EasyChair Preprint

№ 10190

# Review of Machine Learning Techniques for Cryptocurrency Price Prediction

Shubham Bhattad, Stefin Sunnymon, Dallas Vaz and Chhaya Dhavale



May 17, 2023

# REVIEW OF MACHINE LEARNING TECHNIQUES FOR CRYPTOCURRENCY PRICE PREDICTION

Shubham Bhattad[1], Stefin Sunnymon[2], Dallas Vaz[3], Prof. Chhaya Dhavale[4]

[1,2,3,4] *Dept. of Information Technology, Xavier Institute of Engineering, Mumbai, India*

***Abstract: Cryptocurrency is a class of digital asset that is very challenging to monitor and forecast. Predicting cryptocurrency price action and its locus is difficult because it does not coincide with market movements. Our objective is to analyze the machine learning algorithms used in 9 researches and find out the best model which can be used to forecast the prices of time series models. In this work, we compared and analyzed earlier methodologies in which several machine learning models were applied to forecast the trend of cryptocurrency time series data. The outcomes support the machine learning models' ability to predict trends reasonably well. Making long-term predictions and generalizing them based on a small number of models yields low accuracy outputs for a highly volatile asset like cryptocurrency. We suggest using various models to fill this gap. With this method, we'll strive to identify the ideal machine learning algorithm for achieving the best accuracy with the lowest possible error rates.***

***Keywords: Machine Learning, Cryptocurrency, Time Series, Prophet, Arima, Xgboost, LSTM***

## I. INTRODUCTION

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such as image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more.

At a broad level, machine learning can be classified into three types:

i. Supervised learning
ii. Unsupervised learning
iii. Reinforcement learning

a) Supervised learning

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

b) Unsupervised learning

Unsupervised learning is a learning method in which a machine learns without any supervision.The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classified into two categories of algorithms:

i. Clustering
ii. Association

c) Reinforcement learning

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

The market capitalization of all cryptocurrencies is over $870.81 billion. Cryptocurrency is a digital currency form that uses cryptography for security and verifies transactions on the network. People are constantly looking for long-term investments that will yield a decent return on their investment. A Blockchain.com poll estimates that 83 million individuals possess cryptocurrency worldwide, which is a staggering amount. 27 million people, or around 2% of India's total population possess cryptocurrencies. Due to their volatility, cryptocurrencies are high risk investments. Due to this volatility, many individuals avoid the cryptocurrency market.

Investors would benefit greatly from having knowledge about the cryptocurrency they are investing in as well as the future trends of coins in order to make better decisions. There are several websites and newsletters that advise cryptocurrency investors on the best investments to make money, but most of them are for-profit. Having an algorithm would make it easier to analyze coins and make the prediction on adjusted closing price of coins with higher volume. In this approach, the users or investors may stop being terrified of the cryptocurrency market due to its volatility and make smarter judgments about when and how much to invest in a certain crypto coin.

Due to its recent price explosion and breakdown, Bitcoin has recently attracted a lot of media and public interest. But compared to Bitcoin, Ethereum has the second-highest market capitalization and offers significantly more functionality. Ethereum is thought to have been created by Vitalik Buterin, who in 2014 released a white paper introducing it. In 2015, Buterin and Joe Lubin, the creator of the blockchain software firm ConsenSys, introduced the Ethereum platform. Decentralized open-source blockchain platform Ethereum has its own coin called Ether. ETH serves as both a framework for the execution of decentralized smart contracts and a host of other cryptocurrencies.

The paper introduces machine learning models like the Prophet Model, ARIMA, LSTM, XGBOOST, SVM, Logistic Regression, Naive Bayes that may be used to predict the cryptocurrency coin's closing price in the future with accuracy and simple implementation. Data on the cryptocurrencies are initially gathered from the open source websites like yfinance, kaggle.

## II. RELATED WORK

Blockchain technology, which is the underlying framework of cryptocurrencies, has gained a lot of attention and trust because it provides secure transactions and fast data transfer. It also provides authentication of a product and can act as a contract. Investing in crypto currencies has been a challenge for most people due to its volatility.

According to the study in [1], volatility in the crypto market is quite high and it is difficult to predict the prices of the coins like bitcoin, ethereum etc. Analysis between August 30th, 2015 and October 19th, 2017 shows that Bitcoin had a monthly volatility of 21.73% and over that same time span, Ethereum had a monthly volatility of 77.91%. For comparison, the S&P 500 has a historical monthly volatility of about 14%. Paper [2] compares various deep learning-based Bitcoin price prediction models using Bitcoin blockchain information and uses deep learning models such as deep neural networks (DNN), long short-term memory (LSTM) models, convolutional neural networks (CNN) and their combinations.

[3] proposed a price prediction system of different cryptocurrencies using technical trade indicators and neural network algorithms achieving an accuracy of 94.89% under all circumstances of technical trade indication. [4] depicts the use of Recurrent Neural Network which uses the Long Short-Term Memory algorithms. This paper also calculated the Root Mean Square Error of the model which was found to be 3.38%. The main objective of [5] is to forecast the bitcoin price with improved efficiency using deep learning models and minimizing the risks for the investors as well as policy-makers. The paper mentioned two deep learning techniques such as LSTM and GRU as prediction models and the study reveals that the GRU model is the better mechanism for time series cryptocurrency price prediction and takes lower compilation time.

Paper [7] focuses on Bitcoin's future values using the PROPHET and ARIMA methods. According to the results in the paper, the PROPHET model makes predictions quite close to reality, that is up to 94.5% precision and the ARIMA model provides only 68% precision. The paper [8] proposes two time series Machine Learning models: ARIMA and Prophet model. According to the paper, both the ARIMA model and Prophet model have very similar performance with the ARIMA model having a slightly higher R-square score. R-square score for ARIMA is 94% and Prophet is 93%. However this paper also used the Prophet model for future predictions as the Prophet model is simpler and has easy to understand methods and workflow.

Paper [9] used the BTC dataset which contains the OHLC (Open High Low Close) data between the period of Jan 1st 2012 and Dec 31st 2020. In this paper four different models, namely LSTM, ARIMA, XGBoost and Facebook Prophet were used and the ARIMA Model emerges as the best model among the other three.

Table: Summary of Referred Research Papers

| Sr. No. | Author | Paper Name | Techniques/ Algorithms | Summary / Results |
|---|---|---|---|---|
| 1. | Matthew Chen, Neha Narwal and Mila Schultz | Predicting Price Changes In Ethereum (2017) | • Logistic Regression<br>• Naive Bayes,<br>• Support Vector Machines,<br>• Random Forest,<br>• ARIMA,<br>• Recurrent Neural Network,<br>• Neural Network | While all methods achieved above 50% accuracy, The best performance was achieved by the Auto Regressive Integrated Moving Average (ARIMA) model, which is attributed to its features and suitability to time-series data. Other methods fell short due to lack of data, assumptions made by the models about the data, and non-convergence of the models. |
| 2. | Suhwan Ji, Jongmin Kim and Hyeonseung Im | A Comparative Study Of Bitcoin Price Prediction Using Deep Learning (2019) | • Deep neural network,<br>• Long short-term memory,<br>• Convolutional neural network,<br>• Deep residual network | For regression problems, LSTM slightly outperformed the other models, whereas for classification problems, DNN slightly outperformed the other models unlike the previous literature on Bitcoin price prediction. Although CNN and ResNet are known to be very effective in many applications, including sequence data analysis, their performance was not particularly good for Bitcoin price prediction. Overall, there was no clear winner and the performance of all deep learning models studied in this work was comparable to each other. |

| | | | | |
|---|---|---|---|---|
| 3. | Mohammed khalid salman, Abdullahi Abdu Ibrahim | Price Prediction Of Different Cryptocurrencies Using Technical Trade Indicators And Machine Learning (2020) | • Neural network,<br>• Technical trade indicators | Results obtained from predicting bitcoin prices using machine learning based neural networks achieving an accuracy of 94.89% under all circumstances of trade indication. |
| 4. | Samiksha Marne, Delisa Correia, Shweta Churi, Joanne Gomes | Predicting Price Of Cryptocurrency - A Deep Learning Approach (2020) | • RNN<br>• LSTM | It was very evidently observed that the difference between the actual and predicted value is very minute. With every epoch and different ratios of datasets different variations of the graph can be extrapolated. The Root Mean Square Error (RMSE) calculated was 3.3% of the Testing data set. |
| 5. | Temesgen Awoke, Minakhi Rout, Lipika Mohanty, and Suresh Chandra Satapathy | Bitcoin Price Prediction And Analysis Using Deep Learning Models (2020) | • RNN<br>• LSTM<br>• GRU | The MSE value obtained for 7 days ahead from both the models is plotted, and it is clearly observed that GRU is converging faster and steady than the LSTM model. It is also discovered that the variation of actual price and predicted price is more in LSTM than the GRU. |
| 6. | Lekkala Sreekanth Reddy, Dr.P. Sriramya | A Research On Bitcoin Price Prediction Using Machine Learning Algorithms (2020) | • LSTM<br>• RMSE<br>• LASSO | The Linear regression model have more accuracy than the other algorithms. In this paper we conclude that the linear regression algorithm is more efficient than the other algorithms. |

| | | | | |
|---|---|---|---|---|
| 7. | Işıl Yenidogan, Aykut Çayır, C ¸i ˘gdem Arslan | Bitcoin Forecasting Using Arima And Prophet (2018) | • PROPHET<br>• ARIMA | While the PROPHET model makes predictions quite close to reality, that is up to 94.5% precision, the ARIMA model provides only 68% precision. |
| 8. | Gowtham Saini, Dr. M. Shobana | Cryptocurrency Price Prediction Using Prophet And Arima Time Series (2022) | • PROPHET<br>• ARIMA | Both the ARIMA model and Prophet model have very similar performance with the ARIMA model having a slightly higher R-square score. R-square score for ARIMA is 94% and Prophet is 93%. However, the future predictions were chosen to be predicted using the Prophet model because of the model's simplicity and easy to understand methods and workflow. |
| 9. | Yash Wadalkar, Yellamraju V H Sai Tarun, Jaiesh Singhal, Reena Sonkusare | A Comparative Analysis Based Approach For Bitcoin Price Forecasting (2021) | • LSTM<br>• XGBoost<br>• Prophet<br>• ARIMA | ARIMA Model is very accurate while making predictions for both short as well as long term prices. |

## III. COMPARATIVE ANALYSIS OF ALGORITHMS

*A. Prophet Model*

Prophet is a Facebook tool for ASCII text files that forecasts statistical information that aids companies in understanding and maybe forecasting the market. It provides a sophisticated additive model that takes into account the effects of vacations as well as non-linear trends and seasonality. It avoids some of the more serious problems of other methods and is noticeably reasonable at simulating statistics with various seasonalities.

There are two vital terms in Prophet Model:

i. Trend - The trend shows the tendency of the info to extend or decrease over a protracted amount of your time and it filters out the differences due to the season.

ii. Seasonality - Seasonality is the variations that occur over a brief amount of your time and isn't distinguished enough to be known as a "trend".

The “Prophet Equation” fits, as mentioned above, trend, seasonality and holidays. This is given by,

$$y(t) = g(t) + s(t) + h(t) + e(t)$$

where,

- g(t) is growth function and refers to trend (changes over a long period of time)
- s(t) is seasonality function and refers to seasonality (periodic or short term changes)
- h(t) is holiday function and refers to effects of holidays to the forecast
- e(t) is an error function and refers to the unconditional changes that are specific to a business or a person or a circumstance. It is also called the error term.
- y(t) is the forecast.

*B. Auto-Regressive Integrated Moving Average (ARIMA)*

ARIMA stands for autoregressive integrated moving average model and is specific by 3 order parameters: (p, d, q)

- AR(p) Autoregression - a regression model that utilizes the dependent relationship between a current observation and observations over a previous amount. AN automobile regressive (AR(p)) element refers to the utilization of past values within the equation for the statistic.
- I(d) Integration - uses differencing of observations (subtracting AN observation from observation at the previous time step) so as to form the statistical stationary. Differencing involves the subtraction of the values of a series with its previous values d range of times.
- MA(q) Moving Average - a model that uses the dependency between AN observation and a residual error from a moving average model applied to lagged observations. A moving average element depicts the error of the model as a mix of previous error terms. The order alphabetic character represents the quantity of terms to be enclosed within the mode

The ARIMA equation is a regression type equation in which the independent variables are lags of the dependent variable and/or lags of the forecast errors. The equation of the ARIMA model is given as :

$$y'(t) = c + \phi 1 * y'(t-1) + \cdots + \phi p * y'(t-p) + \theta 1 * \varepsilon(t-1) + \cdots + \theta q * \varepsilon(t-q) + \varepsilon t$$

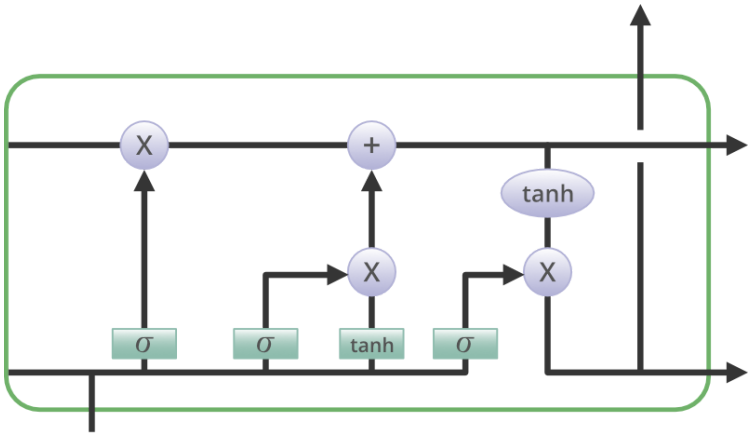
Types of ARIMA Model

- ARIMA:Non-seasonal Autoregressive Integrated Moving Averages
- SARIMA:Seasonal ARIMA
- SARIMAX:Seasonal ARIMA with exogenous variables

*C. Long Short Term Memory (LSTM)*

LSTM networks square measure associate degree extension of continual neural networks (RNNs) principally introduced to handle things wherever RNNs fail. It's a network that works on the current input by taking into thought the previous output (feedback) and storing in its memory for a brief amount of your time (short-term memory). LSTM has been designed in order that the vanishing gradient downside is nearly fully removed, whereas the coaching model is left unedited. A very long time lags in bound issues square measure bridged victimization LSTMs wherever they conjointly handle noise,

distributed representations, and continuous values. With LSTMs, there's no ought to keep a finite variety of states from beforehand PRN within the hidden mathematician model (HMM). With LSTMs, North American nations have access to a wide range of parameters, including learning rates and input and output biases. There is hence no need for fine modifications. With LSTMs, the difficulty of updating each weight is decreased to O(1). LSTM has a chain structure with four neural networks and a variety of cell types for memory units.




Information is retained by the cells and the memory manipulations are done by the gates. There are 3 gates:

i. Input Gate.
ii. Forget Gate.
iii. Output Gate.

*D. XGBoost*

XGBoost is AN implementation of Gradient Boosted call trees. This library was written in C++. It's a specific class of package library made with speed and model performance in mind. Recently, it has dominated the field of applied machine learning. Call trees are generated using this algorithmic technique in serial fashion. Weights are essential to XGBoost. All or all of the independent variables are given weights before being put into the decision tree that forecasts outcomes. The second call tree then receives the doubled load of incorrectly predicted variables from the first.

To estimate the value of a variable quantity, XGBRegressor employs a variety of gradient boosted trees, which are referred to as n estimators in the model. This is frequently accomplished by joining call trees, each of which are weak learners, to produce an integrated strong learner. When expressing a statistic, the model forecasts for a range of improvements using what is known as a lookback amount. As an illustration, if a lookback amount of one is used, the X train (or freelance variable) forecasts future values using lagged values of the statistic regressed on the statistic at time t (Y train).

*E. Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a potentially simple supervised machine learning technique that may be used for regression and/or classification. Although it is frequently used for classification, regression also benefits greatly from it. In essence, SVM identifies a hyper-plane that establishes a boundary

between the different knowledge types. In SVM, we typically plot each knowledge item in the dataset in an N-dimensional space, where N is the number of knowledge features and attributes. Realize the optimum hyperplane for information separation next. So by this, you should already be aware of it.
In SVM, only binary classification will be used (i.e., make a choice from 2 classes). However, there are several methods that may be used for multi-class problems. Support vector machines for problems with multi-class The implicit mapping of their inputs into high-dimensional feature regions allows SVMs to conduct a non-linear classification with efficiency.

*F. Logistic Regression*

Logistic regression is basically a supervised classification algorithm. For a certain collection of characteristics (or inputs), X, the target variable (or output), y, can only take discrete values in a classification issue.
Despite what many people think, logistic regression IS a regression model. In order to determine the likelihood that a certain data input falls into the category designated by the number "1," the programme creates a regression model. Logistic regression models the data using the sigmoid function, much like linear regression assumes that the data follows a linear distribution.

Logistic Regression is given by the formula: $e(z) = 1/(1+e^{-z})$

Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The classification problem itself determines the threshold value, which is a crucial component of logistic regression. The precision and recall levels have a significant impact on the choice of the threshold value. In a perfect world, precision and recall should both equal 1, but this is rarely the case.

*G. Naive Bayes Algorithm*

Naive Bayes classifiers square measure a group of classification algorithms supported by Bayes' Theorem. it's not one rule however a family of algorithms wherever all of them share a typical principle, i.e. each combination of options being classified is freelance of every different. Bayes' Theorem finds the likelihood of an incident occurring given the likelihood of another event that has already occurred. Bayes' theorem is explicit mathematically because the following equation:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

where A and B are events and $P(B) \neq 0$.
Basically, we are trying to find the probability of event A, given that event B is true. Event B is also termed as evidence.
P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

## IV. PROPOSED SOLUTION

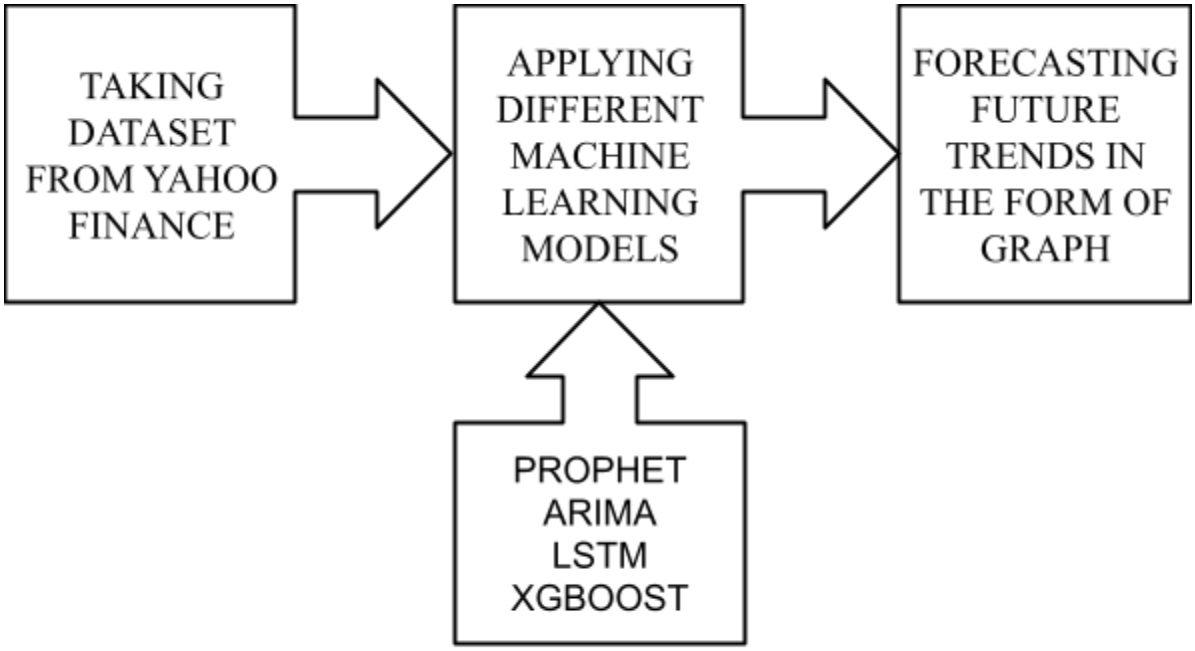



Fig: Proposed Model

## V. CONCLUSION

The review of research papers on the use of machine learning and deep learning algorithms showed that these techniques/algorithms are very helpful in predicting the future prices of the coins and will be helpful for the investors to mitigate their risk. After researching and analyzing all the models like LSTM, XGBoost, Prophet, SVM, Naive Bayes and ARIMA in the various papers, we came to a conclusion that the prophet model outperforms all the other models and has the highest accuracy. Therefore, for future predictions in our project we are choosing the Prophet model because of the model's simplicity and easy to understand methods and workflow and high accuracy. The top few cryptocurrencies coins were analyzed, and it was found that Tether has the biggest transaction volume, but its closing price has constantly been $1. Bitcoin had the second-highest volume and the highest closing price, followed by Ethereum, which corresponds with what actually happened.

In further advancements to this research, we got to know that only price trends don't affect the future prices of cryptocurrencies. So the value depends on various other social, environmental, political and other elements. Capturing the governing elements would improve prediction accuracy of the various coins. For instance, accessing information from Twitter's trending page would provide us insights into how the general public feels about Bitcoin, which would immediately affect demand for it and drive up its price globally. Analyzing the posts on the Reddit site would be helpful in a similar way since it would enable us to go further and understand how the public and investors feel about the market. By

incorporating these insights into the training of models, efficiency and accuracy may be further improved and the investors risk can be mitigated to a great extent.

## VI. REFERENCES


[1] Matthew Chen, Neha Narwal and Mila Schultz, (2017) “Predicting Price Changes in Ethereum”, Stanford University, Stanford, CA 94305

[2] Ji, Suhwan & Kim, Jongmin & Im, Hyeonseung, (2019) “A Comparative Study of Bitcoin Price Prediction Using Deep Learning.”, Mathematics. 7. 898. 10.3390/math7100898.

[3] Mohammed Khalid Salman and Abdullahi Abdu Ibrahim, (2020) “Price Prediction Of Different Cryptocurrencies Using Technical Trade Indicators And Machine Learning” IOP Conf. Ser.:Mater. Sci. Eng. 928 032007

[4] Samiksha Marne, Delisa Correia, et al, (2020) “Predicting Price Of Cryptocurrency - A Deep Learning Approach”, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181

[5] Muniye, Temesgen & Rout, Minakhi & Mohanty, et al, (2020) “Bitcoin Price Prediction and Analysis Using Deep Learning Models.”, Springer 10.1007/978-981-15-5397-4_63.

[6] Lekkala Sreekanth Reddy, Dr. P. Sriramya, (2020) “A Research On Bitcoin Price Prediction Using Machine Learning Algorithms”, International Journal Of Scientific & Technology Research Volume 9, Issue 04, April 2020 ISSN: 22778616

[7] I. Yenidoğan, A. Çayir, O. Kozan, T. Dağ and Ç. Arslan, (2018) "Bitcoin Forecasting Using ARIMA and PROPHET," 3rd International Conference on Computer Science and Engineering (UBMK), 2018, pp. 621-624, doi: 10.1109/UBMK.2018.8566476.

[8] Gowtham Saini, Dr. M. Shobana, (2022) “Cryptocurrency Price Prediction Using Prophet And Arima Time Series”, International Research Journal of Modernization in Engineering Technology and Science e-ISSN: 2582-5208

[9] Yash Wadalkar, Yellamraju V H Sai Tarun, et al, (2021) “A Comparative Analysis Based Approach For Bitcoin Price Forecasting”, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181