maji, "Bitcoin price prediction using machine learning," in 2018 20th International Conference on Advanced Communication Technology (ICACT), pp. 144–147, 2018.

[11] E. Sin and L. Wang, "Bitcoin price prediction using ensembles of neural networks," in 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 666–671, 2017.

[12] S. McNally, J. Roche, and S. Caton, "Predicting the price of bitcoin using machine learning," in 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), pp. 339–343, 2018.

[13] F. Akba, I. T. Medeni, M. S. Guzel, and I. Askerzade, "Manipulator detection in cryptocurrency markets based on forecasting anomalies," IEEE Access, vol. 9, pp. 108819–108831, 2021.

[14] M. Saad, J. Choi, D. Nyang, J. Kim, and A. Mohaisen, "Toward characterizing blockchain-based cryptocurrencies for highly accurate predictions," IEEE Systems Journal, vol. 14, no. 1, pp. 321–332, 2020.

[15] F. Sabry, W. Labda, A. Erbad, and Q. Malluhi, "Cryptocurrencies and artificial intelligence: Challenges and opportunities," IEEE Access, vol. 8, pp. 175840–175858, 2020.

[16] P. Jay, V. Kalariya, P. Parmar, S. Tanwar, N. Kumar, and M. Alazab, "Stochastic neural networks for cryptocurrency price prediction," IEEE Access, vol. 8, pp. 82804–82818, 2020.

[17] S. Tandon, S. Tripathi, P. Saraswat and C. Dabas, "Bitcoin Price Forecasting using LSTM and 10-Fold Cross validation," 2019 International Conference on Signal Processing and Communication (ICSC), 2019, pp.323-328, doi:10.1109/ICSC45622.2019.8938251.