International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 04 | Apr 2025 www.irjet.net p-ISSN: 2395-0072

Facebook User Prediction in India Using Time Series ARIMA Forecasting

Sintu Nitharwal[1*], Dr. Deepa Mordia[2]

[1] Research Scholar, Department of Statistics, University of Rajasthan, Jaipur, Rajasthan,India(302004) (Orcid ID: https://orcid.org/0009-0003-8426-247X)*
[2] Assistant Professor, Department of Statistics, University of Rajasthan, Jaipur, Rajasthan, India(302004) (Orcid ID: https://orcid.org/0009-0007-9433-1220)

Abstract

This study uses statistical and econometric techniques to analyze trends, stationary, and autocorrelation in the time series data of Facebook user growth from April 2009 to September 2024. The PACF (Partial Autocorrelation Function) and ACF(Autocorrelation Function) are analyzed to determine seasonal patterns and to evaluate stationary patterns using the Augmented Dickey-Fuller (ADF) test. The best model for predicting was determined to be ARIMA(2,1,2). The residual diagnostics, such as the Ljung-Box and Box-Pierce tests, which reveal no appreciable autocorrelation in the residuals, verify the suitability of the model. We estimate Facebook user growth using this model for the ensuing two years, and we verify the model's dependability by analyzing residuals.

Key words: Time series modeling, ARIMA, BIC (Bayesian Information Criterion), Augmented Dickey-Fuller, AIC (Akaike Information Criterion), Ljung-Box and Box-Pierce tests.

Introduction

Facebook introduced a new era of social connectivity to a fast-expanding online audience when it entered the Indian market in 2006. At first, the platform was only appealing to tech-savvy metropolitan consumers, but as Smartphone penetration rose and internet connectivity got more reasonably priced, it soon gained popularity. Facebook appealed particularly to Indians, who place a great emphasis on social ties, because of its easy-to-use interface and capacity to connect with friends and family. Facebook further increased its popularity by tailoring its platform to India's different demographics by emphasizing local content and regional languages.

Facebook has become a major factor in India's digital economy over time, moving beyond social networking. It gave businesses new tools to use, allowing even tiny and medium-sized organizations to reach larger markets. By investing in Jio Platforms in 2020, Facebook strengthened its connections with the Indian market and increased its sway over the country's digital environment. With millions of active users, Facebook still influences social interaction, business, and even political debate in India today, connecting people from all over the huge country.

Table 1 represents the 16-year monthly Facebook users in India. The data was taken from "Statcounter GlobalStats" as a secondary source from April 2009 to September 2024.

Table 1: Facebook Users In India (In millions)

Date	Facebook	Date	Facebook	Date	Facebook	Date	Facebook	Date	Facebook	Date	Facebook
2009-04	13.4	2011-12	89.58	2014-08	92.13	2017-04	95.86	2019-12	82.53	2022-08	54.72
2009-05	14.8	2012-01	89.99	2014-09	94.14	2017-05	96.37	2020-01	83.56	2022-09	53.09
2009-06	16.06	2012-02	84.31	2014-10	94.58	2017-06	94.30	2020-02	82.11	2022-10	57.60
2009-07	19.87	2012-03	72.85	2014-11	94.73	2017-07	93.36	2020-03	84.94	2022-11	56.58

2009-08	23.00	2012-04	72.28	2014-12	95.64	2017-08	90.36	2020-04	86.04	2022-12	58.32
2009-09	27.98	2012-05	71.29	2015-01	95.67	2017-09	91.21	2020-05	90.06	2023-01	59.83
2009-10	24.69	2012-06	73.82	2015-02	95.19	2017-10	88.38	2020-06	89.16	2023-02	67.75
2009-11	34.07	2012-07	79.72	2015-03	95.44	2017-11	86.37	2020-07	86.75	2023-03	64.82
2009-12	54.66	2012-08	80.72	2015-04	95.53	2017-12	85.02	2020-08	75.93	2023-04	59.75
2010-01	52.33	2012-09	76.59	2015-05	95.65	2018-01	85.30	2020-09	79.30	2023-05	54.15
2010-02	47.89	2012-10	77.77	2015-06	95.97	2018-02	82.98	2020-10	75.27	2023-05	54.15
2010-03	44.67	2012-11	79.11	2015-07	96.06	2018-03	81.52	2020-11	74.18	2023-06	53.35
2010-04	64.05	2012-12	80.24	2015-08	94.35	2018-04	83.49	2020-12	77.26	2023-07	55.46
2010-05	65.29	2013-01	78.35	2015-09	91.77	2018-05	84.2	2021-01	82.53	2023-08	64.42
2010-06	73.06	2013-02	79.40	2015-10	92.58	2018-06	77.20	2021-02	82.55	2023-09	64.79
2010-07	77.71	2013-03	76.49	2015-11	91.94	2018-07	77.08	2021-03	84.18	2023-10	66.26
2010-08	75.25	2013-04	78.38	2015-12	93.37	2018-08	76.87	2021-04	82.19	2023-11	70.59
2010-09	75.44	2013-05	81.44	2016-01	94.50	2018-09	83.83	2021-05	77.58	2023-12	77.98
2010-10	79.56	2013-06	81.69	2016-02	95.43	2018-10	86.56	2021-06	75.81	2024-01	76.32
2010-11	80.66	2013-07	78.03	2016-03	95.84	2018-11	88.23	2021-07	78.33	2024-02	76.97
2010-12	80.81	2013-08	81.03	2016-04	95.81	2018-12	89.28	2021-08	59.56	2024-03	70.64
2011-01	83.84	2013-09	81.84	2016-05	95.80	2019-01	90.45	2021-09	64.16	2024-04	69.66
2011-02	82.31	2013-10	84.57	2016-06	95.16	2019-02	91.97	2021-10	63.83	2024-05	72.19
2011-03	88.01	2013-11	85.57	2016-07	94.38	2019-03	90.29	2021-11	68.83	2024-06	69.31
2011-04	92.27	2013-12	79.54	2016-08	95.81	2019-04	87.62	2021-12	68.61	2024-07	66.66
2011-05	88.74	2014-01	78.99	2016-09	96.37	2019-05	91.00	2022-01	65.32	2024-08	62.36
2011-	86.62	2014-	79.69	2016-	97.18	2019-	90.36	2022-	57.37	2024-	61.34

06		02		10		06		02		09	
2011-07	82.69	2014-03	80.06	2016-11	97.29	2019-07	89.29	2022-03	62.75		
2011-08	84.22	2014-04	80.86	2016-12	96.87	2019-08	89.79	2022-04	67.41		
2011-09	84.58	2014-05	81.20	2017-01	96.64	2019-09	88.05	2022-05	62.67		
2011-10	80.23	2014-06	81.05	2017-02	96.79	2019-10	83.80	2022-06	55.94		
2011-11	83.82	2014-07	88.16	2017-03	96.27	2019-11	84.68	2022-07	58.50		
Source: https://gs.statcounter.com/social-media-stats/all/india											

ARIMA Model (Box-Jenkins)

A time series is a collection of data that has been observed throughout time. Based on historical data for a single variable, a family of models called ARIMA models may produce accurate forecasts and characterize both stationary and non-stationary time series. In contrast to other forecasting models, This one doesn't make any assumptions about the time series data that need to be forecasted. The Box-Jenkins process uses the following steps to generate ARIMA models:

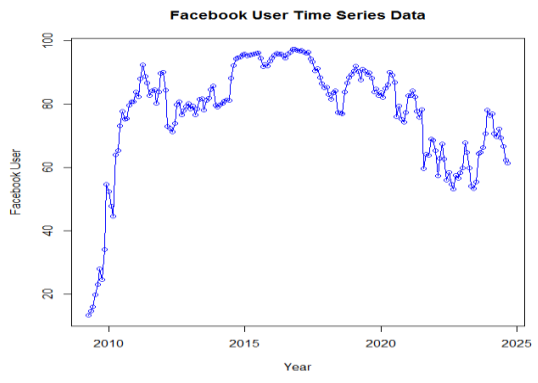
(1) Model recognition,
(2) Parameter selection and estimation,
(3) Diagnostic testing (sometimes called modal validation), and
(4) Model implementation.

To identify the model, it is necessary to determine the orders (p, d, and q) of the AR and MA components. In essence, the question it asks is: Is the data stationary or non-stationary? Which order of differentiation (d) causes the time to become stationary?

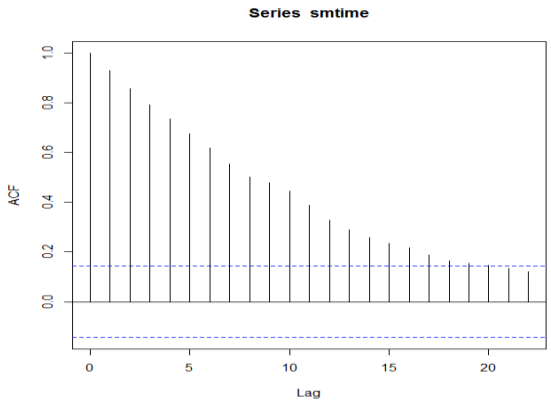
Statistical Analysis

For this study, several statistical and time series packages, including "tseries" and "forecast" are utilized in addition to other conventional programs. The open-source statistical program "RStudio" (version 3.0.1) is also utilized.

To create a forecasting model, the data set in Table 1 is utilized. Graph 1 below shows the line graph of Indian Facebook users.



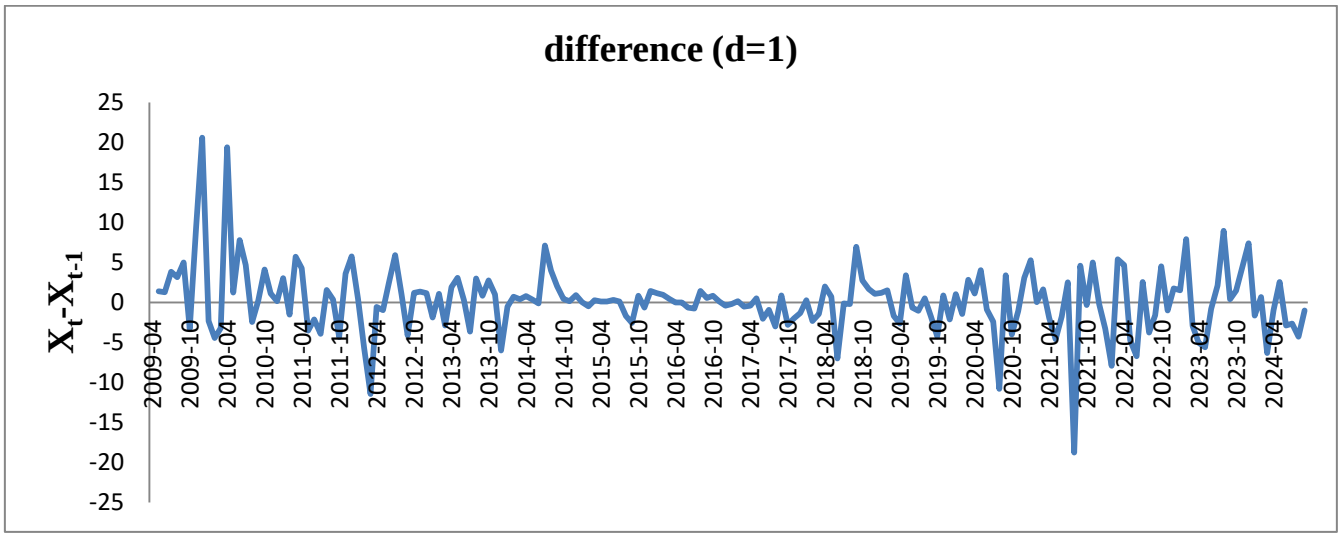
Graph 1: Facebook Users (In Millions) in India from 2009 to 2024



Graph 2: ACF in time series

Graph 1 indicates a rise in trend, and Graph 2 shows high autocorrelation in time series data. Thus, the data set of Facebook users is not stationary. To apply the ARIMA model firstly we convert this series into stationary. We determine the minimum order differencing (d=1) and test for the unit root issue.

$$\Delta X_t = X_t - X_{t-1}$$



Graph 3: Line plot of first-order differences in Facebook user data (d=1)

The graph above (Graph 3) makes it clear that the time series seems to have a stationary mean and variance. However, before continuing, To check if the differenced time series data is stationary (unit root problem), we will utilize the augmented Dickey-Fuller test(ADF).

ADF Test:

We test null and alternative hypothesis at $\alpha = 0.05$ level of significance:

H_o : Non-stationary time series data are present.

H_1 : Stationary time series data are present.

ADF Test Equation:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \epsilon_t$$

The ADF test result:

Dickey-Fuller = -4.1546, Lag order = 5, p-value = 0.01

Here, $p - value < \alpha$.

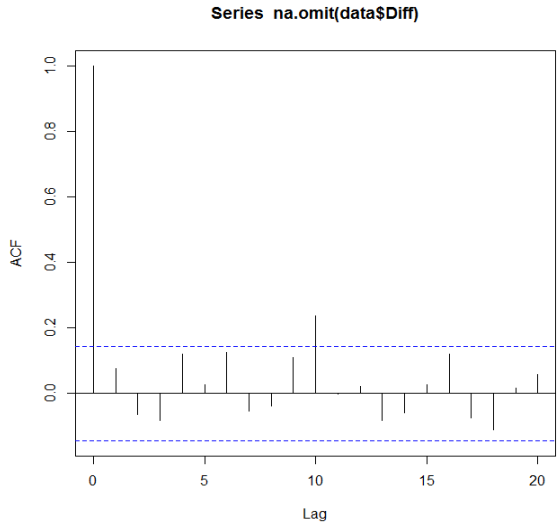
The H_o is rejected. Thus, we concluded that the Stationary time series data are present.

With the help of this test, we can continue developing the ARIMA model by determining appropriate values for p in AR and q in MA. The stationary (d=1) time series correlogram and partial correlogram must be examined in order to do that.

Correlogram and Partial Correlogram

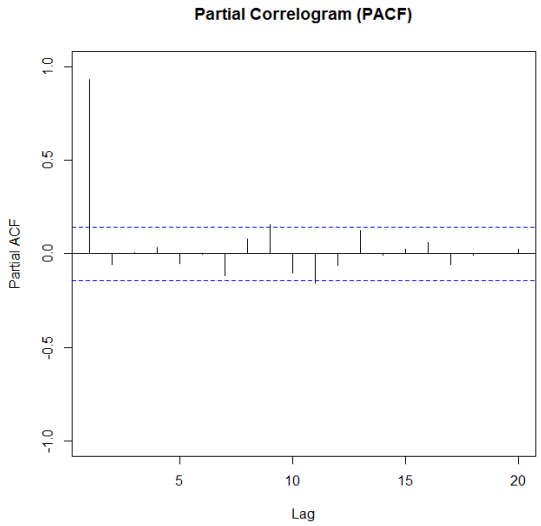
Table 2 below shows the "ACF and PACF coefficients" for lags 1 through 20 of that first-order differenced series.

Table 2: Values of ACF and PACF for Lags 1–20					
Lag	**ACF**	**PACF**	**Lag**	**ACF**	**PACF**
0	1.0000	-	11	-0.0016	-0.23155
1	0.0763	0.80792	12	0.021924	-0.23947
2	-0.06418	-0.17480	13	-0.0821	0.07785
3	-0.08237	0.04081	14	-0.05859	0.01971
4	0.120769	-0.05864	15	0.02813	0.01638
5	0.027199	-0.13624	16	0.120042	-0.03524
6	0.125946	0.09084	17	-0.07332	-0.15727
7	-0.05247	-0.18431	18	-0.11174	-0.11371
8	-0.03758	0.15385	19	0.016844	0.09458
9	0.110355	0.20031	20	0.058213	-0.11408
10	0.237579	-0.14482			



Graph 4

Graph 4 illustrates the autocorrelation function graphic for lags ranging from 1 to 20 in the d=1 time series of Indian Facebook users. ACF at lag 10 (ACF=0.237579) is displayed, and it barely surpasses the significance bounds. We can presume that the lag 10 autocorrelation is erroneous and a result of pure chance.



Graph 5

The PACF graphic for lags ranging from 1 to 20 in the d=1 time series of Facebook users in India is shown in Graph 5. Here also we have two outliers at lag 9 and 11. Since all other PACFs from lag 2 to 20 fall inside the significant limits, this could just be a random occurrence.

Given that the partial PACF and the ACF tail off to zero after lag 2 and lag 3, respectively, after omitting the outlier, we can advise the following possible ARMA (autoregressive moving average) models for the d=1 time series data of Facebook users in India:

1. An autoregressive model of order p=2, or ARMA(2,0), as both the autocorrelation and the partial autocorrelation are zero after lag 2.
2. A MA model of order q=3, or ARMA(0,3) model, since the autocorrelation and partial autocorrelation are both zero after lag 3.
3. Because both partial autocorrelation and autocorrelation have tails, an ARMA(p,q) model, also known as a mix model, has p and q greater than Zero.

As a result, under the following criteria, we are limited to the 3 ARIMA(p,d,q) models that are tentative:

ARIMA(p, d, q): ARIMA(2,1,2), ARIMA(2,1,1), and ARIMA(2,1,0)

Table 3

ARIMA Model	Coefficients				σ^2 (Estimated)	Likelihood	AIC	BIC	AICc
	AR1	AR2	MA1	MA2					
ARIMA(2,1,0)	0.0848	-0.0661			4.1023	523.6487	1053.297	1062.958	1053.363
ARIMA(2,1,1)	0.2735	-0.0865	-0.1894		4.1004	-523.563	1055.126	1068.007	1055.258
ARIMA(2,1,2)	0.6467	-0.9812	-0.6293	0.9153	3.9832	-518.7336	**1047.467**	**1062.569**	**1047.688**

Table 3 above shows a summary of the results from each fitted ARIMA model in our time series (of Facebook users). we select the model with the lowest AIC and BIC values is ARIMA(p=2,d=1,q=2), which is the best predictive model for estimating future values of our time series data.

Consequently, we are fitting our time series to the ARMA(2,2) model for d=1. Given that q is two in MA, this indicates the AR(2) model. Therefore, the model can be written as:

$$X_t = \mu + \left(\alpha_1 * (Z_{t-1} - \mu)\right) + (\alpha_2 * (Z_{t-2} - \mu) + \varepsilon_t$$

where μ is the time series mean and X_t is the stationary time series under examination. As shown in Table 3, ε_t is error with mean zero and constant variance, and ε_t, $\alpha_1 = AR1 = 0.0848$ and $\alpha_2 = AR1 = -0.0661$ are the parameters to be estimated. For a stationary differenced series, the mean (μ) should be 0 or extremely near to 0. In the event that μ is not equal to zero, in our study $\mu = 77.61968$. We use $\mu = 77.61968$ to forecast the future values in the above equation.

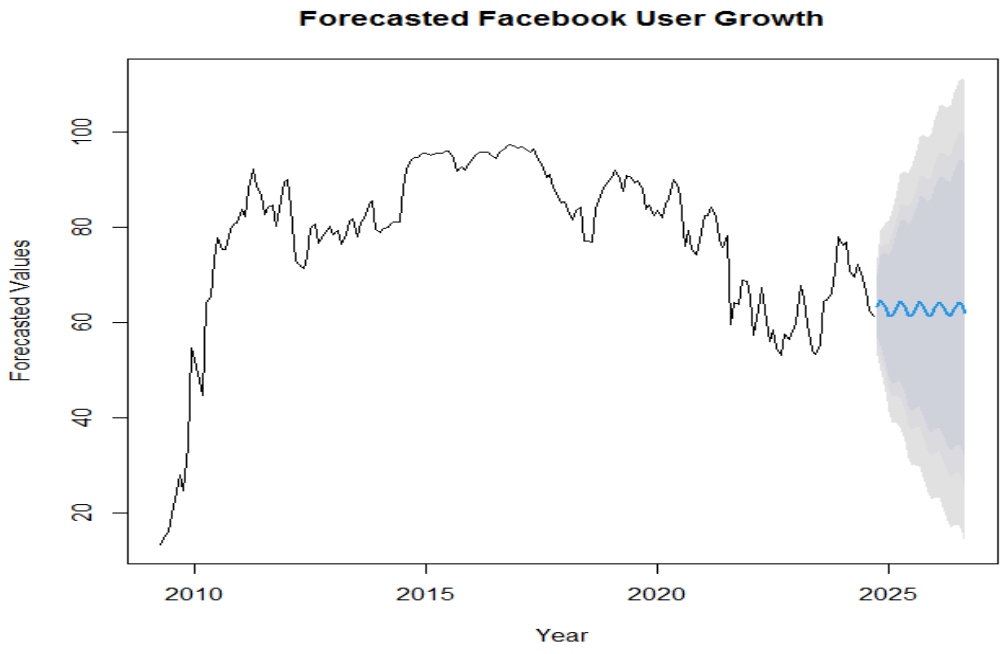
Forecasting:

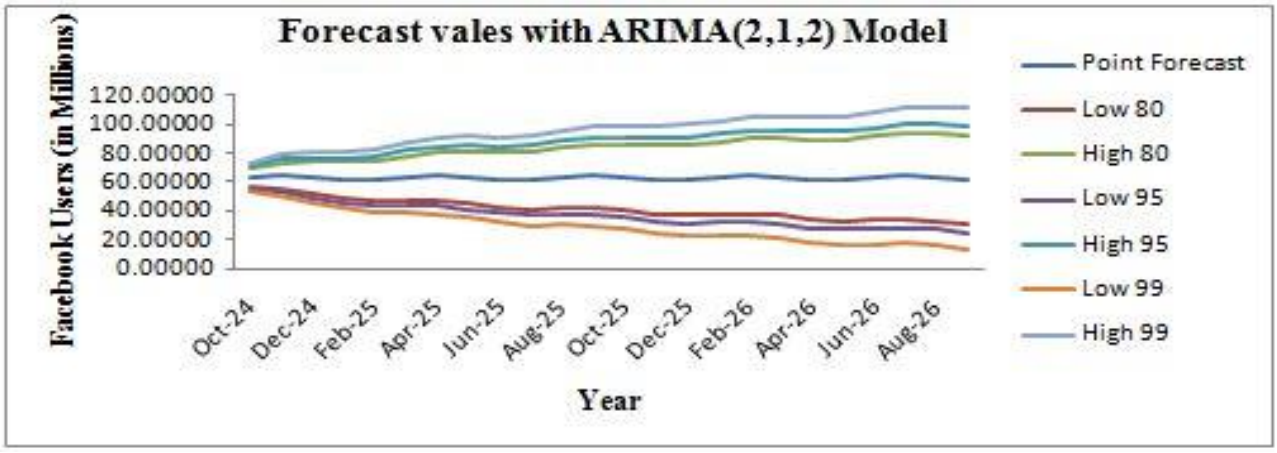
Table 4: 2-years forecasting for Facebook Users in India

Month Prediction	Point Forecast	Low 80	High 80	Low 95	High 95	Low 99	High 99
Oct-24	63.41230	56.86057	69.96404	55.60543	71.21918	53.15233	73.67227
Nov-24	64.60287	55.25661	73.94914	53.46611	75.73963	49.96669	79.23906
Dec-24	63.33938	52.06433	74.61443	49.90433	76.77443	45.68274	80.99603
Jan-25	61.35407	48.59922	74.10892	46.15573	76.55241	41.38006	81.32807
Feb-25	61.30997	47.17452	75.44543	44.46654	78.15341	39.17395	83.44600

Mar-25	63.22952	47.66533	78.79371	44.68364	81.77539	38.85611	87.60293
Apr-25	64.51413	47.58531	81.44295	44.34220	84.68607	38.00372	91.02455
May-25	63.46135	45.38447	81.53823	41.92142	85.00129	35.15308	91.76962
Jun-25	61.52002	42.47631	80.56372	38.82805	84.21199	31.69771	91.34233
Jul-25	61.29761	41.30729	81.28793	37.47767	85.11755	29.99291	92.60231
Aug-25	63.05869	42.04503	84.07235	38.01937	88.09801	30.15144	95.96593
Sep-25	64.41579	42.37515	86.45643	38.15275	90.67883	29.90031	98.93127
Oct-25	63.56538	40.62297	86.50778	36.22782	90.90294	27.63773	99.49303
Nov-25	61.68379	37.96495	85.40263	33.42105	89.94653	24.54025	98.82733
Dec-25	61.30144	36.81871	85.78417	32.12847	90.47442	22.96166	99.64123
Jan-26	62.90047	37.58414	88.21679	32.73420	93.06673	23.25527	102.54566
Feb-26	64.30971	38.13772	90.48170	33.12386	95.49556	23.32456	105.29486
Mar-26	63.65203	36.70999	90.59406	31.54861	95.75544	21.46099	105.84307
Apr-26	61.84392	34.23142	89.45641	28.94160	94.74623	18.60294	105.08489
May-26	61.31997	33.05014	89.58980	27.63439	95.00555	17.04962	105.59032
Jun-26	62.75533	33.76787	91.74278	28.21464	97.29601	17.36117	108.14948
Jul-26	64.19767	34.46316	93.93218	28.76682	99.62852	17.63364	110.76170
Aug-26	63.72199	33.30263	94.14134	27.47509	99.96889	16.08549	111.35849
Sep-26	61.99910	30.97872	93.01947	25.03603	98.96216	13.42140	110.57679

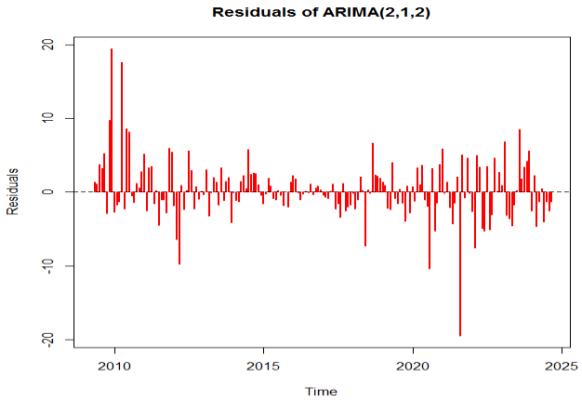
The forecast for our time series ARIMA(2,1,2) model's future values for a 2-year period up to September 2026 is displayed in Table 4. In Graph 6 and 7 below, we illustrate a two-year prediction of Facebook users.



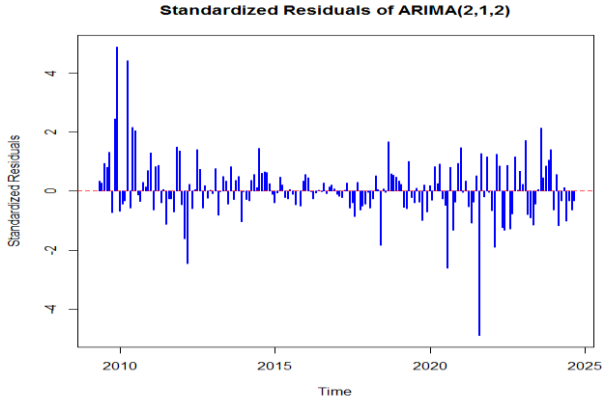


We will then investigate three questions: (1) are the forecast errors of our ARIMA(2,1,2) model normally distributed with $N(0, \sigma^2)$; (2) are there relationships among forecast errors; and (3) are the residuals simply white noise?

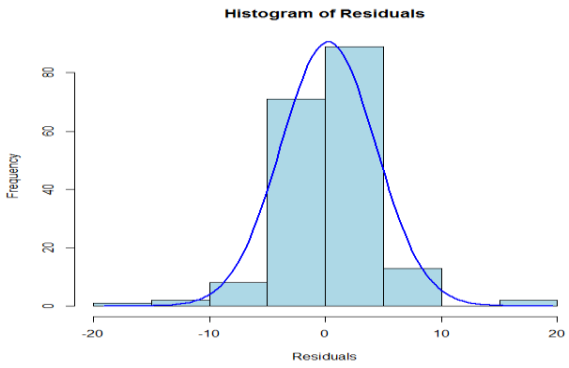
Plotting the errors will allow us to examine the predicted error distribution (standard residuals). Different standard residuals plots and histograms (prediction errors) of the fitted ARIMA(2,1,2) model are displayed in the Graph 8(a),8(b),8(c), and 8(d) below:



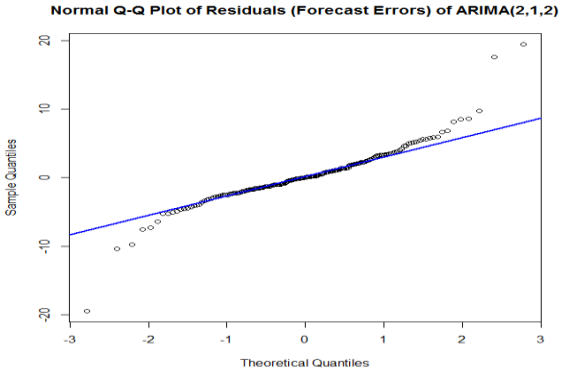
Graph 8 (a): Residuals ARIMA (2,1,2)



Graph 8(b): Standard residual of fitted ARIMA(2,1,2)



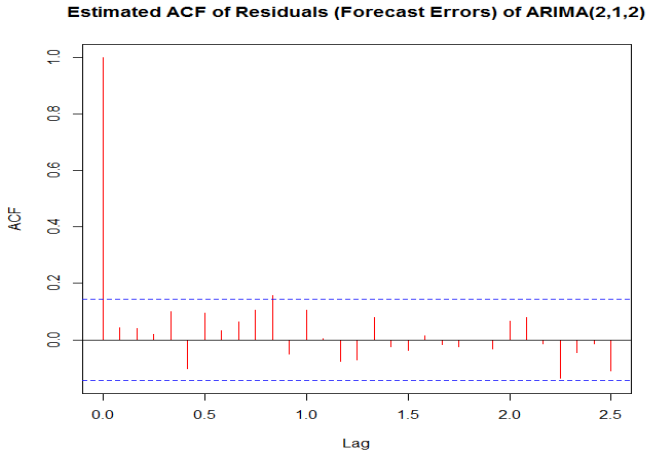
Graph 8(c): ARIMA (2,1,2) Histogram of Residuals (Forecast Errors)

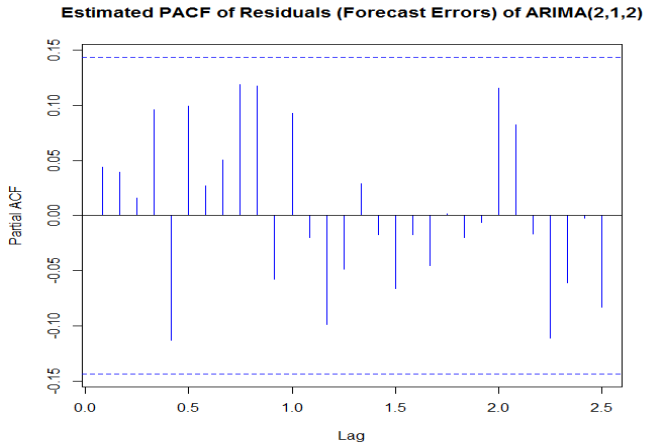


Graph 8(d): Residuals (Forecast Errors)–ARIMA(2,1,2) Normal Q-Q Plot

The fitted model's multiple line plots and Q-Q plots of the standard residuals (shown above in Figures 6(a) through 6(d)) suggest that the mean and variance of standard errors are consistent throughout time. However, there is a greater degree of variation at the beginning and conclusion of the series. Additionally, the histogram verified that errors have a normal distribution with zero as the distribution mean. Additionally, the Q-Q plot validates that the errors are normal.

We will now plot the prediction errors' partial and absolute correlograms (PACF and ACF) to further investigate any relationships between successive forecast errors below Figures 9 and 10.





Graph 9: Residuals (Forecast Errors) ARIMA Estimated ACF (2,1,2)

Graph 10: Estimated PACF of Residuals ARIMA(2,1,2)

Box-Ljung and Box-Pierce Test Statistics:

The null and alternative hypothesis:

H_o : The autocorrelation functions are zero.

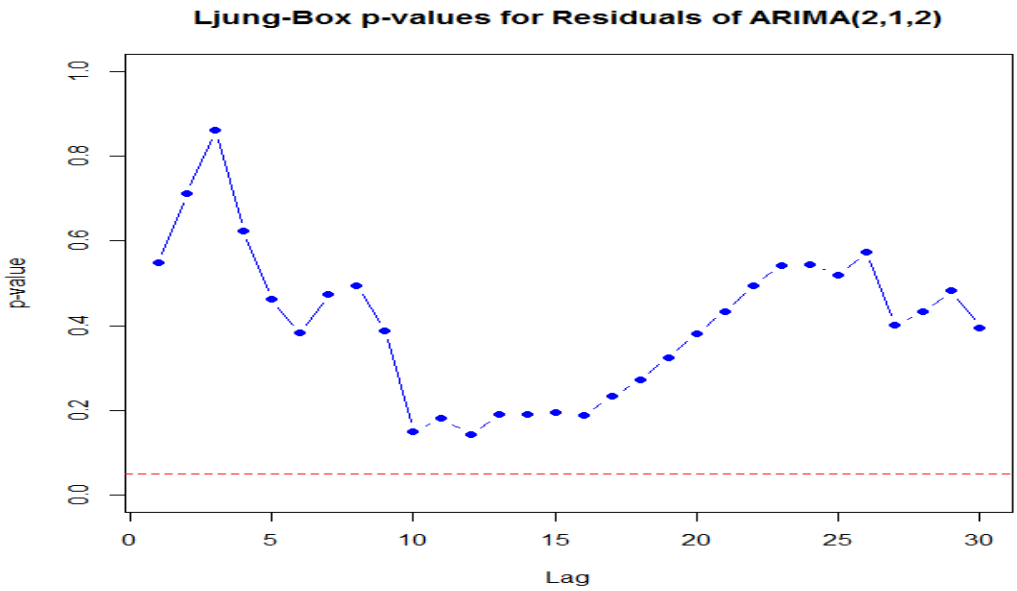
H_1 : The autocorrelation functions are not zero.

The test statistics results are shown in Table 5 below. The Box-Ljung "p-values" for the fitted model are shown in Figure 9 below.

We see that p-value for both test statistic is greater than the level of significance $(\alpha) = 0.05$. Thus, We fail to reject H_o.

Table 5: Statistics for the Box-Ljung and Box-Pierce Tests

Test	X^2	Degree of Freedom	p-value
Box-Ljung	31.419	30	0.395
Box-Pierce	28.553	30	0.5412



Graph 11: p-values for the fitted ARIMA (2,1,2) using Ljung-Box

We accept the null hypothesis, according to which all autocorrelation functions in lags 1 through 30 are zero, based on the statistics and high p-values in the two tests previously indicated. We can conclude that the forecast errors at lags ranging from 1 to 30 exhibit little to no evidence of non-zero autocorrelations in our fitted model.

Conclusion:

The ARIMA model can be used with this set of data because the differenced series passed the stationary condition and the ADF p-value is less than 0.05. The optimal model for forecasting was determined to be ARIMA(2,1,2) based on AIC, BIC, and AICc values. The residuals are mainly randomly spread around zero on the plot, suggesting no discernible patterns, which implies a decent fit. For optimal model accuracy, the residuals appear to be roughly normal based on the Q-Q plot and histogram. The Ljung-Box and Box-Pierce test p-values are generally greater than 0.05 at various lags, indicating that the residuals no longer include any significant autocorrelation. This provides more evidence that the ARIMA(2,1,2) model fits the data quite well.

This time series data on Facebook user growth seems to suit the ARIMA(2,1,2) model the best. It effectively captures the patterns without leaving significant autocorrelation in the residuals, which means that it can reliably be used for forecasting future Facebook user growth. The residuals analysis shows that the model assumptions are met, indicating robust model performance.

References:

[1] Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley.
[2] Chris Chatfield, Haipeng Xing : The Analysis of Time Series: An Introduction with R.
[3] Coghlan A. A Little Book of R for Time Series, Readthedocs.org, 2010. Available online at: http://a-little-book-of-r-for-time-series.readthedocs. org/en/latest/src/timeseries.html
[4] Enders, W. (2014). *Applied Econometric Time Series* (4th ed.). Wiley.
[5] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). OTexts.
[6] https://atsa-es.github.io/atsa-labs/sec-boxjenkins-aug-dickey-fuller.html
[7] Kumar, Manoj & Anand, Madhu. (2014). An Application Of Time Series Arima Forecasting Model For Predicting Sugarcane Production In India. Studies in Business and Economics. 9. 81 - 94.
[8] Nath, & Bhattacharya, Debasis & Correspondence, Debasis & Bhattacharya, & DHAKRE, DIGVIJAY. (2018). Forecasting wheat production in India: An ARIMA modelling approach. 2158-2165.

[9] Paul S.P. Cowpertwait, Andrew V. Metcalfe : Introductory Time Series with R https://books.google.co.in/books/about/Introductory_Time_Series_with_R.html?id=QFiZGQmvRUQC&source=kp_book_description&redir_esc=y

[10] Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples* (4th ed.). Springer.