CDS513 PREDICTIVE BUSINESS ANALYTICS

SEMESTER 2, ACADEMIC YEAR 2022/2023

Assignment 2

Stock Price Prediction – LT Finance Holding LTD

Name: Looi Kah Fung

Matric no: P-COM0049/22

SCHOOL OF COMPUTER SCIENCES

UNIVERSITI SAINS MALAYSIA

Table of Contents

1.0 Intro & Problem Background

Stock prediction and forecasting play a crucial role in financial markets, as investors and traders strive to make informed decision about buying, selling or holding stocks. Predicting the future movement of stock prices is a challenging task that has captivated traders. In this analysis, we delve into stock prediction using the dataset of LT Finance Holding LTD. By leveraging historical price and volume data, the aim of this study is to develop a model that can provide valuable insights and forecasts, aiding investors in their decision-making processes.

By exploring the dataset, we can uncover underlying patterns, trends, and relationships that may help us predict future stock price movement. We will be utilizing the ARIMA (time-series analysis modelling) and Xgboost (machine learning algorithm). The dataset must be a univariate approach – stock price vs time span. We strive to develop a robust forecasting model that can capture the underlying patterns in the stock market and navigate the dynamic landscape of the stock market with greater confidence.

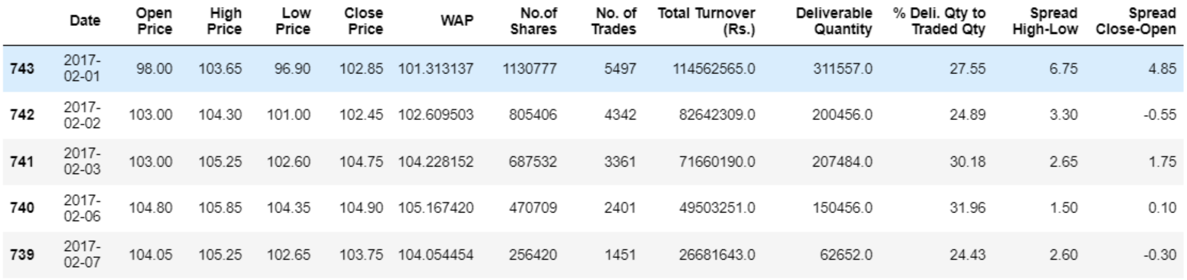
Data interpretation. Date is parsed as date time format %year%month%day and all other attributes are continuous in nature.

1. Date (time span 2017-2020)
2. Open price
3. High price
4. Low price
5. Close price
6. Weighted average price (WAP)
7. No. of shares
8. No. of trades
9. Total turnover (Rs.)
10. Deliverable Quantity
11. % Deli. Qty to Traded Qty
12. Spread High- Low
13. Spread Close-Open

In the context of the stock market, the closing price reflects the final price at which a stock, typically at the end of the trading day. It is widely regarded as the most important price point for time-series analysis and forecasting. The closing price incorporates all the market information and investor sentiment throughout the trading session, while open, high, low prices provide valuable insights into intraday price movements, they may not capture the full sentiment and market dynamics of the entire trading session. By focusing on close price, it allows for the identification of trends, patterns, and seasonality.

2.0 Data Plotting

Figure 2.1 shows Data Frame of the LT Finance Holding LTD stock price wherein Date has been parsed.



	Date	Open Price	High Price	Low Price	Close Price	WAP	No.of Shares	No. of Trades	Total Turnover (Rs.)	Deliverable Quantity	% Deli. Qty to Traded Qty	Spread High-Low	Spread Close-Open
743	2017-02-01	98.00	103.65	96.90	102.85	101.313137	1130777	5497	114562565.0	311557.0	27.55	6.75	4.85
742	2017-02-02	103.00	104.30	101.00	102.45	102.609503	805406	4342	82642309.0	200456.0	24.89	3.30	-0.55
741	2017-02-03	103.00	105.25	102.60	104.75	104.228152	687532	3361	71660190.0	207484.0	30.18	2.65	1.75
740	2017-02-06	104.80	105.85	104.35	104.90	105.167420	470709	2401	49503251.0	150456.0	31.96	1.50	0.10
739	2017-02-07	104.05	105.25	102.65	103.75	104.054454	256420	1451	26681643.0	62652.0	24.43	2.60	-0.30

Figure 2.1 first 5 rows of LT Finance Holding LTD dataset.

Figure 2.2 shows the time series portrayed a seasonality trend when the trend going up from 2017 Jan to 2017 Sept and gradually trending down till 2020 Jan, within these intervals, there is a seasonal up and down trend.

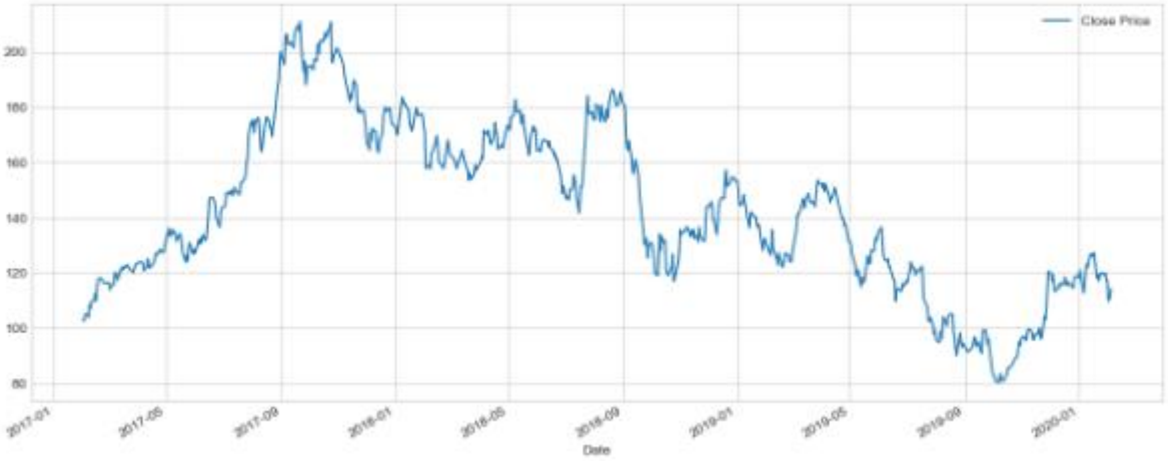
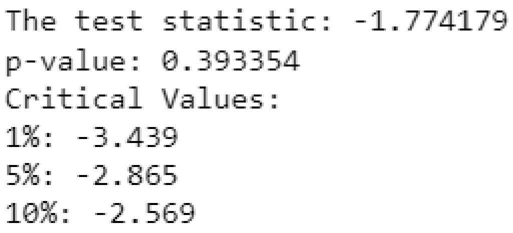


Figure 2.2 Time Series Trend

3.0 Data Pre-processing/ Differencing

The time series dataset has been checked for stationarity with no pre-processing at the first place. The Augmented Dickey-Fuller (ADF) test used in time series analysis to determine whether a given time series is stationary or not. In the provided summary of the ADF test, the test statistic is -1.774179. The p-value associated with the ADF test is 0.393354. In this case, the p-value of 0.393354 is greater than the significance level of 0.05. It is found that the time series dataset to be non-stationary without differencing.

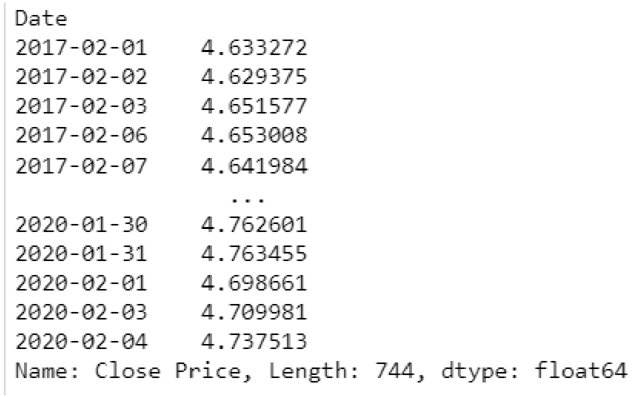


```
The test statistic: -1.774179
p-value: 0.393354
Critical Values:
1%: -3.439
5%: -2.865
10%: -2.569
```

Figure 3.1 ADF Test

Differenced the data set.

First-order differencing is used to transform a non-stationary time series into a stationary one. In involves computing the difference between consecutive observations in the series. In our case, the lagging is one order. From the data frame, we can have a glance of the time series having constant mean, variance, autocovariance over time.



```
Date
2017-02-01    4.633272
2017-02-02    4.629375
2017-02-03    4.651577
2017-02-06    4.653008
2017-02-07    4.641984
                ...
2020-01-30    4.762601
2020-01-31    4.763455
2020-02-01    4.698661
2020-02-03    4.709981
2020-02-04    4.737513
Name: Close Price, Length: 744, dtype: float64
```

Figure 3.2 First-order differencing data frame

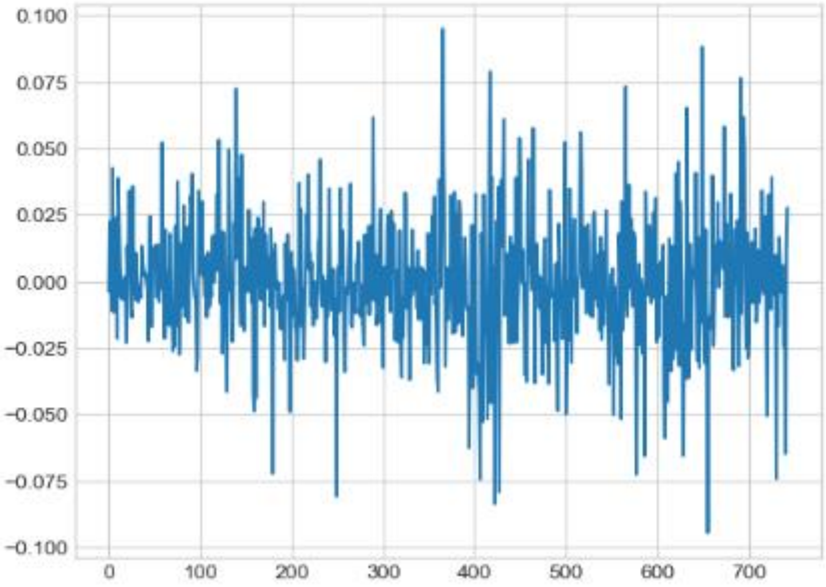
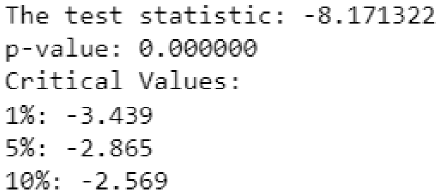


Figure 3.3 First-order differencing plot



```
The test statistic: -8.171322
p-value: 0.000000
Critical Values:
1%: -3.439
5%: -2.865
10%: -2.569
```

t

Figure 3.4 ADF test after first-order differencing

ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots are used in time series analysis to understand the autocorrelation structure of a time series and to determine the appropriate orders for autoregressive (AR) and moving average (MA) models. After first-order differencing, the p,d,q notion, d =1, we observed that ACF values decay slowly and slightly converge to the significance bounds up until 40 lags whereas PACF values decay abruptly and cuts off after lag 1. We can deduce that the ARIMA model is (1,1,0). ACF plot implies high complexity, since all the 40 lags are remaining significant, it has high degree of autocorrelation in the time series. In our case, we would like the model to be as simple as possible.

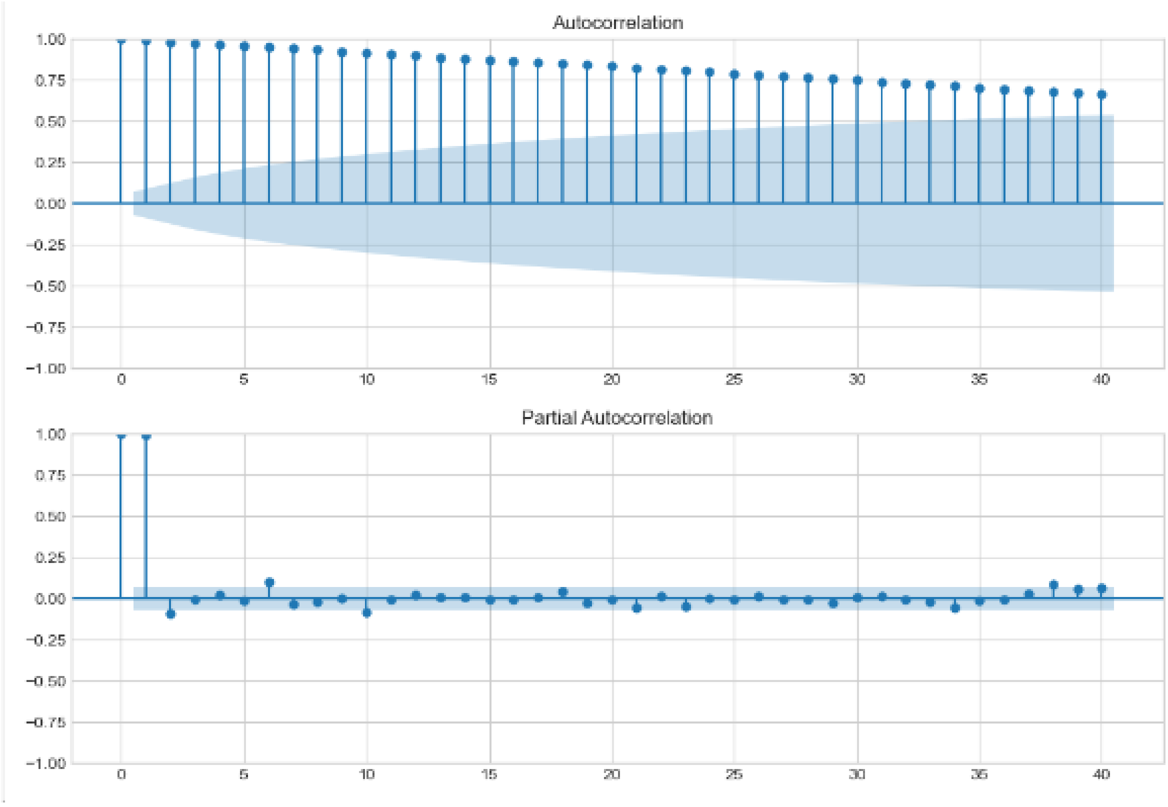
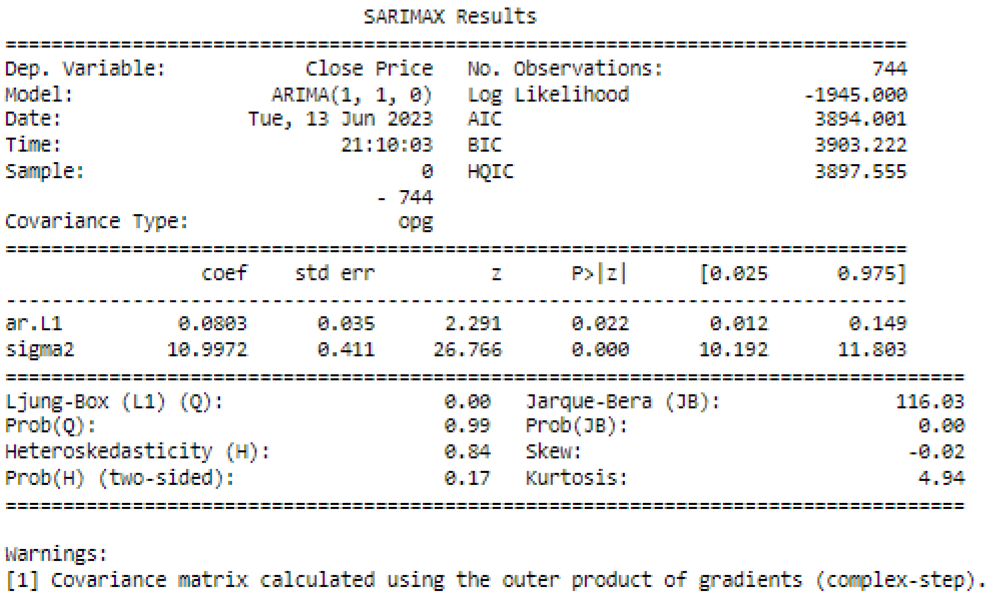


Figure 3.5 ACF & PACF plots

Auto ARIMA

The model is specified as ARIMA (1,1,0) indicating that it includes an autoregressive component of order 1 and a differencing of order 1, but no moving average component. Next important statistical measure is AIC (Akaike Information Criterion), the AIC value for the model is given as 3894.001. It is generally desired to have a lower AIC value, a higher AIC value could indicate the model is not fitting well.



```
                               SARIMAX Results
==============================================================================
Dep. Variable:            Close Price   No. Observations:                  744
Model:                 ARIMA(1, 1, 0)   Log Likelihood               -1945.000
Date:                Tue, 13 Jun 2023   AIC                           3894.001
Time:                        21:10:03   BIC                           3903.222
Sample:                             0   HQIC                          3897.555
                                - 744
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0803      0.035      2.291      0.022       0.012       0.149
sigma2        10.9972      0.411     26.766      0.000      10.192      11.803
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):               116.03
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               0.84   Skew:                            -0.02
Prob(H) (two-sided):                  0.17   Kurtosis:                         4.94
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Figure 3.6 SARIMAX result

ARIMA prediction

The ARIMA (1,1,0) is put on the prediction on beginning 100 data points.

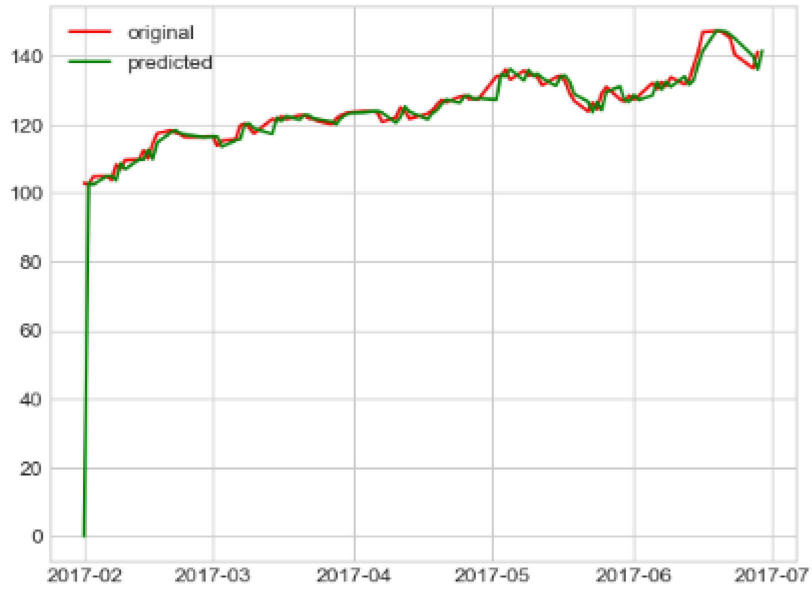


Figure 3.7 ARIMA (1,1,0) prediction on first 100 data points

The ARIMA (1,1,0) is put on the prediction on last 43 data points and zoom out of the time series trend.

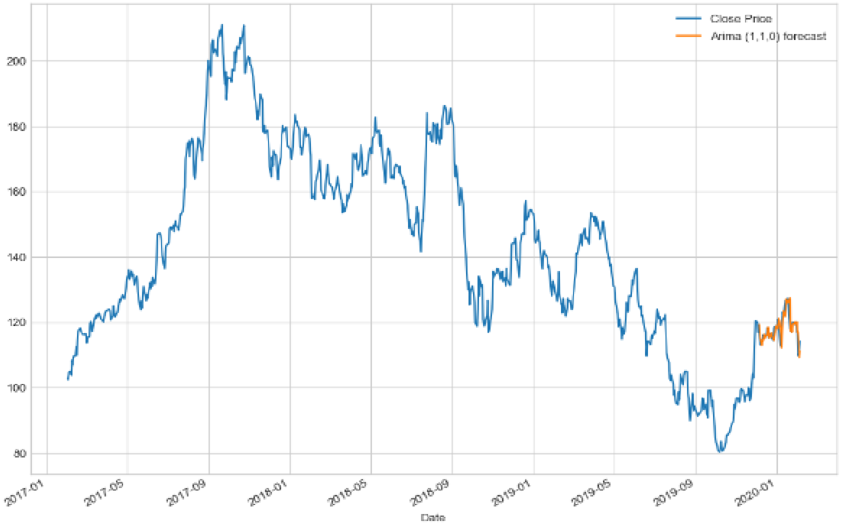


Figure 3.8 ARIMA (1,1,0) prediction on last 43 data points

Xgoobst – Machine learning algorithm

The conceptual framework of applying machine learning algorithm to time series analysis, the goal is to forecast future values based on historical data. The time series dataset is transformed into a supervised learning dataset. Next, the dataset is split into training and test sets by specifying the number of samples to be included in the test set.

Xgboost prediction

The walk forward validation is used to evaluate the performance of of Xgboost on time series data. The walk forward validation involves training the model sequentially on a rolling window of data and making predictions on a subsequent window. The iterative process allows for the model to be retrained and evaluated multiple times, incorporating new information as it becomes available. In our case, it has been iteratively up to 11 times until the end of the dataset is reached.

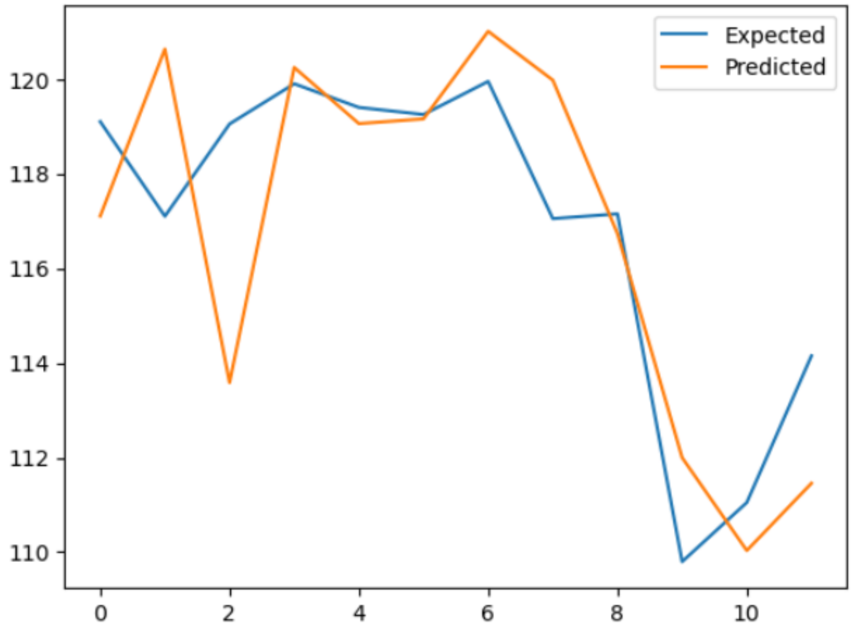


Figure 3.9 Prediction value of Xgboost

4.0 Model Comparison & Analysis

We have computed the time series analysis using 2 distinct approaches which are time series modelling, ARIMA and machine learning algorithm, Xgboost. We compared the performance of two models in forecasting a specific time series dataset. The evaluation is based on 3 commonly used metrics: RMSE, MAE and MSE. ARIMA (1,1,0) model is observed that it achieves an RMSE of 45.230, indicating that, on average, the forecasted value deviate by approximately 45.230 from the actual values, the MAE of 38.259 suggests that the average absolute difference between the predicted and actual values is 38.259. Additionally, the MSE of 204.801 demonstrates that the squared differences between predicted and actual values average to 204.801.

On the other hand, the Xgboost model showcases superior performance in all evaluation metrics. With an RMSE of 1.840, the Xgboost model outperforms the ARIMA model by a significant margin, indicating its ability to provide more precise forecasts and other metrics such as MAE and MSE are substantially lower than ARIMA (1,1,0).

When we delve into the technical depth of the attributes of Xgboost and ARIMA, we can roughly gain some insights why Xgboost outperforms ARIMA for our dataset. Firstly, Xgboost can handle non-linear relationship whereas ARIMA is a linear model that assumes linear relationship between past and future observation. Xgboost leverages an ensemble of decision trees, allowing it to model and capture intricate interactions between variables, resulting in improved forecasting accuracy. Secondly, Xgboost learn and model complex seasonality and trend patterns autonomously. It can adaptively capture and exploilt temporal dependencies and non-stationarity in the data. Thirdly, Xgboost features engineering by considering a broader range of input variables and their interactions while ARIMA in our case only considers a univariate modelling, close price vs time-series.

Table 4.1 Evaluation metrics for time series modelling and machine learning algorithm

Evaluation Metric	Time series modelling	Machine learning algorithm
	ARIMA (1,1,0)	Xgboost
Root means square error (RMSE)	45.230	1.840
Mean absolute error (MAE)	38.259	5.798
Mean squared error (MSE)	204.801	2.408

5.0 Conclusion

The comparison and analysis of the ARIMA (1,1,0) and Xgboost models highlight the latter’s superiority in forecasting the given LT Finance Holding LTD stock price prediction dataset. The significantly lower RMSE, MAE and MSE values are the fool proof of the superiority of machine learning algorithm in the context of time-series analysis. In future work, seasonality (SARIMA) or exogenous variables (ARIMAX) accounting the seasonality and trend into the modelling. SARIMA might provide more hidden insight or hindsight on the time series dataset. As it has shown in the ACF and PACF plot, the plot is non-stationarity when differencing is 0.