

Block Mining reward prediction with Polynomial Regression, Long short-term memory, and Prophet API for Ethereum blockchain miners.

Jeyasheela Rakkini Simon^{1*}, Geetha K²

¹ School of Computing, SASTRA Deemed University, TamilNadu.

² School of Computing, SASTRA Deemed University, TamilNadu.

Abstract. The Ethereum blockchain is an open-source, decentralized blockchain with functions triggered by smart contract and has voluminous real-time data for analysis using machine learning and deep learning algorithms. Ether is the cryptocurrency of the Ethereum blockchain. Ethereum virtual machine is used to run Turing complete scripts. The data set concerning a block in the Ethereum blockchain with a block number, timestamp, crypto address of the miner, and the block rewards for the miner are explored for K means clustering for clustering miners with a unique crypto address and their rewards. Linear regression and polynomial regression are used for the prediction of the next block reward to the miner. The Long Short-Term Memory (LSTM) algorithm is used to exploit the Ether market data set for predicting the next ether price in the market. Every kind of price and volume for every four hours is taken for prediction. The root mean square error of 34.9% is obtained for linear regression, the silhouette score is 71% for K-means clustering of miners with same rewards, with the optimal number of clusters obtained by Gap statistic method.

1 Introduction

Blockchain is a decentralized ledger of transactions with the traits of decentralization, transparency, and anonymity. Ethereum 2.0 has the proof of stake consensus algorithm, which has energy efficiency, high economic incentives, and opens up high revenue-generating capability for a large number of users. The validator nodes add blocks to Proof of Stake (POS) Ethereum blockchain [1] by staking their cryptocurrency, ether on the network and anyone validator can be randomly selected to propose a block. The other validators agree with this validator who is proposing the block. The reward is for both attesting the other validators when they are proposing the block and for the self proposal of the block also. Their own money or ether is at stake when they behave maliciously and this is quite in contrast to the hectic energy-consuming proof of work consensus protocol in the bitcoin blockchain. If the validator does not propose a block when his turn comes or fails to do attestations when other validators propose a block then he is penalized for that also. The validator's stake amount will be fully debited and the validator will be sent out from the proof of stake network if he performs any re-org attack of transactions. The

proof of stake consensus algorithm does not require the miners to invest more in hardware, the initial hardware setup costs, the electricity bills but the monetary stake amount, 32 ether (ETH) has to be staked by the Ethereum miner for active participation as a validator in Ethereum mining process. For those miners who find 32 ETH, a significant amount to be staked upon, then they can join staking pools which aggregate the funds of miners and the resultant rewards are distributed proportionally. There is a decentralization trend in PoS mining since anyone can participate with minimized electricity costs and this opens up revenue opportunities for many people. PoS achieves both scalability and security with sharding of the blockchain.

2 Related Work

The quintessential details about the proof of stake consensus algorithm for the Ethereum blockchain is given in [1]. Qin, R. et al. [2] discusses the speculation of individual miners in joining mining pools. Due to increased hardware cost, electricity, and maintenance costs, the chances or the probability of the individual miners who can successfully mine a block is very less, since they

* Corresponding author: jeyasheelarakkini@cse.sastra.edu

have to produce a hash value early and once their hash value is less than or equal to target hash produced by the system, they gain rewards. So, the miners join mining pools and contribute their hash to the overall mining pool. The choice of the mining pools or choosing the right mining pool to join and go ahead with the mining is not a trivial one and this is discussed by the authors. This is modelled as a risk decision process. The pool manager has to distribute the reward got from successfully mining to all miners who have joined the mining pool. The mechanisms for distributing the reward include the pay per share mechanism, pay per last N share mechanism, and the flat proportional mechanism. This paper discusses the computational result for a decision of selecting a mining pool.

Qin, R., Yuan et al. [3] discusses the three levels of participants in the bitcoin mining, namely mining pools, individual miners, and blockchain systems. The author lists the issues based on artificial societies, computational experiments, and parallel systems (ACP) and proposes a research framework. Easley et al. [4] discuss the strategy adopted by miners, the significant role of mining rewards, transaction fees, price, waiting time, and external constraints that influence the participation of users in the blockchain. Kiayias, A. et al. [5] discusses the two simplified forms of the game wherein the first game the miners mine a block and release it immediately, in the second form of the game the miners mine a block, and they announce it but defer to release it. In both forms of the game, the best response and the Nash equilibrium is demonstrated. The best response whether it matches the expected behaviour of the bitcoin designer is introspected. Wang, W. et al. [6] discusses the decentralized nature of the blockchain from the perspective of the individual nodes who form the backbone network of the blockchain, the decentralized consensus that runs on these nodes and they have surveyed the incentive mechanisms and byzantine fault tolerance mechanisms from a game theory perspective.

Li, J. et al. [7] discusses the users bidding transaction fees for faster confirmation of transactions. Generalized Second Price auction method is used for the bidding process. The metrics that are considered are transaction size, the bidding scores submitted by the user, the scores of quality, and the virtual fees of the transactions. These metrics play a vital role in reducing and saving the user's fees. Saad M. et al. [8] explores the features in the bitcoin blockchain network and Ethereum blockchain and how they change over time and how they relate to the demand and supply economics of both the cryptocurrencies. Jang, H. et al. [9] discusses the highly correlated network indicators that are useful in predicting the bitcoin price with

high accuracy. A Bayesian neural network is used to predict the price of cryptocurrency that takes in to account a sudden decrease in the hash rate that leads to delay in block publishing time and reduction in network throughput. Poongodi M. et al. [10] exploits the ARIMA model for bitcoin data set which is publicly available on the URL <http://coinmarketcap.com> for 4 consecutive years to analyze the volatility of bitcoin and how it affects the miners joining mining pools and dropping their solo mining. Catalini, C. et al. [11] discusses the two metrics such as the cost of verification of transactions and the cost of networking, with the cost of verification of transactions related to the cost of verifying the state of transactions along with their attributes, their previous history, and their current ownership of a digital asset. The cost of networking relates to the cost of operating a decentralized digital market place. Salimitari, M. et al. [12] discusses the group of miners joining a mining pool and giving their share of electricity for mining, and three concepts of dynamic game theory such as Social optimum, Nash equilibrium, and myopic Nash equilibrium is proposed. Vimal, S. et al. [13] discusses the efficiency of transferring p2p files in a decentralized interplanetary file system and awarding the miners who successfully shared their resources.

Gupta, S.S et al. [14] discusses clustering methods are used for fraud detection in cryptocurrency financial transactions. Chawathe, S. S. et al. [15] explores the clustering algorithm for the blockchain data and its impact on miners. The chapter from the book, *mastering-bitcoin* [16], refers to the process of transactions being validated by miners, a new block formed ever since last block creation, 10 minutes ago in blockchain, and the miners mint new coins for each new block creation and the transaction fees from all transactions that are included in the block. The miner who mines for a block by fetching transactions from a mining pool and by solving the target hash given by the bitcoin system or any blockchain system, that is the proof of work system has the crypto address. The miners have to find a hash that is less than or equal to the target hash and the first miner to find it gets the reward in bitcoins. Here we will consider an Ethereum blockchain. But to our surprise for the data posted in <http://xblock.pro/xblock-eth.html> [17], a simple k means of clustering the mining rewards based on the crypto address of the users for each progressing block number, a linear regression model for prediction of the next block reward, the polynomial regression model for the prediction of the next block reward for the miner has not yet been analyzed in the previous works. Here we have also taken a dataset with stock market prediction of ether

from the same url and we have done prediction with LSTM algorithm and Prophet API for ether with ether market data.

3 Problem Definition

Clustering of miner’s crypto address, their rewards per unique miner address and the prediction of miner’s rewards in Ethereum blockchain has not been addressed in previous works and we have implemented with the existing K-means clustering of miner’s crypto address and rewards. The linear regression and Polynomial regression are used for the prediction of miner’s rewards. The LSTM algorithm, Prophet API for prediction of Ether crypto currency in Ether stock market. Here the Ethereum blockchain which has Proof of Stake consensus mining, with Ether currency is first analyzed for mining rewards with the dataset in the website [17] that has attributes, miner crypto address, number of transactions, and the rewards. K means clustering of the miner’s crypto address with rewards is implemented and the rewards for ‘n’(n=3) optimal number of clusters is implemented below in Section 4. The centroids of the cluster are found, the elbow method of K means clustering, and the scatter plot of the clusters is shown with the linear regression line fitted for the scattered values is shown. The polynomial regression is implemented for the same with the prediction for the next block reward for the miner crypto address and values. The LSTM method and the Facebook Prophet API for the prediction of the stock price of Ether is used here. These methods are plotted and analyzed with statistical measures.

4 Results and discussion

We have implemented a K mean clustering in python, given three attributes, timestamp, miner crypto address, and block rewards in the dataset [17]. Then we did a polynomial regression and compared it with the normal linear regression. The Ordinary Least Square (OLS) regression model fit is shown below in Fig.3. The data is scattered and the accuracy of the model with the statistics is shown below. The timestamp for the miner, the crypto address of the miner with the rewards for successfully mining is shown in the head of the data frame with five values in Fig 1.

```

timestamp          miner  VALUES  reward
0 1438269988 0x05a56e2d52c817161883f50c441c3228cfe54dd9f 58 5.000000e+18
1 1438270017 0xdd2f1e6e498202e86d8f5442af596580a4f03c2c 59 5.000000e+18
2 1438270048 0x5088d623ba0fcf0131e0897a91734a4d83596aa0 60 5.160000e+18
3 1438270048 0xc0ebccc5f5689fa8659d83713341e5ad19349448 61 3.750000e+18
4 1438270077 0x5088d623ba0fcf0131e0897a91734a4d83596aa0 60 3.130000e+18
(9999, 4)
timestamp  int64
miner      object
VALUES     int64
reward     float64
    
```

Fig 1. The head of the Ethereum block dataset.

The crypto address being an object data type, the other attributes such as timestamp, VALUES being integer values and reward being float value respectively. The data set is acquired from the block dataset present in <http://xblock.pro/xblock-eth.html>[17].

The summary of the data frame is given below in Fig 2. with the distribution of the data frame columns, count (number of non-null observations), mean, standard deviation min, max, 25% (lower), 50% (median), and 75% (upper) of values. The min+(max-min) * percentile is used for calculating percentiles per column values.

```

count  timestamp  VALUES  reward
mean   1.438286e+09  124.072107  4.807288e+18
std    1.330892e+04  70.835563  5.698741e+17
min    1.438270e+09  1.000000  1.250000e+18
25%   1.438275e+09  68.000000  5.000000e+18
50%   1.438282e+09  97.000000  5.000000e+18
75%   1.438296e+09  148.000000  5.000000e+18
max    1.438317e+09  356.000000  5.310000e+18
    
```

Fig 2. Summary of the block dataset.

Here we have replaced the miner’s crypto address with a unique integer value for each unique crypto address, so the clustering with the unique integer per miner’s crypto address is implemented seamlessly. The linear regression results are shown below in Fig 3, which has the dependent variable “reward”, the R-squared value, the measure of the symmetry of the data about the mean, F- statistic, T- statistic, P-values, Durbin-Watson measure for autocorrelation and Jarque-Bera measure of skewness.

```

=====
OLS Regression Results
=====
Dep. Variable:  reward  R-squared:  0.013
Model:  OLS  Adj. R-squared:  0.013
Method:  Least Squares  F-statistic:  47.39
Date:  Sat, 03 Oct 2020  Prob (F-statistic):  4.40e-21
Time:  16:23:32  Log-Likelihood:  -2.9599e+05
No. Observations:  6999  AIC:  5.928e+05
Df Residuals:  6996  BIC:  5.928e+05
Df Model:  2
Covariance Type:  nonrobust
=====
coef    std err    t    P>|t|    [0.025    0.975]
-----
const   -7.696e+21  7.93e+20  -9.705  0.000  -9.25e+21  -6.14e+21
timestamp  5.354e+12  5.91e+11  9.711  0.000  4.27e+12  6.44e+12
VALUES    -4.131e+14  1.03e+14  -4.020  0.000  -6.15e+14  -2.12e+14
=====
Omnibus:  3741.634  Durbin-Watson:  1.966
Prob(Omnibus):  0.000  Jarque-Bera (JB):  23011.179
Skew:  -2.593  Prob(JB):  0.00
Kurtosis:  10.213  Cond. No.:  1.70e+14
=====
    
```

Fig 3. OLS Regression summary.

The linear regression line fit is shown below in Fig 4. with x-axis as values (integer assigned to unique crypto addresses) and y-axis as rewards of the miners.

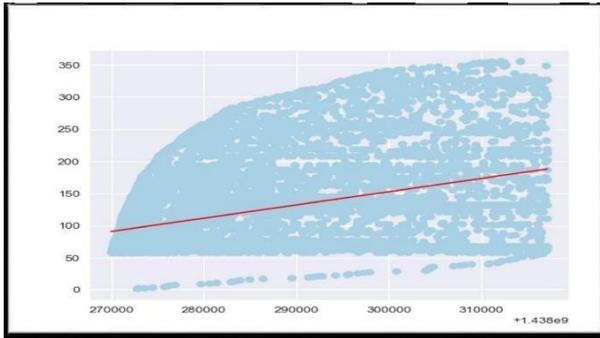


Fig 4. Linear regression fit in a scatter plot.

```
[3.83485401]
[[0.]]
Mean Absolute Error: 0.3014282015259374
Mean Squared Error: 0.12183620349642657
Root Mean Squared Error: 0.34905043116493434
```

Fig 5. Linear regression accuracy metrics.

Then we have done a cluster with K-means for three clusters, to find miners with similar rewards, this is used for analyzing the miners with a similar hash rate of computation. The three clusters are formed with red, green, and cyan colours and are represented in Fig 6.

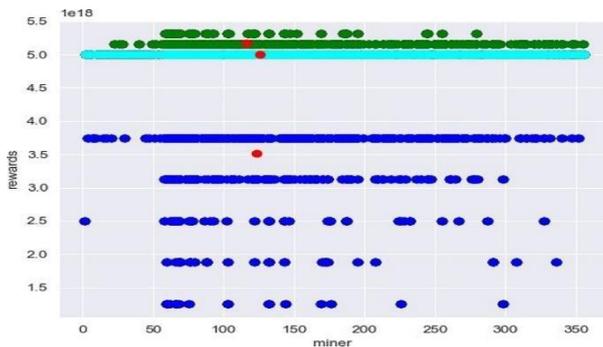


Fig 6. The three clusters.

The cluster of three is formed with values shown below in Fig 7. The number of clusters is user-defined. The centroid of each cluster is determined and the distance of the observations (which are values, rewards) from the cluster is calculated and they are grouped in appropriate clusters.

Cluster	reward	VALUES
1	3.510048e+18	123.567680
2	5.173063e+18	115.926426
3	5.000000e+18	125.676271

Fig.7. Three clusters with reward and values.

The elbow method of k means is shown below in Fig 8. As the number of clusters increases, sum of squared distances decreases in the elbow method.

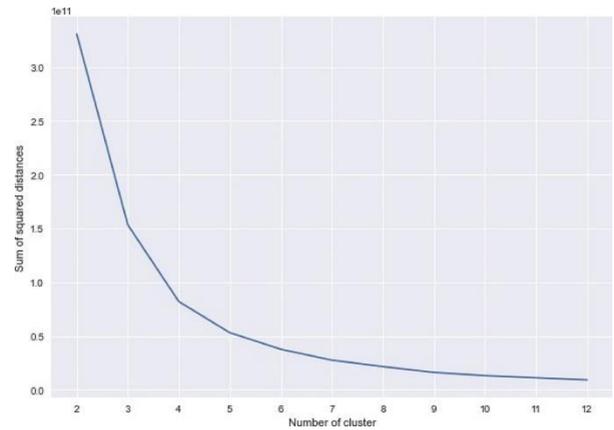


Fig 8. Elbow method of K-means clusters for Ether mining rewards.

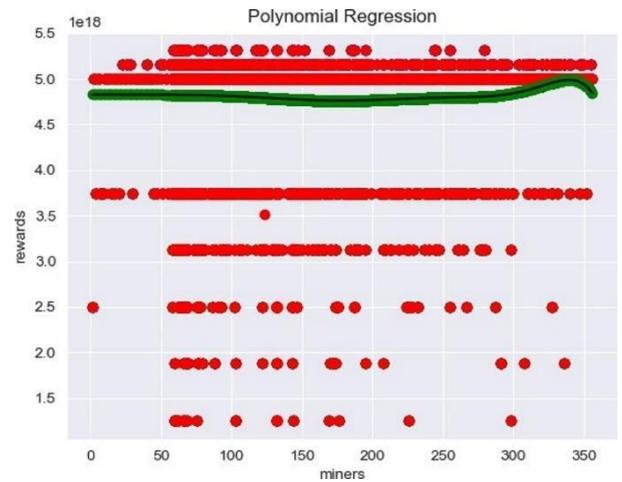


Fig 9. Polynomial Regression with n='3'.

The result of the polynomial regression is shown above in Fig 9. We noticed there was skewness in data and so we did a skew transformation as shown in Fig 10, Fig 11, and Fig 12. The miner's address was more skewed towards the left and the rewards were skewed towards the right. After the removal of the skewness of data with normalization, we have implemented polynomial regression results with a degree of 10.

```
-----
'timestamp' had 'positive' skewness of 0.72
Transformation yielded skewness of 0.00
-----
'VALUES' had 'positive' skewness of 1.26
Transformation yielded skewness of -0.03
-----
'reward' had 'negative' skewness of -2.65
Transformation yielded skewness of nan
In [119]: transformedDF.head(5)
Out[119]:
timestamp  VALUES  reward
0  0.000539  3.420461  3.700765e+152
1  0.000539  3.432457  3.700765e+152
2  0.000539  3.444234  4.792648e+152
3  0.000539  3.455799  3.489676e+151
4  0.000539  3.444234  7.916962e+150
```

Fig 10. Skewness transformed data set.

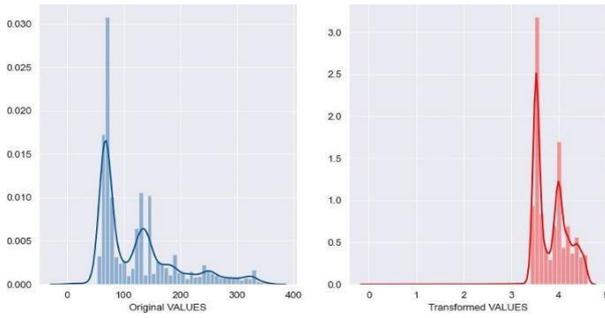


Fig 11. Skewness transformation of the VALUES in Ethereum dataset.

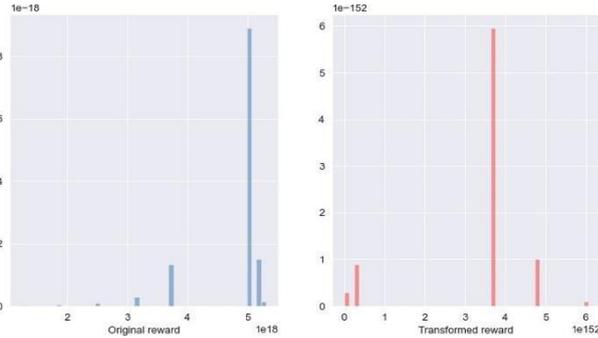


Fig 12. Skewness transformation of the reward attribute in Ethereum dataset.

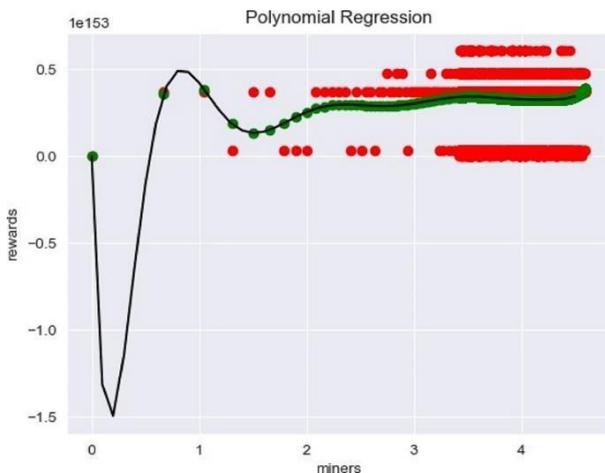


Fig 13. Polynomial Regression for Prediction of next Ethereum block reward.

The root mean square error for polynomial regression is $1.3368300755199987e+152$. R2 score, which is a measure of how close are the data to the fitted regression line is 0.005977411651757691 . The Polynomial regression with the degree of 10 is shown in Fig 13. The curvilinear fit of the polynomial regression fits the line of best fit so that the best approximation between the independent variable and the dependent variable is implemented.

4.1 Ether price prediction with LSTM

The data set we acquired from <http://xblock.pro/mt/> had the following attributes, close: the close price in the period, date: the timestamp at the beginning of this period, high: the

highest price in the period, low: the lowest price in the period, open: the open price in the period, quote Volume: the quote volume in the period, base volume: the base volume in the period and weighted Volume: the average price for those base volume and quote volume. The plot of the original values is shown in Fig 14 with x-axis as time in days and y-axis as the price of the Ether., and the prediction with LSTM for the future values with the given observed values is shown in Fig 15. The original dataset, training data set and testing data set is shown in the graph in Fig 15.

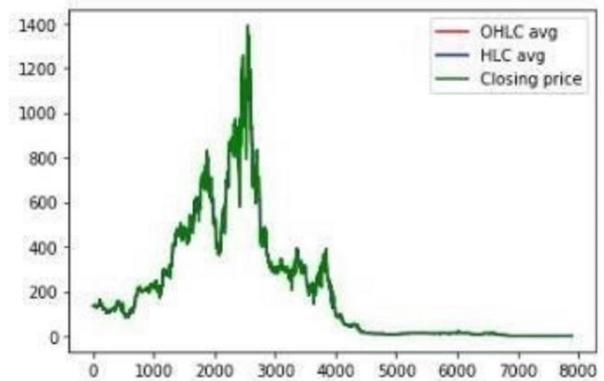


Fig 14. Plot of original values.

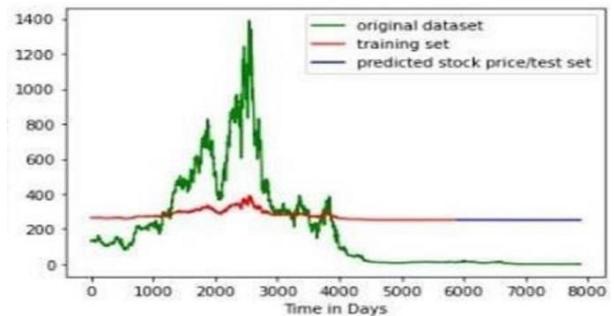


Fig 15. LSTM Prediction with training set and testing set

The LSTM model has predicted lower values than the original values of ether, cryptocurrency with themean square error of 3645 %. The Ether market prices in the future are predicted by Facebook Prophet API in Fig 16. The column 'ds' has dates for which prediction has to be done using this API. The predicted value is shown in column 'yhat'. 'yhat_lower' and 'yhat_upper' represent uncertainty intervals.

	ds	yhat	yhat_lower	yhat_upper
1799	2021-04-12	4521.662794	-30957.902882	40698.046442
1800	2021-04-13	4741.287733	-28437.709043	39877.217167
1801	2021-04-14	4709.327598	-29577.932056	39545.338922
1802	2021-04-15	5000.149948	-29537.530429	42966.354571
1803	2021-04-16	4619.096229	-28157.431980	39048.772218

Fig 16. Prophet API prediction of Ether prices in the future.

The x-axis has the 'ds' column with the year as values and y-axis have ether, cryptocurrency values. The prediction of the ether prices, the trend of the ether price, weekly and yearly is shown in Fig. 17 and Fig. 18 respectively. The trend of the Ether prices, yearly seasonality, weekly seasonality is shown with Prophet API for the time series. The future data frame for a time series of 365 days is taken. Here prediction is done on a data frame with the date and closing values.

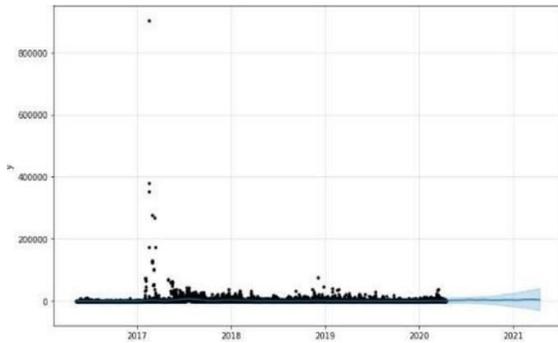


Fig 17. Prophet API prediction of Ether prices.

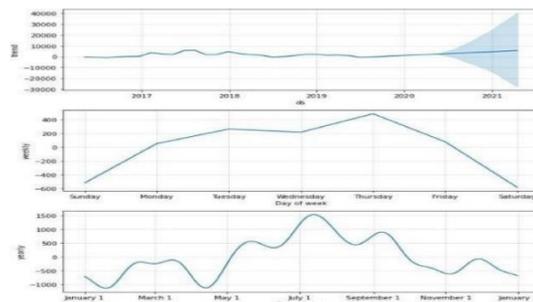


Fig 18. Forecast components yearly, monthly and trend-wise for Ether price.

Thus, Prophet API for forecasting components and future ether values prediction is implemented in python language. The mean absolute error of 1326.71 is obtained with Prophet API. The prediction of Ether price, which is a Time Series problem, is also done with LSTM.

5 Conclusion and Future work

Thus, we have devised a k means clustering algorithm for clustering the unique addresses of the miners based on their rewards. The linear regression and polynomial regression to implement prediction of the mining reward for the next block with the next sequential block number and time period is implemented. The LSTM model to predict the stock price of ether is also implemented. The future work would be the implementation of pattern mining algorithms, centralization versus decentralization traits in Ethereum mining pools with deep learning models.

References

- <https://www.coindesk.com/proof-of-stake>.
- Qin, R., Yuan, Y., & Wang, F. Y. (2018). Research on the selection strategies of blockchain mining pools. *IEEE Transactions on Computational Social Systems*, 5(3), 748-757.
- Qin, R., Yuan, Y., Wang, S., & Wang, F. Y. (2018, June). Economic issues in bitcoin mining and blockchain research. In *2018 IEEE Intelligent Vehicles Symposium (IV)* (pp. 268-273). IEEE
- Easley, D., O'Hara, M., & Basu, S. (2019). From mining to markets: The evolution of bitcoin transaction fees. *Journal of Financial Economics*, 134(1), 91-109.
- Kiayias, A., Koutsoupias, E., Kyropoulou, M., & Tselekounis, Y. (2016, July). Blockchain mining games. In *Proceedings of the 2016 ACM Conference on Economics and Computation* (pp. 365-382).
- Wang, W., Hoang, D. T., Hu, P., Xiong, Z., Niyato, D., Wang, P & Kim, D. I. (2019). A survey on consensus mechanisms and mining strategy management in blockchain networks. *IEEE Access*, 7, 22328-22370.
- Li, J., Yuan, Y., & Wang, F. Y. (2019). A novel GSP auction mechanism for ranking Bitcoin transactions in blockchain mining. *Decision Support Systems*, 124, 113094.
- Saad, M., Choi, J., Nyang, D., Kim, J., & Mohaisen, A. (2019). Toward characterizing blockchain-based cryptocurrencies for highly accurate predictions. *IEEE Systems Journal*, 14(1), 321-332.
- Jang, H., & Lee, J. (2017). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *Ieee Access*, 6, 5427-5437.
- Poongodi, M., Vijayakumar, V., & Chilamkurti, N. (2020). Bitcoin price prediction using ARIMA model. *International Journal of Internet Technology and Secured Transactions*, 10(4), 396-406.
- Catalini, C., & Gans, J. S. (2016). *Some simple economics of the blockchain* (No. w22952). National Bureau of Economic Research.
- Salimitari, M., Chatterjee, M., Yuksel, M., & Pasilio, E. (2017, October). Profit maximization for bitcoin pool mining: A prospect theoretic approach. In *2017 IEEE 3rd international conference on collaboration and internet computing (CIC)* (pp. 267- 274). IEEE.
- Vimal, S., & Srivatsa, S. K. (2019). A new cluster p2p file sharing system based on ipfs and blockchain technology. *Journal of Ambient Intelligence and Humanized Computing*, 1-7.
- Gupta, S. S. (2017). *Blockchain*. John Wiley & Sons, Inc.

15. Chawathe, S. S. (2019). Clustering blockchain data. In *Clustering Methods for Big Data Analytics* (pp. 43-72). Springer, Cham.
16. <https://www.oreilly.com/library/view/mastering-bitcoin/9781491902639/ch08.html>.
17. <http://xblock.pro/xblock-eth.html>.