IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Utilizing Machine Learning and Deep Neural Network Model for Crypto Currency Price Prediction

Mayank Chaudhary
Department of Computer Science and Engineering
ASET, Amity University U.P
Noida, India

Dr. Anant Kumar Jayswal
Department of Computer Science and Engineering
ASET, Amity University U.P
Noida, India

***Abstract* -** For several decades now, the accurate and precise forecasting of the values of crypto currency index values has been a significant topic of study. In this paper, we present a hybrid modeling strategy based on two parameters which are moment correlation coefficient and root mean square error (RMSE) for predicting crypto currency prices by constructing deep learning and machine learning based models which are multivariate linear regression, MARS, artificial neural network (ANN), random forest, support vector machine (SVM), bootstrap aggregation, decision tree, extreme gradient boosting (XG Boost) and long short term memory (LSTM). We utilized the crypto index data from the online crypto currency exchange from April 1, 2021, to March 31, 2023, pertaining to the analysis we're doing.. With the assistance of the training data, which included crypto index records from April 1, 2021, to December 31, 2022, we created eight regression models. Using these models, we forecasted the crypto index open values across the time frame from January 01, 2023, to March 31, 2023, and found that the Long Short Term Memory (LSTM) model was the most precise.

Keywords* - *Crypto currency, Correlation, RMSE, Multivariate Linear Regression, MARS, ANN, Random Forest, SVM, Bootstrap Aggregation, Decision Tree, XG Boost, LSTM, crypto index.

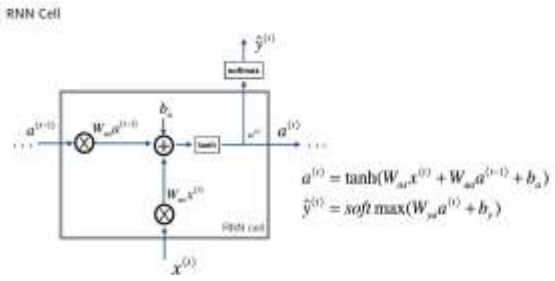
I. INTRODUCTION

For a considerable amount of time, the topic of predicting the price of crypto currencies has caught the attention of academics and they have shown that with the right variables and modelling, we can reasonably predict the values of various crypto currencies in the Indian market. The people's prediction of price increases against price decreases because the market for crypto currency prices to fluctuate. It follows that any signal that provides accurate forecasts of future prices will be discovered and reflected in the price, which is virtually inevitable. Finding these small-scale economies, using them while they endure, and discovering new ones is a complete business and is a problem that requires lots of work by analysing and predicting information that is collected and distributed by large scale entities in the market. This sector employs thousands of people full-time and has access to far more data than is generally available to the general public. In general, market returns are non-normal and this volatility is highly autoregressive, asymmetric. Most every financial distribution is leptokurtotic and thus either modelled through a Students-T, Generalized Normal, or Pareto Family Distribution.

Despite people saying that there are too many unknowns, parametric models still out perform non-parametric models, ML models, and Neural Networks in most scenarios but you do need to curb expectations, though. You can't predict the price point of a crypto currency at a certain time very accurately, but you can predict periods of volatility in which crypto currency prices fluctuate more frequently and this is very useful to an investor.

Artificial Neural Networks (ANN) like a decision tree where some input data is analyzed in one step and then depending on the result, sent to the next appropriate analysis. They are like those graphs in magazines which are made for fun/humour where each bubble is a question and you follow your response in the appropriate line to the next bubble until you get your result. This can be used to recognize a complex object by breaking it into individually recognizable elements or patterns, kind of like how you play 20 questions to narrow down the object with each question's response also narrowing down the next possible set of questions you could ask. A decision tree is a type of flow chart where basically you have a choice and from each choice you have branches that lead to the outcomes of those choices which may, themselves be (or lead to) more choices which may have more branches. When you draw this out, it looks like an upside-down tree. It is often used in programming by breaking larger problems into simple yes/no questions. With random forests, both training and scoring are easily parallelizable. A single tree is fast to train, and you can train each tree independently of the others. The identification of a separating hyper plane in a linear space serves as the basis for SVM. By using this method, every piece of data is represented as a point in an n-dimensional space, and each component's estimation is an estimation of a particular coordinate. We have support vectors and non-support vectors, which is why it is named Support Vector Machines. Bootstrap aggregation or bagging is when you resample multiple times from some data sample to lower the deviation of the original sample from the actual. Bagging is used to reduce overfitting in predictive modelling. Instead of making predictions based on a single high-variance model, the idea of bagging is to train a multitude of high-variance models (using subsampling with replacement) and aggregate their prediction. Extreme Gradient Boosting or XGBoost is really good for tabular data and if you want to apply XGBoost to non-tabular data, then you have to design your own features. Extreme Gradient Boosting learns regimes and extrapolates from known data and is a perfectly viable option for shorter timeframes. Factors like balance sheet information & company health play important roles in filtering companies that can weather market downturns and establish a baseline of which companies investors will flock to during these bear markets, but they don't play as large of a deterministic role as you'd expect them too.

ML "algorithms" are basically just non-linear regressions. Long and short term memory (LSTM) should be able to identify the complex sequence function without doing much of tuning and the main advantage of LSTM is that it can decide what to forget and remember; even what an input were many time steps ago. Financial datasets have an extremely high noise-to-signal ratio, potentially high-dimensional, and fat tailed. Thus odds are that any relation found by a statistical model (which is what ML models are) is just a regression overfit to the noise in the data. Additionally, any model based on data alone will underestimate risk due to the fat tails. You need to come up with a prediction task based on this vector, for example predicting the next element in the sequence by outputting a distribution over a dictionary containing the letters and characters in your strings. Finally, you need to define a loss (e.g. softmax cross entropy) and set up an optimizer (e.g. Adam) that minimizes said loss.


RNN Cell
$\hat{y}^{\langle t \rangle}$
softmax
$a^{\langle t-1 \rangle}$
$W_{aa}a^{\langle t-1 \rangle}$
b_a
tanh
$a^{\langle t \rangle}$
$W_{ax}x^{\langle t \rangle}$
RNN cell
$x^{\langle t \rangle}$
$a^{\langle t \rangle} = \tanh(W_{ax}x^{\langle t \rangle} + W_{aa}a^{\langle t-1 \rangle} + b_a)$
$\hat{y}^{\langle t \rangle} = soft\max(W_{ya}a^{\langle t \rangle} + b_y)$

In this RNN, there is a sequence of the hidden layer activations for the first 3 layers, which forms the hidden state of the 3rd and 4th layers. The hidden state of the 2nd and 3rd layers is used to form the final hidden state of the 3rd layer. The 3rd and 4th layers are then fed into the hidden state of the 1st layer, and so on, until the sequence is all the hidden state of the 1st layer. The sequence of activation values for the 1st layer is a sequence of the last two hidden state activations.

II. LITERATURE REVIEW

The Indian crypto market is one of the biggest emerging markets in the world, and ranks in the top five in the world. Predicting the crypto market in the Indian context will provide a unique perspective as the market is semi-efficient.

Sen et al. in [1] presented a new and advanced method for forecasting the crypto currency values by using the breakdown of time series. Numerous studies on return predictability have revealed that there is some predictability, with rather consistent results.

Sen et al. in [2] showed us the breakdown of data by using the power of time series of crypto currency values and the effects it has on the forecasting precision.

Sen et al.'s argument in [3] concerning asset allocation hypothesis is supported, for illustration, by the development of habits that recognize that gains rely on risk tolerance (as well as in competitive exchanges) and that threat resistance fluctuates as time passes makes it possible to estimate gains.

Sen et al. in [4] explained that each crypto currency moves a bit differently than each other crypto currency, based on countless, ever-changing conditions.

Sen et al. in [5] proposed that in most machine learning cases, both sides of the equation want the same thing. For example: if I am predicting the customer's propensity to buy a product, it's most likely that the company wants to sell the product and customer wants to buy it - we have consent on both sides.

Sen et al. in [6] suggested that autoregressive forecasts require multiple seasons of data and we should expect annual and weekly (weekday vs. weekend) seasonality for logistics, with inflection points due to holidays.

Sen et al. in [7] recommended that instead of looking at the daily items sold, one should look at monthly items sold or even yearly items sold. This should smooth out and flatten and form either an upward curve (if the company is selling more over time) or go sideways in the long term, showing the company isn't upping their sales. You may be able to find seasonality in actual seasons, like summer time vs. winter time sales with multiple years of monthly data. Once you see a pattern, you can use a forecasting tool.

Sen et al. in [8] realized that it's very hard to define what constitutes demand. It may sound straightforward when you just take the aggregate of realized past orders as past demands and forecasts from there but that's not realistic from business point of view because you didn't take into account of out-of-crypto currency events, promotional events, etc.

Sen et al. in [9] showed that the predictive model depends on some characteristics of the data. If your data is static, so methods like Random Forest, Decision tree and so on will work great. But if your data is related to a dynamical system, just adding static features to a model will not solve the problem, no matter how many features.

Sen et al. in [10] suggested that financial datasets have an extremely high noise-to-signal ratio, potentially high-dimensional, and fat tailed. Thus odds are that any relation found by a statistical model (which is what ML models are) is just a regression overfit to the noise in the data.

Mehtab et al. in [11] provide that any model based on data alone will underestimate risk due to the fat tails. You need to come up with a prediction task based on this vector, for example predicting the next element in the sequence by outputting a distribution over a dictionary containing the letters and characters in your strings.

Jaffe et al. in [12] presented research based on the usage of variables and the method used to model the issue of forecasting the values of crypto currencies.

Fama et al. in [13] proposed that a good way to really test the accuracy is checking for higher or lower. If your model predicted higher than the previous day, and the next day is higher, you have a point. If you can achieve more than 51%, then you have something there.

Chui et al. in [14] represented that these machine learning models occasionally give correct results. Basu et al. in [15] suggested that LSTM takes an input as well as a "memory vector" that is supposed to "remember" parts of all the previous vectors. Basically it's useful for things like time series prediction and is an original sort of monotonous frontal cortex affiliations (RNNs) - mind networks that award examination circles to pass on information.

Rosenberg et al. in [16] proposed that in order to achieve significant indications for crypto currency market investment; one must spend time and money in acquisition of the right data. The informative representations provided by these technologies may aid those considering investing in making the most effective investment decisions.

Geron et al. in [17] suggested that when you train your LSTM model, you need to train it to predict a target (which is the next day's price) using the features (price, volume, etc.). Then after the close today, you feed that model all of today's data (features) and it will predict the target (tomorrow's price).

According to Brownlee et al. in [18], LSTMs or GRUs may experience the vanishing gradient problem if they include an excessive number of past observations. There is an efficiency compromise when using RNN and the techniques that are used to generate highly sophisticated CNNs.

Mehtab et al. in [19] represented that in LSTMs you have an output gate that decides (based on the current input) to which extend the cell state (which is separate from the hidden state, but with the same dimensionality) should be exposed at each dimension/neuron.

Thus, predicting the crypto currency market becomes a very challenging task as there is always scope for surprises which may lead to a significant rise or fall in the market.

III. DATASET DESCRIPTION

We obtained the daily crypto index data from Kaggle [20] for the time frame of April 1, 2021, through March 31, 2023. The factors are the following: the date, the initial value, the maximum value, the minimum value, the final value, and the amount traded on a specific day.

For the accurate forecasting and prediction, the regression method was used. Open served as the response variable while the other factors served as predictors in order to achieve this goal. The data was pre-processed before being used to train and test the regression models. Utilizing training data for the period of April 1, 2021, through December 31, 2022, we will create eight machine learning based regression models and one deep learning based model. We will forecast the crypto index's open values between January 01, 2023, and March 31, 2023, using these models.

IV. PROPOSED METHODOLOGY

The following machine learning techniques were developed and evaluated in the current research: (i) Support vector machine (SVM); (ii) Artificial neural network (ANN); (iii) Multivariate linear regression; (iv) Bootstrap aggregation (Bagging); (v) Multivariate adaptive regression spline (MARS); (vi) Extreme gradient boosting (vii) decision trees; (viii) Random Forest and Long Short Term Memory (LSTM), a deep learning-based model.

LSTM takes an input as well as a "memory vector" that is supposed to "remember" parts of all the previous vectors. Basically it's useful for things like time series prediction and is an original sort of monotonous frontal cortex affiliations (RNNs) - mind networks that award examination circles to pass on information. RNN affiliations of the relationship allotment is reliant upon the obligation to the relationship plan opening nearby the condition of the relationship in the past time allotment. Regardless, RNNs experience the shrewd effects of an issue known as dispersing and detonating incline issue, in which an affiliation either quits learning or keeps on learning at an exceptionally. LSTM networks defeat the issue of scattering and detonating propensity issues by certainly failing to review a couple of past unessential data, and thusly such affiliation shows really legitimate for showing consistent information, similar to messages.

We employ two measures to assess the performances of these models which are moment correlation coefficient and root mean square error (RMSE). The models with greater correlation coefficient values and lower root mean square error (RMSE) are thought to be more accurate and precise.

V. PERFORMANCE RESULTS

The outcomes of prediction are thoroughly discussed in this section. All machine learning models testing and training results are shown. Figure 1 shows the graph of comparison between these eight algorithms based on using product moment correlation coefficient as the parameter and as discussed above, higher correlation is considered better. Since test performance is what counts, we see that multivariate regression, MARS, and random forest are better than other models if we take correlation into account, between the test dataset's open values, both real and anticipated. Figure 2 shows the graph of comparison between these eight algorithms based on using root mean square error (RMSE) as the parameter and it should be low. The multivariate regression and the random forest, nevertheless, produced the lowest values of RMSE.

TABLE I. Support Vector Machine (SVM)

Crypto currency	Training		Testing	
INDEX	RMSE Correlation	0.72 0.98	RMSE Correlation	8.36 0.54

TABLE II. Artificial Neural Network (ANN)

Crypto currency	Training		Testing	
INDEX	RMSE Correlation	12 0.6	RMSE Correlation	19.36 0.42

TABLE III. Multivariate Linear Regression

Crypto currency	Training		Testing	
INDEX	RMSE Correlation	0.25 0.98	RMSE Correlation	0.40 0.98

TABLE IV. Bootstrap Aggregation (Bagging)

Crypto currency	Training		Testing	
INDEX	RMSE Correlation	1.71 0.98	RMSE Correlation	3.68 0.95

TABLE V. MULTIVARIATE ADAPTIVE REGRESSION SPLINE

Crypto currency	Training		Testing	
INDEX	RMSE Correlation	0.40 0.98	RMSE Correlation	0.81 0.98

TABLE VI. DECISION TREE

Crypto currency	Training		Testing	
INDEX	RMSE Correlation	2.4 0.9	RMSE Correlation	10.32 0.14

TABLE VII. EXTREME GRADIENT BOOSTING (XG BOOST)

Crypto currency	Training		Testing	
INDEX	RMSE Correlation	0.34 0.98	RMSE Correlation	1.82 0.96

TABLE VIII. RANDOM FOREST

Crypto currency	Training		Testing	
INDEX	RMSE Correlation	0.26 0.98	RMSE Correlation	0.41 0.98

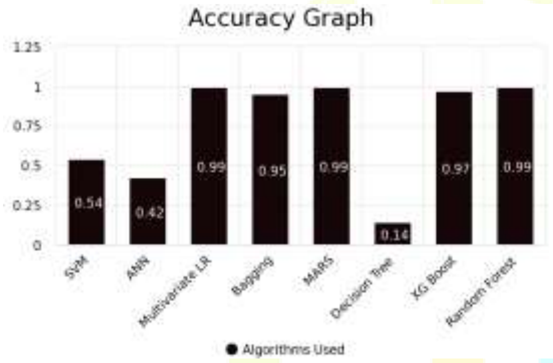


Fig. 1. Comparison graph on the basis of Correlation

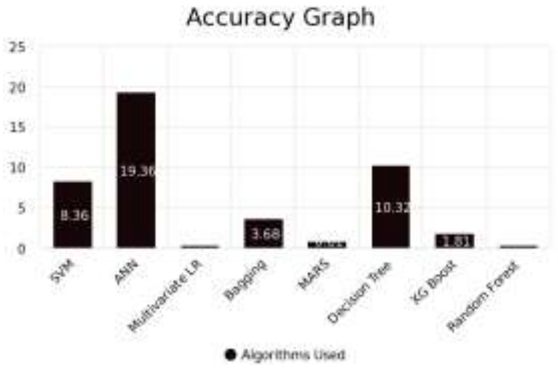


Fig. 2. Comparison graph on the basis of RMSE

TABLE IX. LONG SHORT TERM MEMORY (LSTM)

Crypto currency	Training		Testing	
INDEX	RMSE Correlation	0.01 0.99	RMSE Correlation	0.02 0.99

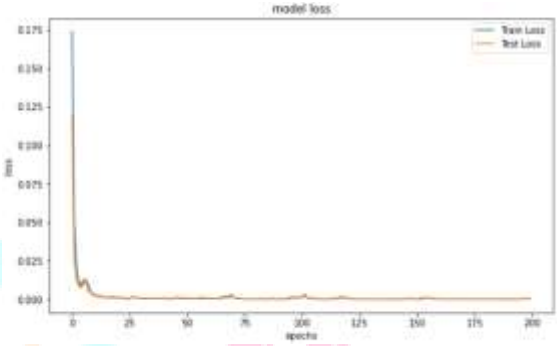


Fig. 3. LSTM train-test loss

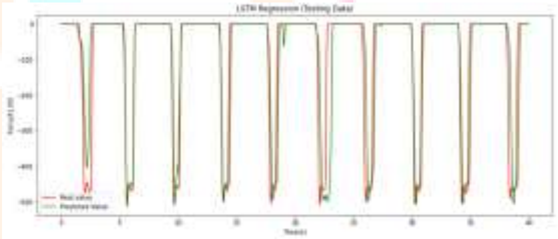


Fig. 4. LSTM regression accuracy

You can clearly see that the resulting prediction by the LSTM is the smoothed true price from the previous time-step, i.e. the prediction is just trailing the ground truth and time series appears fairly straightforward with a clear weekly seasonality.

The tables from 1 to 8 and figure1 and 2 shows that the multivariate linear regression and random forest regression are the most precise out of the eight machine learning models based on correlation and RMSE, however, if we compare these results from table 9 and figure 3 and 4, we can clearly see that the long short term memory (LSTM) model is the most accurate of all the models tested in this paper for predicting the crypto index price values.

VI. CONCLUSION

This work has provided many methods for forecasting crypto currency index values and movement patterns over one week utilizing eight machine learning regression models and one recurrent neural network based model. We created, improved, and then evaluated values from April 1, 2021, to March 31, 2023. The performances of the multivariate regression and the random forest regression were the most accurate according to the findings of all machine learning models tested in this paper. However, if we compare these results from the results Long Short Term Memory (LSTM), we can clearly see that the LSTM model is the most accurate of all the models tested in this paper for predicting the crypto index price values.

REFERENCES

[1]Sen, J. and Datta Chaudhuri, T.: An Alternative Framework for Time Series Decomposition and Forecasting and its Relevance for Portfolio Choice - A Comparative Study of the Indian Consumer Durable and Small Cap Sector. Journal of Economics Library, 3(2), 303-326 (2016)

[2]Sen, J. and Datta Chaudhuri, T.: Decomposition of Time Series Data of Crypto currency Markets and its Implications for Prediction - An Application for the Indian Auto Sector. In Proc. of the 2nd Nat. Conf. on Advances in Business Research and Management Practices (ABRMP), Kolkata, India, pp. 15-28 (2016)

[3]Sen, J. and Datta Chaudhuri, T.: An Investigation of the Structural Characteristics of the Indian IT Sector and the Capital Goods Sector - An Application of the R Programming Language in Time Series Decomposition and Forecasting. Journal of Insurance and Financial Management, 1(4), 68-132 (2016)

[4]Sen, J. and Datta Chaudhuri, T.: A Time Series Analysis-Based Forecasting Framework for the Indian Healthcare Sector. Journal of Insurance and Financial Management, 3(1), 66-94 (2017)

[5]Sen, J. and Datta Chaudhuri, T.: A Predictive Analysis of the Indian FMCG Sector Using Time Series Decomposition-Based Approach. Journal of Economics Library, 4(2), 206 – 226 (2017)

[6]Sen, J.: A Time Series Analysis-Based Forecasting Approach for the Indian Realty Sector. Int. Journal of Applied Economic Studies, 5(4), 8 - 27 (2017)

[7]Sen, J.: A Robust Analysis and Forecasting Framework for the Indian Mid Cap Sector Using Time Series Decomposition. Journal of Insurance and Financial Management, 3(4), 1- 32 (2017)

[8]Sen, J. and Datta Chaudhuri, T.: Understanding the Sectors of Indian Economy for Portfolio Choice. Int. Journal of Business Forecasting and Marketing Intelligence, 4(2), 178-222 (2018)

[9]Sen, J. and Datta Chaudhuri, T.: A Robust Predictive Model for Crypto currency Price Forecasting. In: Proc. of the 5th Int. Conf. on Business Analytics and Intelligence, Bangalore, India, December 11-13 (2017)

[10]Sen, J.: Crypto currency Price Prediction Using Machine Learning and Deep Learning Frameworks. In: Proceedings of the 6th International Conference on Business Analytics and Intelligence, Bangalore, India, December 20 – 22 (2018)

[11]Mehtab, S. and Sen, J.: A Robust Predictive Model for Crypto currency Price Prediction Using Deep Learning and Natural Language Processing. In: Proceedings of the 7th Int. Conf. on Business Analytics and Intelligence, Bangalore, India, December 5 – 7 (2019)

[12]Jaffe, J., Keim, D. B. Keim, and Westerfield, R.: Earnings Yields, Market Values and Crypto currency Returns. Journal of Finance, 44, 135 - 148 (1989)

[13]Fama, E. F. and French, K. R.: Size and Book-to-Market Factors in Earnings and Returns. Journal of Finance, 50(1), 131 - 155 (1995)

[14]Chui, A. and Wei, K.: Book-to-Market, Firm Size, and the Turn of the Year Effect: Evidence from Pacific Basin Emerging Markets. Pacific-Basin Finance Journal, 6(3-4), 275-293 (1998)

[15]Basu, S.: The Relationship between Earnings Yield, Market Value and Return for NYSE Common Crypto currencies: Further Evidence. Journal of Financial Economics, 12(1), 129 – 156 (1983)

[16]Rosenberg, B., Reid, K., Lanstein, R.: Persuasive Evidence of Market Inefficiency. Journal of Portfolio Management, 11, 9 – 17 (1985)

[17]Geron, A.: Hands-on Machine Learning with Scikit-Learn Keras & Tensorflow, O'Reilly Publications, USA (2019)

[18]Brownlee, J.: Introduction to Time Series Forecasting with Python, (2019)

[19]Mehtab, S. and Sen, J.: Crypto currency Price Prediction Using Convolutional Neural Network on a Multivariate Time Series. In: Proc. of the 3rd Nat. Conf. on Machine Learning and Artificial Intelligence (NCMLAI), New Delhi, INDIA (2022)

[20]Kaggle Website: Crypto currency Market Data (2021 - 2023) | Kaggle