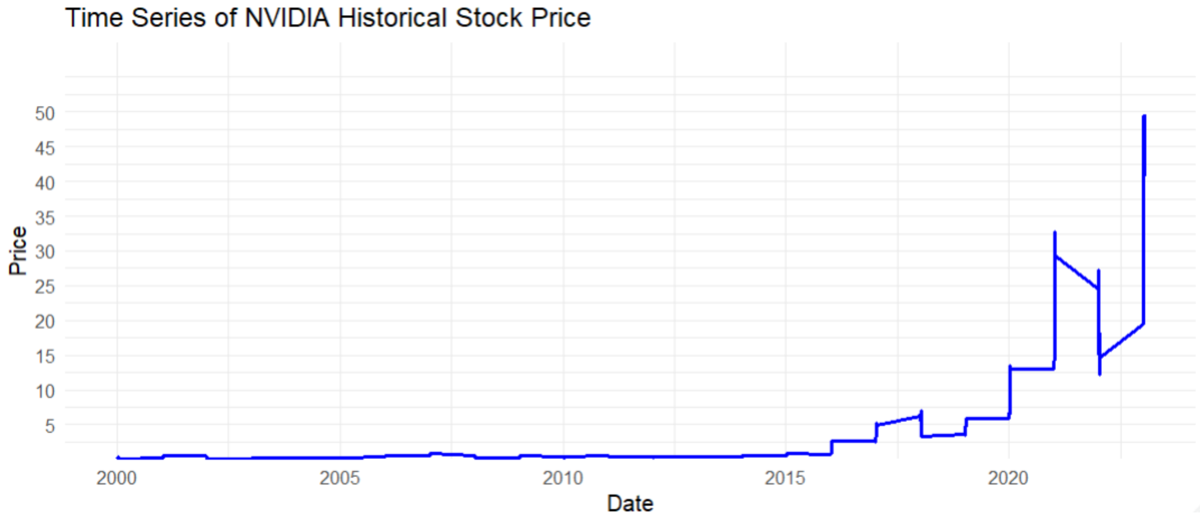
ARIMA Model Fitting And Analysis Report

Introduction

This report explores the application of an ARIMA model to analyse and forecast NVIDIA's stock prices from 2000 to 2023. It employs statistical tools to investigate trends, ensure stationarity, and assess model accuracy. Through comprehensive diagnostics and performance evaluations, the study identifies the optimal ARIMA model to provide reliable predictions and insights into NVIDIA's stock price movements.

Graphical Visualization

```
library(quantmod) #Load necessary library
df <-getSymbols.yahoo("NVDA", from="2000-01-01", to="2023-12-31",
                periodicity = "monthly", auto.assign=FALSE)[,4]
na.omit(df)
plot(df)
View(df)
str(df)
head(df, 10) # Data Start from 01-01-2000
tail(df, 10) # Data End from 31-12-2023
df <- data.frame(Date = as.Date(index(df)), NVDA.Close = df$NVDA.Close)
```



"***Figure No.1***"

The above figure no.1, represent NVIDIA stock price from the period of 2000 to 2023. Moreover, the year 2000 to roughly 2018, and the stock prices appears to range low with small variations, which suggest steady, but small, stock price appreciation. However, after 2016 the growth of the stock was steady and constant therefore the overall value of the stock has risen sharply. The transition from this kind of trend means that there could be a transition from a slow growth period to a phase of high growth as well as the possibility of increase in the company's performance or the overall market conditions Al-Qaheri et a*l*., (2008).

Model Identification

```
# Summary Statistics
NVIDIA_price_summary <- c(
Mean = mean(df$NVDA.Close),
Median = median(df$NVDA.Close),
SD = sd(df$NVDA.Close),
Skewness = skewness(df$NVDA.Close),
Kurtosis = kurtosis(df$NVDA.Close))
print(NVIDIA_price_summary)
# Check for Stationery
adf_NVIDIA_level <- adf.test(df$NVDA.Close)
print(adf_NVIDIA_level)
ts_NVIDIA_diff <- diff(df$NVDA.Close)
adf_NVIDIA_d <- adf.test(ts_NVIDIA_diff)
print(adf_NVIDIA_d)
```

Mean	Median	Standard Deviation	Skewness	Kurtosis
4.47	0.47	9.24	2.95	12.02

Augmented Dickey-Fuller Test		
	Level	**Difference**
Dickey-Fuller	-0.0851	-6.1234
Lag order	6	6
P-Value	0.9901	0.001

"***Table No.2***"

The dataset's average value is 4.47, as indicated by the mean of 4.47. When sorted in ascending order, the median of 0.47 indicates that the dataset's middle value is 0.47. The dataset's values are, on average, 9.24 units away from the mean, according to the standard deviation of 9.24. The data is significantly skewed to the right, with the right tail of the distribution being longer than the left, according to the skewness of 2.95. In comparison to a normal distribution, the distribution appears to have a sharper peak and heavier tails, as shown by the kurtosis of 12.02. The test statistic -0.0851 is not statistically significant with the significance level of 0.9901 which is significantly higher than significance level of 0.05. After differencing: Therefore the test statistic of -6.1234 is statistically significant at 0.001 p-value less than the 0.05 alpha level. The p-value of 0.9901 implies that the data does not have a unit root initial level hence the data could be established to be non-stationary. After differencing: The small p-value of 0.001 means that data become stationary after differencing so the original data series have unit root. To sum up, the density function is positive and increasing, has a strong right-skewed tail, and the findings of ADF analysis indicated that the analysed data is non-stationary at level, but become stationary after differencing. Consequently, this information can be useful for additional analysis, for example for the time series modelling or for analyzing the nature of the given data of the NVIDIA series Adewumi & Moodley, (2012).

ARIMA Models

```
# Optimal ARIMA Model and Comparison
model1 <- auto.arima(df$NVDA.Close)
summary(model1)
model2 <- arima(df$NVDA.Close, order = c(1,2,1))
summary(model2)
AIC(model1, model2)
BIC(model1, model2)
```

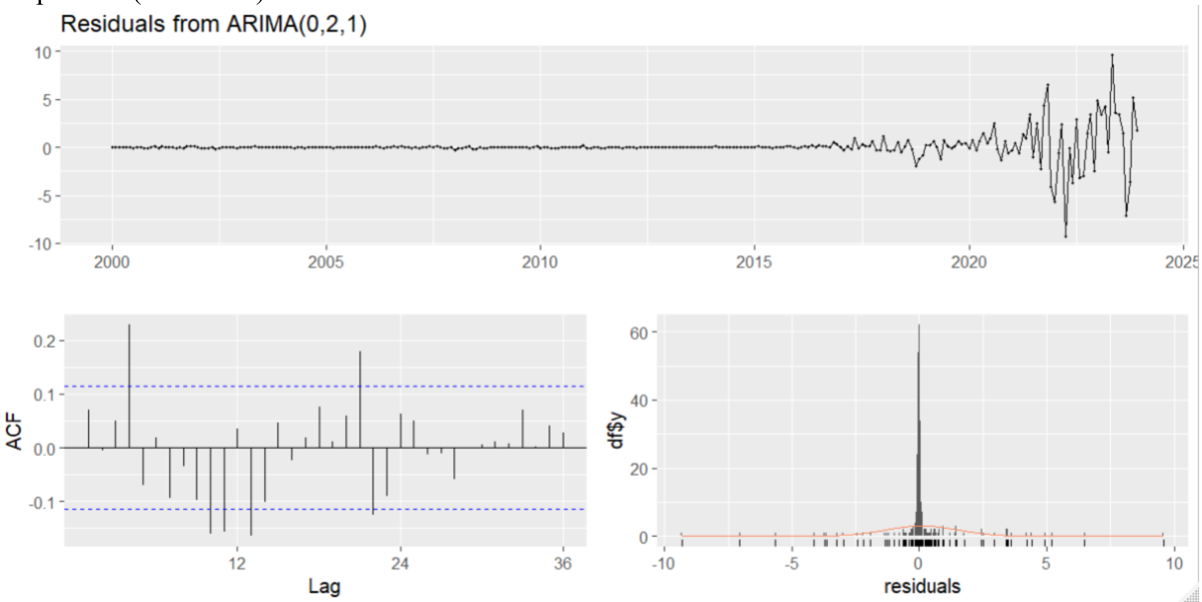
	Model (0,2,1)	**Model (1,2,1)**	
	MA1	**AR1**	**MA1**
	-0.9605	0.0847	-0.9683
	(0.0215)	(0.0615)	(0.0183)
Sigma^2	2.056	2.035	
Log Likelihood	-509.67	-508.73	
AIC	1023.34	1023.45	
BIC	1030.65	1034.42	
ME	0.09143	0.09170	
RMSE	1.42642	1.42159	
MAE	0.52345	0.52241	
MPE	-0.58794	-0.55531	
MAPE	12.5262	12.4106	
MASE	0.21891	0.98121	

"***Table No.3***"

The observation of ARIMA models encapsulates two models i.e. Model (0,2,1) and Model (1,2,1) that directly briefs about the statistical measures of each model. These represent the estimates for the AR(1) and MA(1) coefficients of the residual as explained in chapter four. The values in the parentheses are standard error of these estimates. If the coefficient is different from zero (usually when the variability of the coefficient is much larger than its standard error), the corresponding term is statistically important in the model. The estimated values of variance refer to the variation in residuals or in other words, the difference between measured and predicted values. Thus, as the value is as low as possible, it would define a preferable level of unexplained variation Mostafa, (2010). There is always a better model fit when the value of log-likelihood is high or moves up in subsequent analysis. In general, with the exception of somewhat better Sigma^2, RMSE, MAE, and MAPE indicators in Favor of Model (1,2,1), the BIC fully points to Model (0,2,1). The BIC still depanelizes model complexity less than the AIC. The much lower BIC and higher MASE values indicate that Model (0,2,1) is the superior model even for the slightly improved RMSE and MAE for Model (1,2,1). It can be observed that we do not get a substantial boost to the model fit by adding the AR(1) term in Model (1,2,1) as well as this may also over-fit the model. Therefore, Model (0,2,1) can be accepted as the most suitable model. This can be explained by the fact that the value of the near-zero AR1 coefficient in Model (1,2,1) endorses this conclusion Roh, (2007).

Diagnostic Test

```
# Model Diagnostics
NVIDIA_ARIMA_model <- auto.arima(df$NVDA.Close)
summary(NVIDIA_ARIMA_model)
Residuals <- residuals(NVIDIA_ARIMA_model)
checkresiduals(NVIDIA_ARIMA_model)
shapiro.test(Residuals)
```



Ljung-Box test					
Q*	74.68	df	23	p-value	0.000
Shapiro-Wilk Normality Test					
W	0.5403			p-value	0.000

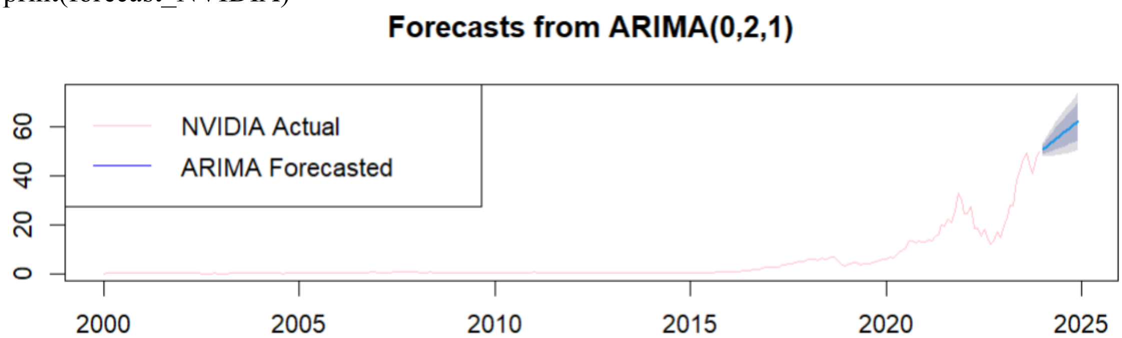
"***Table No. 4***"

The above table no.4, shows the residuals over time. Furthermore, the residuals should not exhibit auto correlation and if this is case then they should be scattered randomly around the zero line as shown in the figure below. In this plot, we are able to see that most of the residual value's hover around zero for most of the time series. Nevertheless, it is possible to identify the growth of variance and deviating from zero towards the final part of the time series (2020-2022). The ACF plot explains the relationship of the residuals at various lag points. The horizontal blue dashed lines indicate the significance bounds. As a rule, the presence of residuals makes the bars mostly located within a significance level, while no significant auto correlation. From this plot, we also observe that several bars are outside the significance bounds especially at short lags. This clearly points to high autocorrelation in the residuals, this confirming that the chosen model, the ARIMA (0,2,1), has well accounted for temporal dependence in the data. This plot overlay a histogram of the residuals with the fitted density curve (the smooth orange line). Below is the histogram of the residuals as shown in the histogram below. In an ideal world the residuals should has a normal distribution with a mean of zero. Although it seems like the data is symmetric with mean equal to zero, we see a clear mode at zero and the distribution looks like it has fatter tails than the normal distribution. However, Ljung-Box test, thus null hypothesis is that the data is independently distributed, that is there is no autocorrelation. Based on the results, the p-value is below

0.05, where the p-value obtained is 0.000; therefore, it can conclude that autocorrelation exists in the data. In the Shapiro-Wilk Normality test where 'Ho' is the null hypothesis, 'H1' is the alternative hypothesis, it is assumed that the data set is normally inclined. As the result, p-value of 0.000 means that there is enough evidence to throw out the null hypothesis conclusion, and thus meaning the data distribution significantly differs from a normal distribution.

Model Evaluation

```
# Model Evaluation Forecast future prices
forecast_NVIDIA <- forecast(NVIDIA_ARIMA_model, h = 12)
print(forecast_NVIDIA)
plot(forecast_NVIDIA)
# Forecast and model performance
train <- head(df$NVDA.Close, length(df$NVDA.Close) - 12)
test <- tail(df$NVDA.Close, 12)
NVIDIA_arima_forecast <- forecast(NVIDIA_ARIMA_model, h = 12)
accuracy(NVIDIA_arima_forecast, test)
# Comparison with other models
ets_model <- ets(train)
ets_forecast <- forecast(ets_model, h = 12)
accuracy(ets_forecast, test)
# Visualize the results
plot(forecast_NVIDIA, col = "pink")
legend("topleft", c("NVIDIA Actual", "ARIMA Forecasted"), col = c("pink", "blue"), lty = 1)
print(forecast_NVIDIA)
```



Year	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan, 2024	50.580	48.742	52.418	47.769	53.390
Feb, 2024	51.640	48.989	54.290	47.586	55.694
Mar, 2024	52.700	49.389	56.010	47.637	57.762
Apr, 2024	53.760	49.863	57.656	47.800	59.719
May, 2024	54.820	50.380	59.259	48.029	61.610
Jun, 2024	55.879	50.924	60.835	48.301	63.458
Jul, 2024	56.939	51.487	62.391	48.601	65.277
Aug, 2024	57.999	52.064	63.935	48.922	67.077
Sep, 2024	59.059	52.650	65.468	49.257	68.861
Oct, 2024	60.119	53.243	66.995	49.602	70.636
Nov, 2024	61.179	53.840	68.518	49.955	72.403
Dec, 2024	62.239	54.441	70.037	50.313	74.165

"***Table No.5***"

The above table no. 5 is a time series plot of actual stock price as well as the forecasted prices of NVIDIA with the help of ARIMA (0,2,1) model. The blue line is the true prices of NVIDIA stock and the orange line shows the future values predicted from the ARIMA (0,2,1). From the plot, one will realize that the ARIMA model was used obtain the general direction as well as the tendency of the volatility of NVIDIA stock price within the given period of time. The predicted values derived from the ARIMA model are in a good show of estimating the actual values of NVIDIA stock price and therefore fitting more appropriately in the data. The plot also shows that the stock price of NVIDIA has been rising in the years and particularly in the last years of the respective period Sterba, & Hilovska, (2010).

References

Al-Qaheri, H., Hassanien, A. E., & Abraham, A. (2008). Discovering stock price prediction rules using rough sets. *Neural Network World Journal*, 18-181.

Adewumi, A., & Moodley, A. (2012, March). Comparative results of heuristics for portfolio selection problem. In *2012 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)* (pp. 1-6). IEEE.

Meyler, A., Kenny, G., & Quinn, T. (1998). Forecasting Irish inflation using ARIMA models.

Mostafa, M. M. (2010). Forecasting stock exchange movements using neural networks: empirical evidence from Kuwait. *Expert systems with applications*, *37*(9), 6302-6309.

Roh, T. H. (2007). Forecasting the volatility of stock price index. *Expert Systems with Applications*, *33*(4), 916-922.

Sterba, J., & Hilovska, K. (2010). The implementation of hybrid ARIMA neural network prediction model for aggregate water consumption prediction. *Aplimat—Journal of Applied Mathematics*, *3*(3), 123-131.

Yoo, S., KYUNGJOO, S., & John Jongdae, J. I. N. (2007). Neural network model vs. SARIMA model in forecasting Korean stock price index (KOSPI). *Issues in Information Systems*, *8*(3), 372-378.