Stock Price Prediction: Comparison between ARIMA and LSTM

1. Introduction

The two datasets used for analysis are the daily stock prices from September 3, 2019, to August 31, 2021, of *Southwest Airlines Co.* and *Netflix Inc.* from *Yahoo Finance*. For each dataset, the date and close prices in US dollars are being analyzed and predicted using two methods, which are ARIMA (Autoregressive integrated moving average) and LSTM (Long Short-Term Memory). ARIMA is a very classical approach for time series analysis in Statistics and Economics, while LSTM is an artificial recurrent neural network architecture in deep learning. The goal of this project is to explore which approach predict and forecast time series better.

Since during the pandemic, traveling is limited and people tend to spend more time at home watching TVs, which lead to the very opposite performance on the stock markets for the Airline industry and Streaming Media industry. Therefore, at the early stage of data selection, a strong interest was raised, in how stock prices of the companies in two very different industries performed before and during COVID-19 pandemic, and in how, the diversity impacts, if there is any, the model performance and forecasts between ARIMA and LSTM.

Figure 1 shows the stock prices of *Southwest Airlines Co.* and *Netflix Inc.* across the dates being analyzed. An obvious V-shaped recovery on *Southwest Airlines* plot and an increasing trend on *Netflix* plot can be seen. No obvious seasonality pattern is found on two plots.

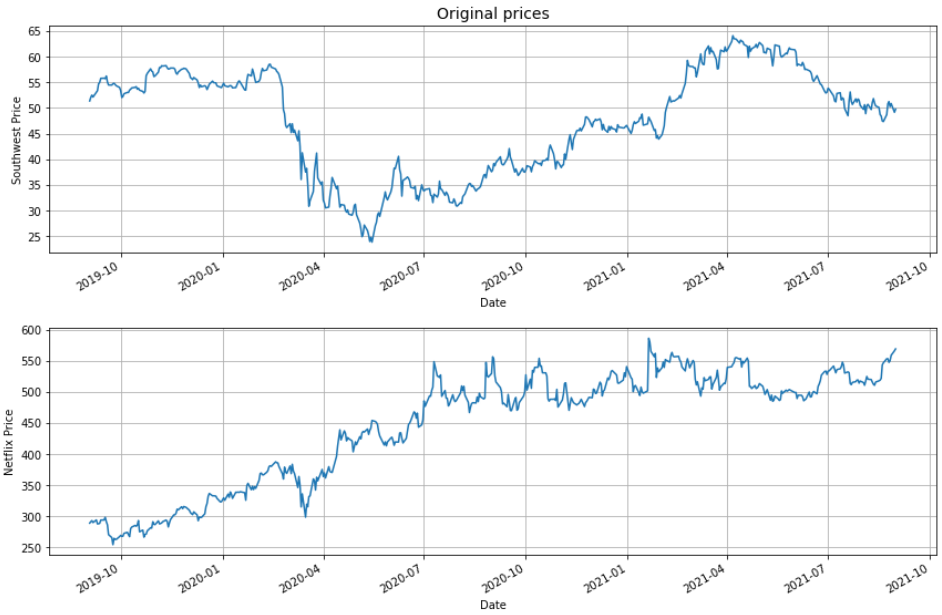


Figure 1 Stock Prices of Southwest Airlines and Netflix

2. ARIMA Model

Since there are obvious trends in plots for both datasets, in order to eliminate the non-stationary behaviors, methods such as differencing and log transformation can be performed. The ACF plots (*Figure 2*) show a rapid decay for both dataset after 1st order differencing and log transformation, in comparison to the slow decay of the sample ACFs for the original datasets, which suggests that the time series of the two datasets show signs of stationary after using these two methods. The Augmented Dickey-Fuller test is also

performed for both datasets and the results show that log transformation and 1st order differencing together are necessary for both stocks. In order to generate the best models for ARIMA method, comparisons among different p and q values are then performed for both datasets.

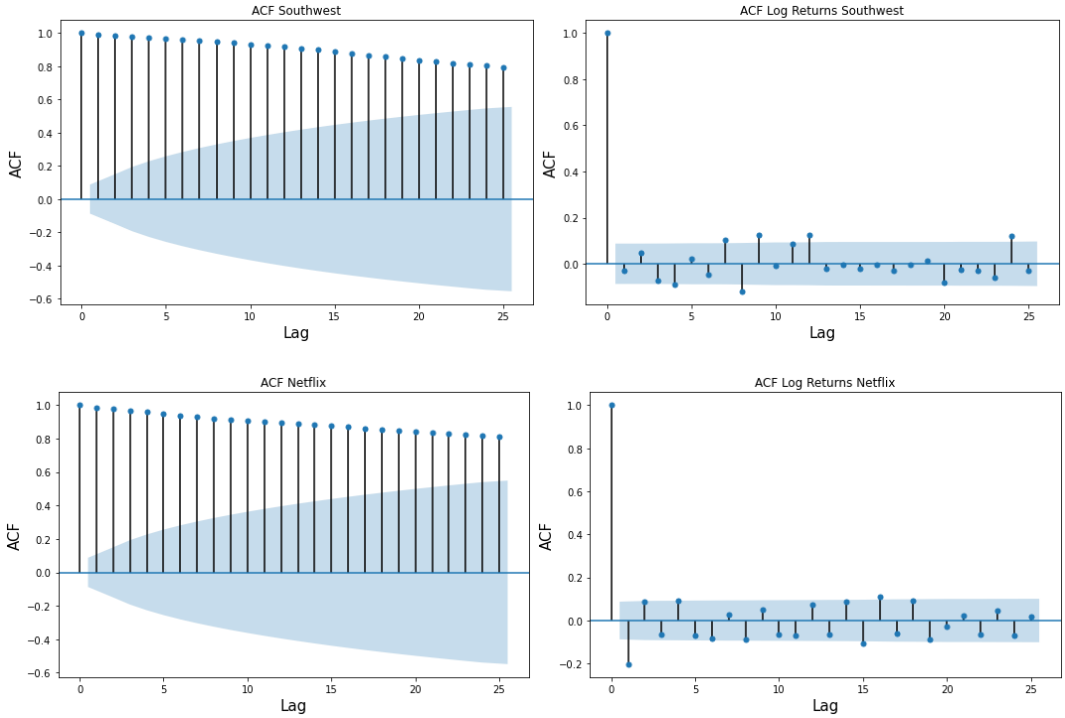


Figure 2 ACF plots for Southwest (top) and Netflix (bottom)

2.1 Model Selection

2.1.1 Southwest Airlines Dataset

To select the best ARIMA model for Southwest airlines, the log transformed dataset is split into 80-20 train and test sets. A function from the 'statsmodel' library is used to select the best models based on training set data, regarding AIC (Akaike Information Criterion) and BIC (Bayes Information Criterion) separately among all pairs of (p, q) under 4th orders. Looking at *Table 1*, ARIMA(0,1,0) has the lowest AIC and BIC values than all other (p, q) combinations. The p-value for other combinations also suggest that higher orders of AR and MA are unnecessary (see Appendix). Therefore, the best model selected for Southwest Airlines Dataset is ARIMA(0,1,0). The residual diagnostic plots and ACF plots (see Appendix) also prove that with ARIMA(0,1,0) and log transformation, the dataset becomes stationary.

Table 1 ARIMA Parameter Comparison

p	d	q	AIC	BIC
0	1	0	-1598.550	−1590.557
1	1	0	-1596.926	−1584.936
0	1	1	-1596.880	−1584.891
1	1	1	-1595.825	−1579.839

2.1.2 Netflix Dataset

The same process is performed for the Netflix Dataset to select the best parameters for the ARIMA model. *Table 2* shows that ARIMA(2,1,2) has the lowest AIC, while ARIMA(1,1,0) has the lowest values. The p-values (see Appendix) of parameters are mostly under the significance level for ARIMA(2,1,2). However, because using AIC usually selects models with higher orders than BIC does, if not the same, and the AIC value of ARIMA(2,1,2) is not smaller than that of ARIMA(1,1,0) by a large amount, ARIMA(1,1,0) is the most suitable model for Netflix Dataset.

Table 2 ARIMA Parameter Comparison

p	d	q	AIC	BIC
0	1	0	-2270.804	−2262.363
1	1	0	-2289.838	−2277.177
2	1	2	-2292.120	−2266.797

2.2 Out-of-Sample Forecasting Using ARIMA

After selecting the best (p, d, q) combinations for ARIMA models for both datasets, a forecasting on the out-of-sample test set is performed for Southwest airlines (*Figure 3 top*) and Netflix (*Figure 3 bottom*) data. The price values are inverse log transformed values (original prices). As we can see from the plots, the predictions for both datasets using ARIMA model are off the actual trends. Specifically, the prices for Southwest Airlines stocks have a decreasing trend for the forecasting period, however, the forecasted results show a slightly increasing trend. For Netflix, the forecast is an upward straight line, while actual out-of-sample prices experienced two small V-shaped recoveries or mostly just fluctuated around a constant value.

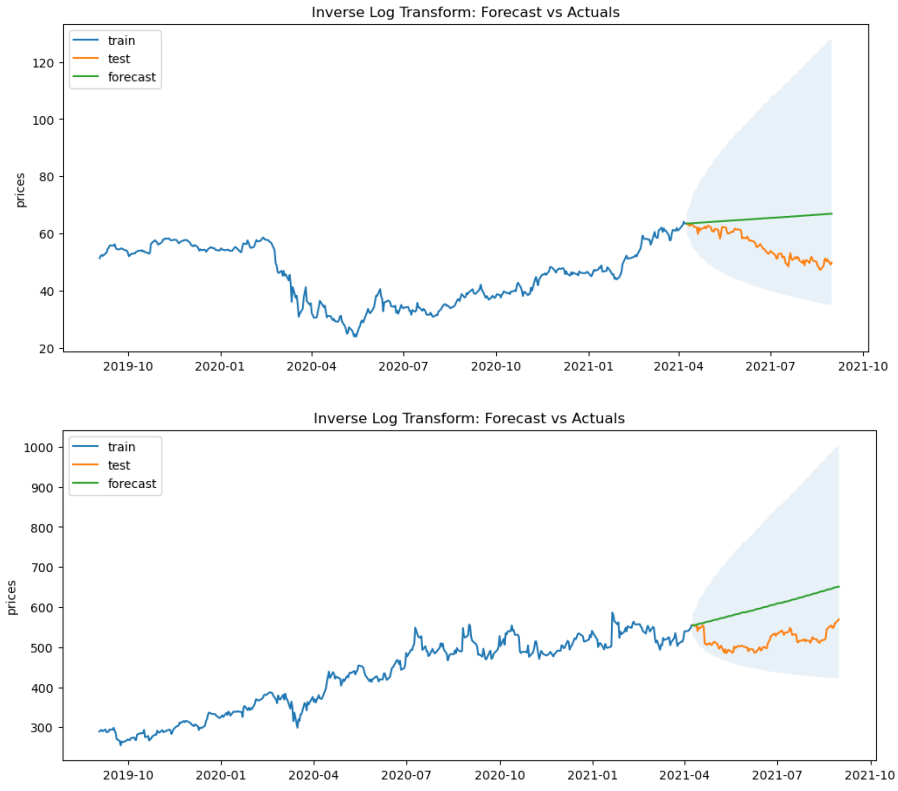


Figure 3 Forecasting

3. LSTM

3.1 Model Selection

3.1.1 Southwest Airlines Dataset

For a better comparison of out-of-sample performance between ARIMA and LSTM, the dataset is split into an 80-20 train-test set for LSTM approach as well, and the values are transformed by MinMaxScaler. When selecting the best LSTM model, different values for parameters are trained to aim for the threshold of a loss function that can converge to or under 0.002. The best LSTM model that generates the lowest MSE for the train set has one layer of LSTM and returns a vector of dimension 4; the optimizer uses SGD (Stochastic gradient descent) at a learning rate of 0.01, and a momentum of 0.9. As shown in *Figure 4*, the loss function decreases over time and eventually converges to 0.001876. The Mean Squared Errors for the train set and the test set are 0.127 and 1.378 accordingly. In comparison, another LSTM model using optimizer 'adam' generates Mean Squared Errors of 0.112 and 3.93 for train and test set, but the loss function does not converge well within 100 epochs and does not forecast the out-of-sample test set as well as the best model.

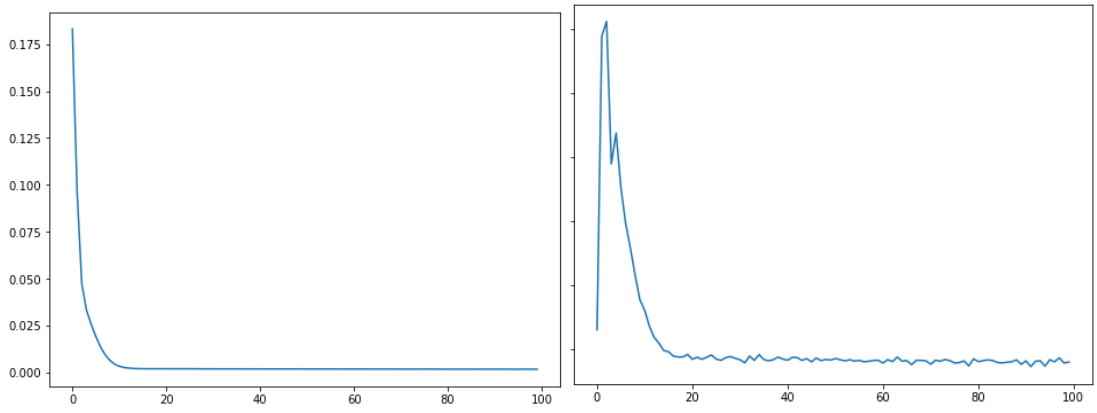


Figure 4 Loss Function for Train (left) and Test (right)

3.1.2 Netflix Dataset

For the Netflix dataset, the same process is performed as Southwest Airline data. The best LSTM model is the same as the one used for Southwest data using SGD as optimizer. The Mean Squared Errors for the train set and the out-of-sample test set are 0.15 and 106.08 accordingly. The Loss Function (*Figure 5*) for the train set decreases over time and converges to 0.00183 around epoch 15. In comparison, the LSTM model using SGD optimizer at a learning rate of 0.02 has Mean Squared Errors 0.149 and 112.6 but loss function for the test set does not converge well and fluctuates over time. The LSTM model using the 'adam' optimizer has MSEs of 0.149 and 161.033 and loss function converges to 0.00246. For the Netflix dataset, LSTM has sign of over fitting when comparing the MSEs for the train set and the test set. But since the forecasting still has high accuracy, this single layer LSTM is still used for analysis.

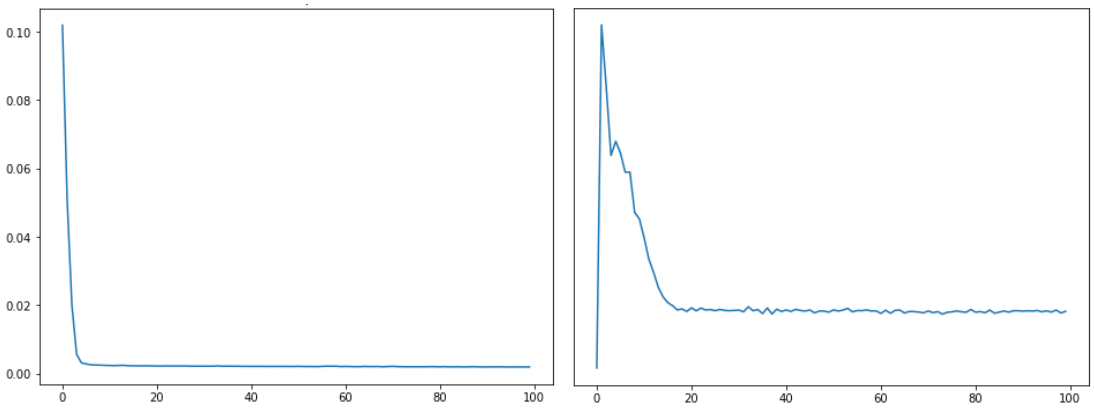


Figure 5 Loss Function for Train (left) and Test (right)

3.2 Out-of-Sample Forecasting Using LSTM

After finding suitable LSTM models for both datasets, a forecasting on the out-of-sample test set is separately performed (*Figure 6*). As we can see from the plots, the forecasting for both datasets using LSTM model predict the actual trends very accurately. Specifically, the predictions have a decreasing trend for the forecasting period, which almost fit the actual price's decreasing trend exactly. For Netflix, the forecast also predicts well the two small V-shaped recoveries, the same as what actual out-of-sample prices experienced. Overall, the LSTM model does a good forecasting on out-of-sample time series data.

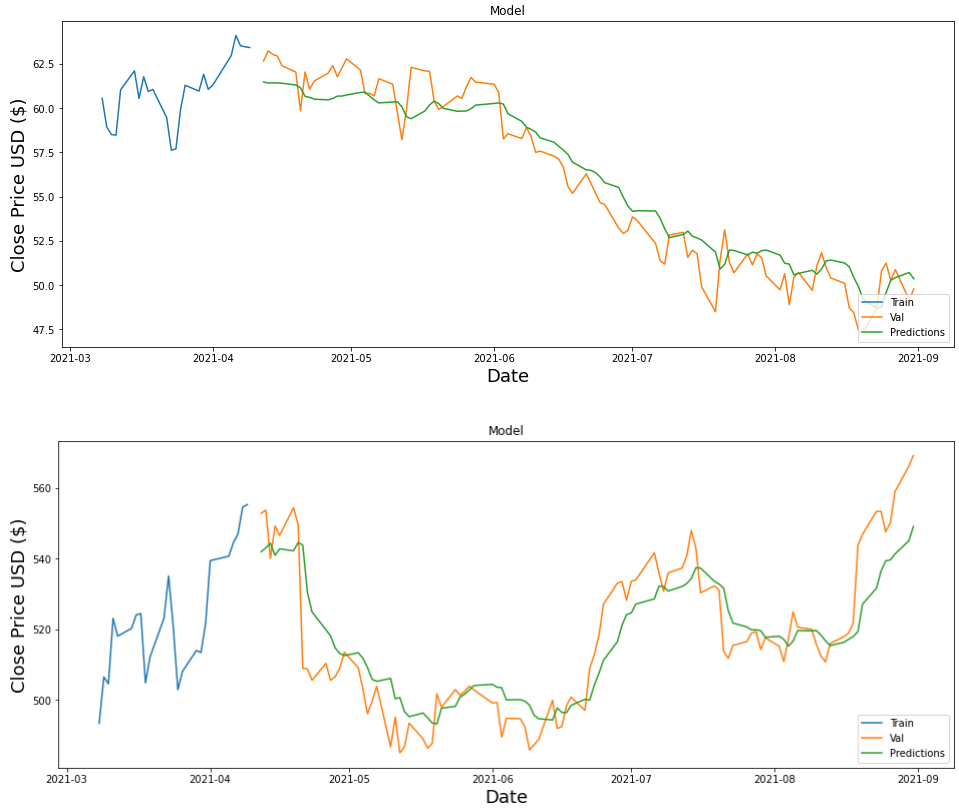


Figure 6 Forecast with LSTM for Southwest (top) and Netflix (bottom)

4. Discussion and Conclusion

Through the analysis from all the previous sections, we can see that LSTM has a much better performance than ARIMA model on forecasting out-of-sample time series data for two datasets, daily stock prices of *Southwest Airlines Co.* and *Netflix Inc.,* even though the datasets have very different pattern and trend and in completely unrelated business industries. The LSTM's forecast is almost exactly the same as the actual prices, while ARIMA models are almost completely off the directions of actual prices for both stocks. If someone follows ARIMA's forecast exactly when planning on stock purchases, it may lead to some great financial loss, while in the opposite, LSTM may become a useful tool and provide some insights when planning on stock purchases, if appropriate parameters and layers are passed.

Although in this project, LSTM has much more outstanding prediction performance than ARIMA model, they still both have their advantages and disadvantages. ARIMA, as the classical time series analysis approach, is much simpler than LSTM and thus has less calculation cost than implementing a more complex model of RNN that can have several stacking layers and hidden components, and when two approaches yield similar outcomes, the simpler one should always be considered first. Also, LSTM often faces the problem of overfitting during training, which requires more comparison between different number of layers and parameters for a better solution. On the other hand, LSTM as a more complex deep learning method, has more potential, has much higher forecasting accuracies, less bias and can achieve desired results if given sufficient data. However, since stock markets can be easily influenced by many other outside factors, such as COVID pandemic, or some government policies, time series analysis and neural networks both have difficulties in recognizing and responding to those unexpected events.

Appendix

Southwest Airlines ARIMA models

ARIMA Model Results

Dep. Variable:	D.Close	No. Observations:	402
Model:	ARIMA(0, 1, 0)	Log Likelihood	801.275
Method:	css	S.D. of innovations	0.033
Date:	Mon, 13 Dec 2021	AIC	-1598.550
Time:	12:03:41	BIC	-1590.557
Sample:	1	HQIC	-1595.386

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0005	0.002	0.319	0.749	-0.003	0.004

ARIMA Model Results

Dep. Variable:	D.Close	No. Observations:	402
Model:	ARIMA(0, 1, 1)	Log Likelihood	801.440
Method:	css-mle	S.D. of innovations	0.033
Date:	Mon, 13 Dec 2021	AIC	-1596.880
Time:	13:10:27	BIC	-1584.891
Sample:	1	HQIC	-1592.133

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0005	0.002	0.328	0.743	-0.003	0.004
ma.L1.D.Close	-0.0269	0.047	-0.573	0.566	-0.119	0.065

Roots

	Real	Imaginary	Modulus	Frequency
MA.1	37.2335	+0.0000j	37.2335	0.0000

ARIMA Model Results

Dep. Variable:	D.Close	No. Observations:	402
Model:	ARIMA(1, 1, 0)	Log Likelihood	801.463
Method:	css-mle	S.D. of innovations	0.033
Date:	Mon, 13 Dec 2021	AIC	-1596.926
Time:	13:09:53	BIC	-1584.936
Sample:	1	HQIC	-1592.179

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0005	0.002	0.329	0.742	-0.003	0.004
ar.L1.D.Close	-0.0305	0.050	-0.613	0.540	-0.128	0.067

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-32.7698	+0.0000j	32.7698	0.5000

ARIMA Model Results

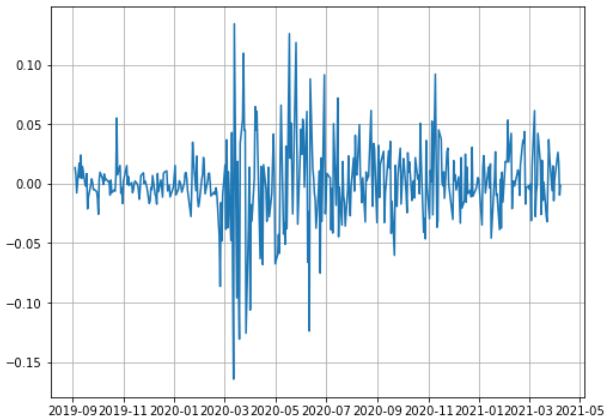
Dep. Variable:	D.Close	No. Observations:	402
Model:	ARIMA(1, 1, 1)	Log Likelihood	801.913
Method:	css-mle	S.D. of innovations	0.033
Date:	Mon, 13 Dec 2021	AIC	-1595.825
Time:	13:10:53	BIC	-1579.839
Sample:	1	HQIC	-1589.496

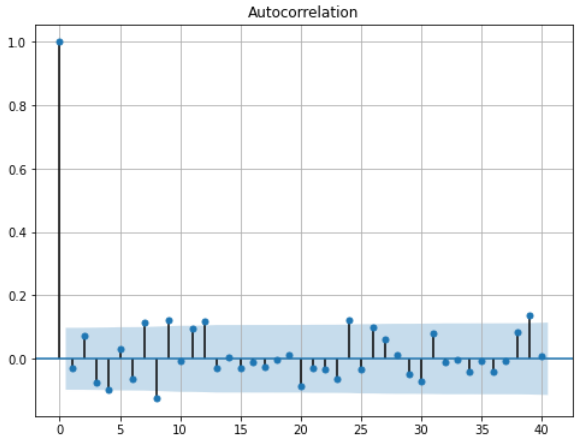
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0005	0.002	0.330	0.741	-0.003	0.004
ar.L1.D.Close	-0.4835	0.270	-1.788	0.074	-1.013	0.046
ma.L1.D.Close	0.4342	0.275	1.578	0.115	-0.105	0.973

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-2.0683	+0.0000j	2.0683	0.5000
MA.1	-2.3030	+0.0000j	2.3030	0.5000

Diagnosis plot for ARIMA (0,1,0):





Netflix ARIMA models

ARIMA Model Results

Dep. Variable:	D.Close	No. Observations:	503
Model:	ARIMA(2, 1, 2)	Log Likelihood	1152.060
Method:	css-mle	S.D. of innovations	0.024
Date:	Mon, 13 Dec 2021	AIC	-2292.120
Time:	01:40:32	BIC	-2266.797
Sample:	1	HQIC	-2282.186

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0013	0.000	2.763	0.006	0.000	0.002
ar.L1.D.Close	0.2825	0.155	1.823	0.068	-0.021	0.586
ar.L2.D.Close	0.5650	0.129	4.389	0.000	0.313	0.817
ma.L1.D.Close	-0.4837	0.168	-2.878	0.004	-0.813	-0.154
ma.L2.D.Close	-0.4499	0.155	-2.898	0.004	-0.754	-0.146

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1037	+0.0000j	1.1037	0.0000
AR.2	-1.6035	+0.0000j	1.6035	0.5000
MA.1	1.0473	+0.0000j	1.0473	0.0000
MA.2	-2.1223	+0.0000j	2.1223	0.5000

ARIMA Model Results

Dep. Variable:	D.Close	No. Observations:	503
Model:	ARIMA(1, 1, 0)	Log Likelihood	1147.919
Method:	css-mle	S.D. of innovations	0.025
Date:	Mon, 13 Dec 2021	AIC	-2289.838
Time:	01:40:33	BIC	-2277.177
Sample:	1	HQIC	-2284.871

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0013	0.001	1.465	0.143	-0.000	0.003
ar.L1.D.Close	-0.2022	0.044	-4.635	0.000	-0.288	-0.117

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-4.9459	+0.0000j	4.9459	0.5000