adani | Institute of Digital Technology Management

**PGDM in Big Data Analytics**

**Trim II (Batch 2023-25)**

**Course Name: Time Series and Forecasting Techniques & Analysis**

Course Code: 210C208

SUB: TSFT [Autoregression (AR) 2 model]

**Submitted to: Dr. Dhyani Mehta**

**Submitted by:**


| Enrolment number | Name |
|---|---|
| 20231014 | Devang Jain |
| 20231032 | Lavanya Purohit |
| 20231037 | Meet Patel |
| 20231044 | Pratiksha Khampariya |
| 20231045 | Reet Neema |
| 20231060 | Yuvraj Singh Rathore |


Dated: 13th January 2024

# Tabel of Contents

# Index Price Forecasting: Autoregressive Model


## ABSTRACT

Predicting Index prices is a significant subject within the realms of finance and economics, capturing the attention of researchers over the years in the quest for improved predictive models. The autoregressive model, widely examined in literature for tune series prediction, serves as a focal point. This report details a comprehensive approach to constructing a stock price predictive model utilizing the AR 2 model. The model is developed using published Index data sourced from the National Stock Exchange (NSE).



Keywords: Short-term prediction, stock market, AR-2 model, and NIFTY BANK Index price prediction.


## INTRODUCTION

Bank Nifty is an index that represents the performance of the banking sector in the National Stock Exchange of India (NSE). It is one of the popularly traded indices in the Indian stock market and is composed of the most liquid and large capitalized banking stocks listed on the NSE.

Bank Nifty provides market participants, including investors, traders, and fund managers, with a benchmark to gauge the overall performance of the banking sector. The index is calculated using the free-float market capitalization weighted methodology, which means that the level of the index reflects the total market value of all the stocks in the index relative to a particular base period.

As the index comprises major banking stocks, movements in Bank Nifty are influenced by factors such as interest rates, economic conditions, government policies, and global economic trends. Traders and investors often use Bank Nifty futures and options to hedge their portfolios or speculate on the future direction of the banking sector.

The financial service in nifty banks are AU Small Finance Bank Ltd, Axis Bank Ltd, Bandhan Bank Ltd, Bank of Baroda, Federal Bank Ltd, HDFC Bank Ltd, ICICI Bank Ltd, IDFC First Bank Ltd, IndusInd Bank Ltd, Kotak Mahindra Bank Ltd, Punjab National Bank, State Bank of India.

The prediction will continue to be an interesting area of research making researchers in the domain field always desiring to Improve existing predictive models. The reason is that institutions and individuals are empowered to make investment decisions and can plan and develop effective strategies for their daily and future endeavours. Index price prediction is regarded as one of the most difficult tasks to accomplish in financial forecasting due to the complex nature of the stock market Many investors want to get their hands on any forecasting technique that will reduce investment risk and ensure Simple profits from the stock market. This continues to be a driving force behind the development of new prediction models by academics.

This report discusses how a stock price can be predicted using the Autoregressive model of order 2, which is a type of time series model commonly used in statistics and econometrics. In the context of Index price prediction or time series analysis, the AR (2) model represents a linear regression of the current value of a time series based on its two most recent past values. The general form of an AR (2) model can be expressed mathematically as:

$$Y_t = C + \alpha_1 * Y_t - 1 + \alpha_2 * Y_t - 2 + \varepsilon$$

Here:

$Y_t$ is the value of the time series at time t,
c is a constant or intercept term,
$\alpha_1$ and $\alpha_2$ are coefficients associated with the two lagged values ($Y_t - 1$ and $Y_t - 2$),
$\varepsilon$ is the error term or white noise.

The AR (2) model posits that the present value of a time series is influenced by a linear dependence on its two preceding values, characterizing it as a second-order autoregressive process. The coefficients $\alpha_1$ and $\alpha_2$ are derived from historical data to grasp the inherent patterns within the time series, enabling predictions of future values.

Day-to-day closing prices have been studied for the last three years of NIFTY BANK INDEX price data.

$$RT = C(1) + C(2)*RT(-1) + C(3)*RT(-2) + \varepsilon$$

## Coefficient Labels:

| VARIABLE | COEFFICIENT |
|---|---|
| C | C (1) |
| RT (-1) | C (2) |
| RT (-2) | C (3) |

## ASSUMPTIONS OF LINEAR REGRESSION:

**1. Linearity:**

- The relationship between the dependent variable RT (RETURN) and the independent variables RT (-1) and RT (-2) is assumed to be linear.

**2. Fixed values of independent variables:**

- There should also be no autocorrelation between independent variables and residuals.

**3. Randomness of Residuals:**

- The residuals (ε) are assumed to be independent, meaning that the value of the residual for one observation is not dependent on the value of the residual for any other observation.
- The residuals (ε) should be normally distributed and have normal bell-shaped curve.

**4. Homoscedasticity (Constant Variance):**

- The variance of the residuals is assumed to be constant across all levels of the independent variables.
- All residuals consecutively should be normally distributed.

**5. Autocorrelation of Residuals:**

- There should not be any autocorrelation between residuals.
- This can also be interpreted in this manner that; no two residuals should be autocorrelated with each other.

**6. No Perfect Multicollinearity:**

- There should not be any correlation between residual and variables.

This report delves into a specific aspect of the Bank Nifty Index Data showing the relationship between three components: RT, RT (-1) and RT (-2).

## DEFINING VARIABLES

| Variable | Variable Code | Description |
|---|---|---|
| Return | RT | This variable represents the ratio of current return in respect to previous return.<br>= (current closing price - previous closing price)/previous closing price |
| Return(-1) | RT (-1) | This variable represents one value previous than RT. |
| Return(-2) | RT (-2) | This variable represents two value previous than RT |

**Endogenous Variable:**
RT

**Exogenous Variables**:
RT(-1)
RT(-2)

Mathematically we can define this relationship as:

**RT = f (RT (-1), RT (-2))**

## DESCRIPTIVE STATISTICS:

| | RETURNS |
|---|---|
| Mean | 0.000664 |
| Median | 0.000840 |
| Maximum | 0.082562 |
| Minimum | -0.057872 |
| Standard Deviation | 0.012281 |
| Skewness | 0.076634 |
| Kurtosis | 7.562995 |
| Jarque-Bera | 643.5713 |
| Probability | 0.000000 |

As we can see the median is greater than the mean in returns. So, there is a positive skewness present as per the result shown above which is 0.076634.

As the kurtosis is more than three shows that the distribution of returns is leptokurtic in nature which indicates that there are more extreme values in the data compared to a normal distribution. The peak of the distribution tends to be higher and sharper, leading to more values clustering around the mean than those of normal distribution.

- The positive skewness indicates that the distribution is skewed to the right, and the high kurtosis suggests heavy tails and a peaky distribution.
- The extremely low p-value in the Jarque-Bera test indicates that the data is not normally distributed.
- The presence of extreme values (as indicated by the maximum and minimum values) might contribute to the skewness and kurtosis.
- The wide range between the maximum and minimum values, along with the positive skewness, suggests that the data may have a long-right tail with some extreme values pulling the mean to the left.
- The small standard deviation suggests that the values are relatively close to the mean on average.

## CORRELATION:

| | RT | RT (-1) | RT (-2) |
|---|---|---|---|
| **RT** | 1.000 | 0.084959 | -0.053413 |
| **RT (-1)** | 0.084959 | 1.000 | 0.084911 |
| **RT (-2)** | -0.053413 | 0.084911 | 1.000 |

In the above table, we came to know that the correlation between RETURN and it's first difference (RT (-1)) is 0.084959 which signifies that there is very weak positive correlation.

The correlation between RETURN and it's second difference (RT (-2)) is -0.053413 which shows that there is very weak negative correlation.

Apart from that the correlation between RETURN (-1) and RETURN (-2) is 0.084911 which shows there is another very weak positive correlation.

- The correlation coefficients between the current return and the returns at lag 1 and lag 2 are very close to zero. This suggests that there is very little linear association between the current return and the returns at the previous two time points.

- The weak correlations imply that the current return is not strongly predictable based on the returns at the two previous time points.

## NORMALITY TEST FOR VARIABLES

| VARIABLES | JARQUE-BERA | CORRELOGRAM | ADF TEST |
|---|---|---|---|
| RT (P Value) | 643.5713 (0.000) | 5.3375 (0.021) | -24.95154 (0.000) |
| RESULTS | NOT NORMAL | NORMAL | NORMAL |

**RT (RETURN):**

### a) Jarque-Bera (JB) Test:

**H0: The data is normally distributed.**
**H1: The data is not normally distributed.**

This test shows that the probability value of 0.00, which is less than the significance value of 0.05 and we can say that there is significant evidence to reject the null hypothesis. Hence, we accept the Alternate Hypothesis.

### b) Correlogram:

**H0: There is no autocorrelation and partial autocorrelation between the data.**
**H1: There is autocorrelation and partial autocorrelation between the data.**

The correlogram tells us whether errors at one time point are related to errors at another time point. A probability value of 0.021 is less than 0.05, indicating a significant relationship between errors at different times and we can reject the null hypothesis.

## c) ADF Test:

**H0: Data has a unit root.**
**Probability > 0.05 Data is Not Normally Distributed & Trend or A.C.**

**H1: Data do not have unit root.**
**Probability < 0.05 Data is Normally Distributed & NO Trend or No A.C.**

A probability value of 0.00 is less than 0.05 and we can conclude that there is enough evidence to reject the null hypothesis saying that the data has unit roots and accepting the alternate hypothesis. Thus, data does not have a unit root indicating that the data is normally distributed.

# AUTO-REGRESSIVE MODEL

| VARIABLE | COEFFICIENT | PROBABILITY |
|---|---|---|
| C | 0.000624 | 0.1670 |
| RT(-1) | 0.090056 | 0.0145 |
| RT(-2) | -0.061011 | 0.0974 |

RT (RETURN) = 0.000624 + 0.090056*RT(-1) + -0.061011*RT(-2)

| | |
|---|---|
| $R^2$ | 0.010920 |
| Adjusted $R^2$ | 0.008233 |
| F-Stat | 4.063038 |
| Probability(F-Stat) | 0.017584 |
| Durbin Watson Stat | 2.005570 |

- $R^2$ (Coefficient of Determination):

  In the context of time series models, $R^2$ represents the proportion of the variance in the dependent variable (returns in this case) that is explained by the explained sum of square. A value of 0.010920 means that a very small percentage (1.092%) of the variability in returns is explained by the AR (2) model.

- Adjusted $R^2$:

  Adjusted $R^2$ takes into account the degree of freedom and adjusts $R^2$ accordingly. A value of 0.008233 is unusual and may indicate potential issues with overfitting or model complexity. It's rare to have a negative adjusted $R^2$, and this might warrant further investigation.

- F Statistic and Probability (F Stat):

  The F-statistic tests the overall significance of the model. In your context, an F-statistic of 4.063038 with a probability of 0.017584 suggests that the AR(2) model is statistically significant. This means that the combined effect of the lagged returns in predicting current returns is significant.

- Given these statistics we can say that the AR(2) model, as described by the provided statistics, is providing a significant improvement in explaining the variation in returns compared to a simple mean. The low $R^2$ and significant F-statistic suggest that the model may be effective in predicting returns for Bank Nifty based on the provided features.

## NORMALITY TEST FOR ERRORS

| Sr. No | Tests | Probability Values | RESULT |
|---|---|---|---|
| 1 | Jarque-Bera | 0.000 | NOT NORMAL |
| 2. | Correlogram | 0.885 | NORMAL |
| 3. | Augmented Dickey-Fuller Test | 0.000 | NORMAL |
| 4. | BGLM | 0.103 | NORMAL |
| 5. | DW Test | 2.005 | NORMAL / NO AUTOCORRELATION |

### I. Jarque-Bera (JB) Test:

**H0: The data is normally distributed.**
**H1: The data is not normally distributed.**

This test shows that the probability value of 0.00, which is less than the significance value of 0.05 and we can say that we reject the null hypothesis and conclude that the errors are not normally distributed. Hence, we accept the alternative Hypothesis.

### II. Correlogram:

**H0: There is no autocorrelation and partial autocorrelation between the data.**
**H1: There is autocorrelation and partial autocorrelation between the data.**

The correlogram tells us whether errors at one time point are related to errors at another time point. From the table the probability value is more than 0.05, indicating that there is no significant relationship between errors at different times and we fail to reject the null hypothesis and accept the alternate hypothesis.

### III. ADF Test (Augmented Dickey-Fuller Test):

**H0: There is a unit root.**
**H1: There is no unit root.**

A probability value of 0.0000 is less than 0.05 and we can conclude that there is enough evidence to reject the null hypothesis saying in the data which has unit roots and accepting the alternate hypothesis. Thus, data does not have a unit root, which means stationery data, indicating that the data is normally distributed.

## IV. BGLM Test (Box-Godfrey LM Test):

**H0: There is no autocorrelation in the residuals as lag = 0**
**H1: There is autocorrelation in the residuals because of lag = 0.**

A probability value of 0.103 is greater than 0.05, which indicates that there is enough evidence to accept the null hypothesis which indicates there is not correlated in the residuals within the data. and we will be rejecting the alternative hypothesis which indicates there is correlated in the residuals within the data.

## V. Durbin-Watson (DW) Test:

Ho: No (+) & (-) Autocorrelation

Probability Value: 2.0055
Accept: DU < DW < 4-DU

DW = 2.005 As we can see our DW value lies between du and 4-du, it means there is no Autocorrelation between our errors.

## HETEROSCEDASTICITY TEST

| Sr. No | Test | P Value(F-stats) | P Value (Chi-square) |
|---|---|---|---|
| 1. | Glejser | 0.0060 | 0.00610<br>- heteroskedastic |

**Glejser Test:**

- **Probability Value:** 0.006
- **Interpretation:** the low probability value (close to 0) ) indicates that there is strong evidence to reject the null hypothesis of homoscedasticity. Accept the alternate hypothesis of heteroskedasticity
- **Chi-Square Value:** 0.0061

**Interpretation:** The chi-square value represents a measure of the difference between the expected and observed values in the test. In this case, the relatively low chi-square value (close to 0) suggests that the model's residuals significantly deviate from the assumption of constant variance. The test provides evidence supporting heteroskedasticity.

# MULTI-COLLINEARITY TEST

| Variable | Coefficient Variance | Centered VIF |
|---|---|---|
| C | 2.04E-07 | NA |
| RT (-1) | 0.001351 | 1.007262 |
| RT (-2) | 0.001351 | 1.007262 |

**1. Coefficient Variance:**

- For RT (-1) and RT (-2), the numbers (0.001351 and 0.001351) are small. This means that if these variables were related to other variables, the impact on the results of our analysis would be tiny.
- For variable C, the number (0.000000204) is extremely small, suggesting that even if C is linked to other variables, it won't really affect our analysis much.

**2. Centered VIF:**

- Both RT (-1) and RT (-2) have VIF values close to 1 (1.007262), which is good. There is no multi-collinearity. It means these variables are not strongly connected to each other in a way that would cause problems in our analysis.
- For variable C, the VIF is not available (NA), which might be a sign of a more serious issue. It could also mean that C is too closely related to other variables. making our analysis difficult.

# CONCLUSION

Equation: RETURN = 0.000624 + 0.090056*RT(-1) + -0.061011*RT(-2) + ε

**The AR(2) model for predicting the returns of Nifty Bank does not appear to be statistically significant, with low explanatory power (low $R^2$ and adjusted $R^2$) and non-significant coefficients. The model does not provide a meaningful improvement in explaining the variability in returns compared to a simple mean. The findings suggest that the lagged returns in the AR(2) model are not effective predictors for the current returns of Nifty Bank stock based on the provided features. Further refinement of the model, exploration of additional features, or consideration of alternative modelling approaches may be necessary. Additionally, checking for the presence of outliers or influential observations could be important for model improvement.**

**DATA SOURCE**
https://www.nseindia.com/reports-indices-historical-index-data

- BANK NIFTY INDEX DATA