International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-05, Issue-09, Dec 2019

Ethereum Blockchain Analytics using Machine learning ARIMA model

[*]P.Pavan Kumar, [#]Deepika Gowlikar

[*,#]Lecturer, GCIT, RUB, Bhutan, [*]pavan2312@gmail.com , [#]deepika.gowlikar@gmail.com

Abstract - Blockchain can be considered as a distributed, cryptographic and decentralized database for storing and retrieving information. Blockchain transactions are immutable, which means that the data contained in it can not be deleted by any means and will continue to grow with time. The massive volume of Blockchain data can be analyzed to gain valuable insights into the future. This paper presents the predictive analytics of Ethereum blockchain, by studying the growth of Ethereum blocks on a daily basis. We visualize, predict and forecast future trends of Ethereum using AutoRegressive Integrated Moving Average (ARIMA) model of machine learning. We have chosen ARIMA model as it uses time series data efficiently to forecast the future points of the time series and also helps to understand the nature of the dataset.

Keywords: ***ARIMA, Blockchain, Ethereum, Forecast, Machine learning, Predictive analytics.***

I. INTRODUCTION

In recent years, Blockchain technology is becoming increasingly predominant as "the internet of value". It is totally revolutionizing various sectors like financial services, banking, government, healthcare, information technology, and many others. Gartner Inc. forecasts that amalgamation of blockchain into any bussiness can enhance its value upto $176 billion by 2025[1]. Today small and large corporations are adopting blockchain to integrate into their businesses, adding huge amounts of data to the blockchain platform. Hence blockchain can be viewed as a distributed data repository that can be analyzed to reveal valuable insights into the behavioral patterns of data. The enormous volume of data stored by blockchain can be used by data science or big data to predict future outcomes.

While data science focuses on harnessing data for making predictions, blockchain ensures the integrity of data by maintaining a decentralized ledger. Blockchain can equip data science with valid and structured data. Blockchain has high scope in analytics. Contemporary businesses have integrated data analytics into their business models for many years now. According to Forbes, enterprise usage of data analytics has increased from 17% in 2015 to 59% in 2018[2]. Predictive analytics is a type of data analytics that is gaining momentum in the present time. It concentrates on predicting future events based on past data using machine learning tools.

In this paper, the time series modeling of Ethereum data is done using the ARIMA model of machine learning. We have chosen the ARIMA model, as it is extensively used and regarded as the most efficient forecasting technique in time series which has a linear dependence on its past values. [3]. For our study, we have retrieved data from Ethereum mainnet and analyzed the trends of basic blockchain data element – blocks, over a period of 4 years (2015 – 2019, till September).

The rest of the paper is organized as follows. Section II gives a brief overview of Ethereum platform, smart contracts, and ARIMA model. Section III discusses the datasets of Ethereum blockchain. Section IV shows the implementation of the ARIMA model through machine learning techniques. Section V presents the prediction analysis and forecast of Etherem platform. Section VI outlines the related work in Blockchain technology. Finally, Section VII concludes the paper.

II. BACKGROUND

A. *Ethereum Blockchain*

Like any other Blockchain platform, Ethereum is also a common record of the complete transaction log with every node on the network storing a copy of the log. Ethereum primarily stores the most recent state of each smart contract along with the other transaction details. Blocks are added to the chain by the miner nodes with cryptographic proofs and timestamps. Once a miner has validated a block, it receives a reward of an Ethereum cryptocurrency asset called *Ether*, as an incentive for providing its computational resources to the network [4].

Ether is used to transfer funds between different addresses on the network and also to compensate for *Gas*. Gas is used by the each node in the network for each transaction and also for each smart contract creation. Thus gas fuels the Ethereum network. The size of contract or transaction determines the amount of Gas needed.

B. *Ethereum Smart Contracts*

A smart contract is a self-executing computer program that

 DOI : 10.35291/2454-9150.2019.0548

can ease exchange of money or anything of value [5]. Termed as *smart*, these contracts can automatically implement their code when specific conditions are met. Smart contracts are created using contract-based, high-level programming language - Solidity. Solidity uses EVM (Ethereum Virtual Machine) to compile contracts written in it, and to generate byte code, which is executed on the Ethereum nodes by the EVM instance running on each of them [6].

C. *AutoRegressive Integrated Moving Average (ARIMA) model*

Time series is a series of readings recorded at regular time intervals for a particular variable. ARIMA is a statistical method that works on the idea that the historic data of a variable can be used to forecast its future. This model was first formulated by George Box and Gwilym Jenkins in 1970 and hence this model is also popularly known as Box-Jenkins Model [7]. This model works on a given time series based on its own past values (lags) and the lagged forecast errors.

The ARIMA model is based on the terms: p, d, q where,

- p is the order of the AR term
- q is the order of the MA term
- d is the number of differencing required to make the time series stationary

If a time series, has seasonal patterns, then we need to add seasonal terms P,D,Q and it becomes SARIMA (Seasonal ARIMA) [8].

The ARIMA model consists of the following stages:
Stage 1: Identification of appropriate model
Stage 2: Parameter estimation
Stage 3: Validation of the model
Stage 4: Implementation of the model selected.

III. DATASETS

We have collected data from the Ethereum mainnet over a period of 4 years (2015 – 2019, September). The data was retrieved using an open-source, REST-based API called **eth.events**.

This API provides functionalities to search, filter and group events from multiple Ethereum based blockchains. It is based on Elasticsearch and PostgreSQL [9]. We queried the Ethereum mainnet to get data on the number of blocks being created per day.

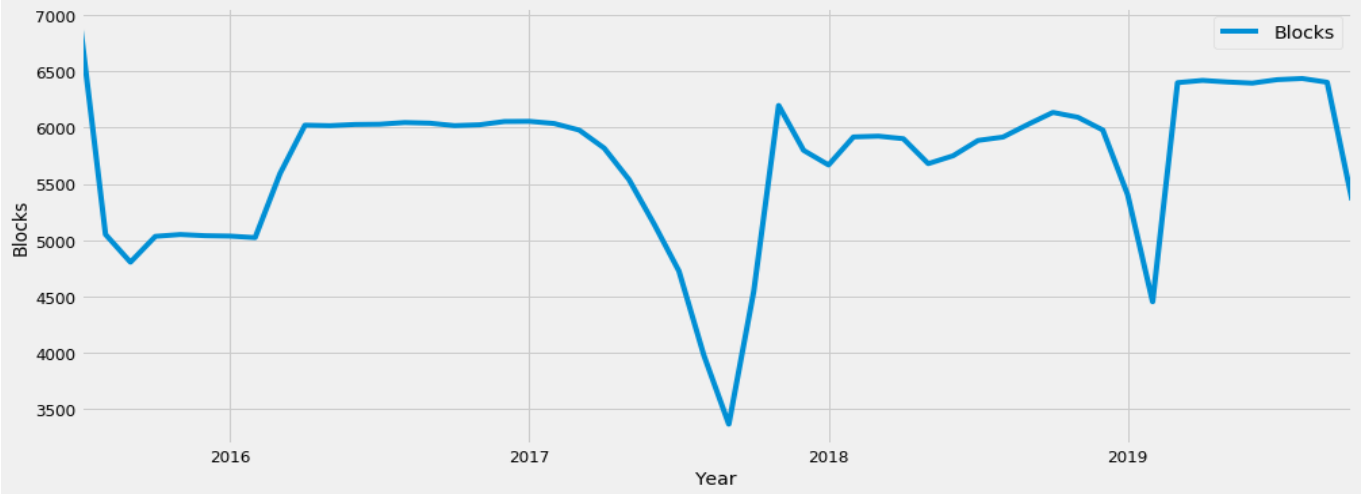

Figure 1: Data Visualization of number of Blocks retrieved from Ethereum mainnet during the period of 2015, July – 2019, September

IV. METHODOLOGY

A. *Model Identification*

In ARIMA(p, d, q) model, the three parameters p, q, d are explained as follows:

- p signifies the autoregressive part of the ARIMA model. It shows the effect of past values (lags) in the model. İt stands for the number of lags to be used as predictors.
- d is the integration part of the ARIMA model. It shows the amount of differencing to be applied to the time series to make it stationary. If d = 0 then time series is stationary.
- q is the moving average part of the ARIMA model. It allows us to set the number of lagged forecast errors in the model.

In Seasonal-ARIMA(p,d,q)(P, D, Q)s, the parameters p, d, q are the non-seasonal parameters, and P, D, Q are the seasonal parameters of the time series. The term s is the periodicity of the time series(12 for a yearly period).

B. *Parameter selection*

We have used grid search to iteratively generate different combinations of parameters- p,d,q,P,D,Q,s. We fit these parameters in the SARIMAX() function from the statsmodels module in Python 3.5 and select the combination with lowest AIC value. AIC (Akaike Information Criterion) value is used to measure how well a model fits the data while taking into account the overall

 DOI : 10.35291/2454-9150.2019.0548

complexity of the model. A model that has large AIC value fits the data using lot of features, where as the model with lower AIC value achieves same goodness of fit using less features. Therefore, we are select the model that yields the lowest AIC value.

We got the following output of SARIMAX() in our experiment:

SARIMAX(0, 0, 0)x(0, 0, 0, 12)12 - AIC:1028.971259996795
SARIMAX(0, 0, 0)x(0, 0, 1, 12)12 - AIC:791.2262825439599
SARIMAX(0, 0, 0)x(0, 1, 0, 12)12 - AIC:656.5070738127782
. .
SARIMAX(1, 1, 1)x(1, 0, 1, 12)12 - AIC:583.5025432466327
SARIMAX(1, 1, 1)x(1, 1, 0, 12)12 - AIC:425.44840654520755
SARIMAX(1, 1, 1)x(1, 1, 1, 12)12 - AIC:399.2115874229374

SARIMAX(1, 1, 1)x(1, 1, 1, 12) returns the lowest AIC value of 399.21. Therefore we consider this to be an optimal choice out of all the models.

(a) Fitting the ARIMA model: Using the optimal parameter values into SARIMAX model produces the following diagnostics:

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3903	0.214	1.826	0.068	-0.029	0.809
ma.L1	-0.9975	3.494	-0.286	0.775	-7.845	5.850
ar.S.L12	-0.0718	0.404	-0.178	0.859	-0.863	0.719
ma.S.L12	-0.8691	2.798	-0.311	0.756	-6.353	4.614
sigma2	2.478e+05	1.07e+06	0.231	0.817	-1.85e+06	2.35e+06

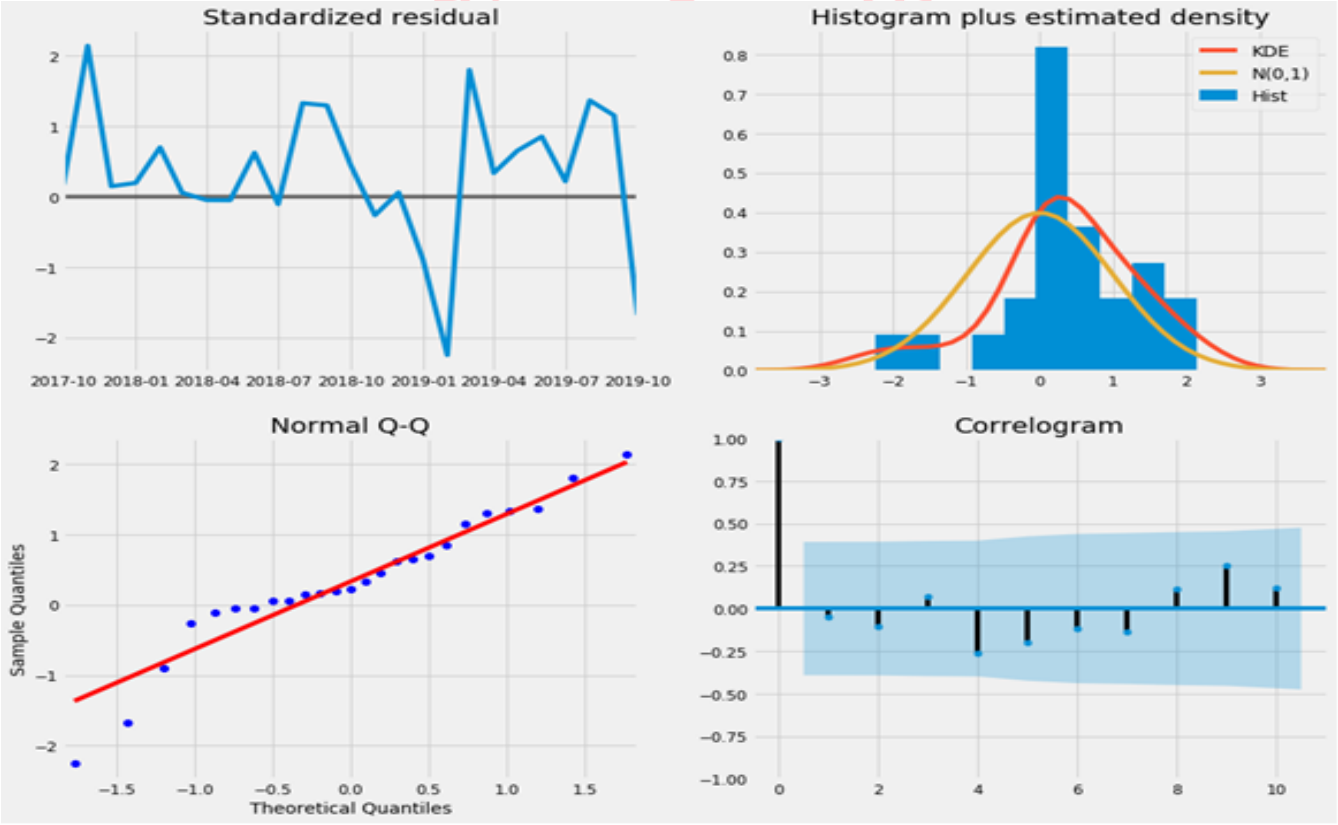

Figure 2: Statistical diagnostic plot of data

C. *Validation of model*

Augmented Dickey-Fuller (ADF) Test is one of the widely used statistical methods to test whether the time series is stationary or non-stationary [10]. The p-value from the test interprets the data as follows:

p-value > 0.05: data is non-stationary.

 DOI : 10.35291/2454-9150.2019.0548

p-value ≤ 0.05: data is stationary.
For our data we obtained the followng statistics using adfuller() function of statsmodels module in Python 3.5.

Test Statistic	-3.394412
p-value	0.011151
#Lags Used	1.000000
Number of Observations Used	1527.000000
Critical Value (1%)	-3.434640
Critical Value (5%)	-2.863435
Critical Value (10%)	-2.567779

since p-value < 0.05 our data is stationary and can used for furthur analysis.

D. *Implementing the model*

(a) Validating Forecast: To assess the accuracy of our forecast, we compared the predicted values to real values of the time series. We have set the forecast to start at 2019-01-01 to the end of the data (2019-09-30). Overall, our forecasts align with the true values very well, showing an overall increase in the trend.

The get_prediction() and conf_int() functions of statsmodels module in Python 3.5 help us to obtain the values for the forecast of the time series.

Figure 3: Number of Blocks forecast to start at 2019–01–01 to 2019-09-30

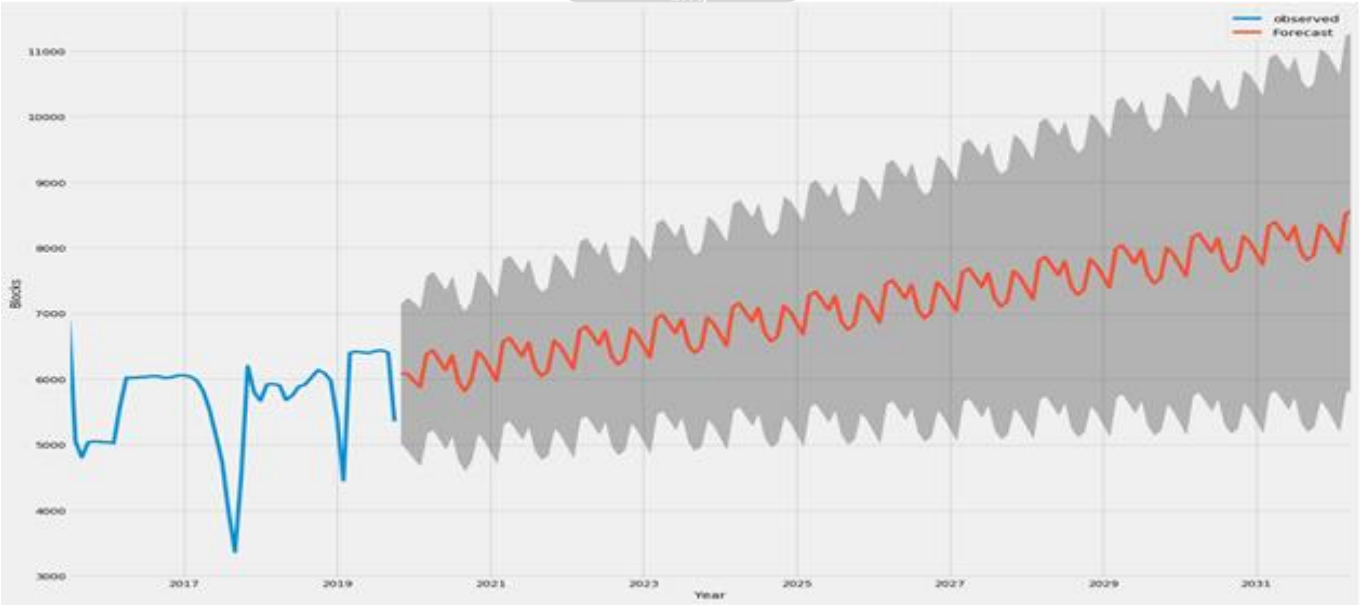

Figure 4: Future forecast of number of blocks in Ethereum mainnet

 DOI : 10.35291/2454-9150.2019.0548

(b) Future Forecast: The get_forecast() attribute of the time series object can calculate forecasted values for a specified number of steps ahead. Our forecasts show that the number of blocks in Ethereum mainnet is expected to increase at a steady pace of 33.33% by the year 2030.

(c) Measuring the accuracy of the prediction: The Root Mean Square Error (RMSE) conveys that our model was able to forecast the average number of blocks in the test set within 708.75 of the real number of blocks. The Mean Squared Error of our forecasts is 502321.1 and the Root Mean Squared Error is 708.75

V. RESULTS

The application of ARIMA model to the Ethereum mainnet data shows that in the next 11 years (2030), Ethereum blocks will continue to grow steadily with growth rate of 33.33%. Thus predicting that Ethereum may become most substantially used blockchain platform.

VI. RELATED WORK

Acheampong [11] discusses how the Blockchain technology could be used to impact Machine Learning systems. To generate good models in Machine Learning large amount of data is needed. Greater volumes of data support more generalized conclusions and effective prediction models. Blockchain can support Machine learning models by providing huge volumes of shared and trusted data. Because of its decentralized data sharing ability, Blockchain can equip Machine Learning with a voluminous amount of data at practically no expense. Blockchain data can help in building more efficient and reliable Machine learning systems at a very low cost.

Akcora et al. [12] discuss various analytical methods and tools that can be applied to Blockchain data to improve financial applications. The authors prescribe the k-chainlets model to analyze the Blockchain network. The model utilizes subgraphs of the Blockchain graph as the building blocks. Using this model all the Blockchain transactions can be modeled in a lossless manner. The chainlets can be used as the predictors of the cryptocurrency price. Other use-cases of Blockchain data analytics are e-crime detection, identifying illegal transactions involving human trafficking, money laundering, ransomware, etc.

VII. CONCLUSION AND FUTURE WORK

Predictive analytics of Ethereum data shows that, it will continue to rise in future with growth rate of 33.33%. In this paper, we studied the data dynamism of Ethereum using the datasets such as blocks, created in the last 4 years. We have observed that since its inception in July 2015, Ethereum is rising remarkably and may become one of the most highly used blockchain platforms. The comparison and parameterization of the ARIMA model have been done using AIC. The accuracy of the model has been verified using the root mean square method. In the future, we plan to predict the growth rate of Ethereum smart contracts and transactions.

REFERENCES

[1] Gartner_Inc "Forecast: Blockchain Business Value, Worldwide, 2017-2030", in *Gartner*, https://www.gartner.com/en/documents/3627117, 2017, Accessed 15 Sep 2019.

[2] L. Columbus, "Big Data Analytics Adoption Soared In The Enterprise In 2018", in *Forbes*, https://www.forbes.com/sites/louiscolumbus/2018/12/23/big-data-analytics-adoption-soared-in-the-enterprise-in-2018/#54e06996332f, 2018, Accessed 16 Sep 2019.

[3] A. A. Adebiyi, A. O. Adewumi, C. K. Ayo, "Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction", in *Journal of Applied Mathematics*, vol.2014, Mar.2014

[4] A. Hertig, "What is Ether?" in *CoinDesk*, https://www.coindesk.com/information/what-is-ether-ethereum-cryptocurrency, Accessed 18 Sep 2019.

[5] A. Rosic, "What is Ethereum? The Most Comprehensive Guide Ever!", in *Blockgeeks*, https://blockgeeks.com/guides/ethereum, Accessed 20 Sep 2019.

[6] R. Molecke, "How To Learn Solidity: The Ultimate Ethereum Coding Tutorial", in *Blockgeeks*, https://blockgeeks.com/guides/solidity, Accessed 20 Sep 2019.

[7] J. Brownlee, "A Gentle Introduction to the Box-Jenkins Method for Time Series Forecasting", in *Machinelearningmastery*, https://machinelearningmastery.com/gentle-introduction-box-jenkins-method-time-series-forecasting, 2017, Accessed 25 Sep 2019.

[8] "ARIMA Model – Complete Guide to Time Series Forecasting in Python", in *machinelearningplus*, https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python, 2019, Accessed 24 Dec 2019.

[9] "Welcome to eth.events's documentation!", in *docs.eth.events*, https://docs.eth.events/en/latest, 2018, Accessed 24 Dec 2019.

[10] "Augmented Dickey-Fuller Test in Python", in *insightsbot*, http://www.insightsbot.com/blog/1MH61d/augmented-dickey-fuller-test-in-python, 2018, Accessed 20 Oct 2019.

[11] F.A. Acheampong, "Big Data, Machine Learning and the BlockChain Technology: An Overview", in *International Journal of Computer Applications*, vol.180, no.20, Mar. 2018.

[12] C. G. Akcora, M.F. Dixon, Y.R. Gel, and M. Kantarcioglu, "Blockchain Data Analytics", in *Journal of IEEE Intelligent Informatis*, 2019.

 DOI : 10.35291/2454-9150.2019.0548