Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia

Vol. 18, 2023, page 39 – 47

Time Series Analysis On Sales Data of Petroleum in Malaysia

Nur Syiffa Sam, Adina Najwa Kamarudin*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: adina.najwa@utm.my

Abstract
In Malaysia, the energy sector is an important driver of economic growth. We all know that oil and natural gas are the two major industries of the energy market. Malaysia is the second-largest oil producer in Southeast Asia and the world's third largest exporter of liquefied natural gas (LNG). Petronas has held exclusive ownership rights over all exploration and production activities related to oil and natural gas in the country since its inception. Throughout the years, there are several issues that affect the sales data of petroleum in Malaysia since the prices are influenced by global supply and demand. This project is based on the time series analysis of monthly sales of petroleum in Petronas Malaysia from the year of 2017 to 2022. The significance of this project is to study the trend and determine the best forecasting model between Exponential Smoothing and Autoregressive Integrated Moving Average (ARIMA) for monthly sales of petroleum in Petronas. The performance of the models are then compared by the values of Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The outcome showed that, in comparison to Exponential Smoothing, the chosen ARIMA model forecasts more accurately.

Keywords: Liquefied natural gas; Petronas Malaysia; petroleum; Exponential Smoothing; ARIMA

1. Introduction

Shell is one of the top petroleum companies in Malaysia, having located and drilled the first oil well near Malaysia on Canada Hill in Miri, Sarawak in 1910. By December 1910, it was producing 83 barrels per day (bbls/d). Following Malaysian independence from the UK in 1957, the oil and gas industry became subject to the Petroleum Mining Act 1966 (Act 95). At first, Shell was the main operator, followed by Esso, who controlled upstream production, downstream refining, and downstream sales. Esso and Conoco were granted oil and gas concessions off the east coast of the Peninsular after other international petroleum corporations went to offshore Peninsular Malaysia in the late 1960s.

When Petronas was founded, four of the 19 oil fields that had been discovered in Malaysia were already producing 90,000 to 99,000 bbls/d. Petronas was established in 1974 under the 1965 Companies Act as a result of economic and political pressures and rising economic nationalism in Malaysia. The 1974 Petroleum Development Act gave Petronas ownership, exclusive rights, and authority over Malaysia's hydrocarbon resources. The importance of Petronas as a source of money for the government has increased significantly ever since. However, the chances of tapping this resource to support public expenditure, continually raise GDP, and encourage local entrepreneurship are quickly dwindling due to a number of factors, including the structural decline of the fossil fuel industry, the periodic price shocks, and the expanding size of the Malaysian economy compared to the O&G sector.

This has put a lot of pressure on Petronas, the government-owned oil company, to keep making large payments to the public coffers. GlobalMonitor (2020) claims that the first half of 2020 saw exceptional occurrences that had never before been seen in the sector that marked the need for oil and gas in Malaysia. Malaysia has been compelled by the proliferation of COVID-19 to issue lockdown orders, travel bans, and remain at home directives. The use of automobiles, industry, shipping, and transportation has significantly decreased as a result of movement restrictions. This had an immediate effect on the demand for gas and oil in Malaysia. In Malaysia, the demand collapse was predicted to be greater than 30% in April 2020 (Global Monitor, 2020). The rapid decline in crude oil prices was caused by a lack of physical demand.

2. Time Series Forecasting

In today's highly competitive world, accurate sales forecasting plays a vital role in every successful business. It is a prognosis of what will happen in a specific sector during a specific time period, which can help avoid overproduction and reduce overstock (Yasaman Ensafi et al.,2022; Islam & Amin, 2020; Nguyen et al.,2021). Additionally, by examining the sales trends of a company's overall sales or the sales of a particular product, it is possible to identify certain key characteristics that will have an influence on future sales. Time-series forecasting issues cannot be resolved with a single optimum method, so a variety of strategies may be used (Yasaman Ensafi et al.,2022; Zhang & Kline, 2007). The approaches, for instance, include ARIMA, ANN, Holt-Winters method, exponential smoothing, Grey-Markov method and structural time series models (Haiges et al., 2017).

2.1 Petroleum Sales Forecasting

Natalie Burclaff (2022) argues that oil and natural gas are significant players in the energy sector, with production and transportation of oil and gas being complex, expensive and reliant on cutting-edge technology. Oil and gas industry is often divided into three segments:

1. **upstream**, the business of oil and gas exploration and production;
2. **midstream**, transportation and storage; and
3. **downstream**, which includes refining and marketing.

Petronas is a fully integrated oil and gas firm that is entirely controlled by the government of Malaysia. It is the largest agency in Malaysia, with 103 wholly-owned subsidiaries, 19 partially owned garments, and 57 associated companies. D.M.H Kee et al. (2020) also claims that state oil provides a significant source of revenue for the Malaysian government, with half of the government budget dependent on dividends and a real government balance of 5% of the gross domestic product in 2011. Department of Statistics Malaysia (2022) stated that the sales value of manufactured refined petroleum products in Malaysia as of 2021, was approximately 175.56 billion Malaysian ringgit. Additionally, Petronas has posted a strong performance for the first half of 2022 due to elevated oil and gas prices arising from a widening supply gap and dwindling system capacity.

The Malaysian oil and gas market is expected to record a Compound Annual Growth Rate (CAGR) of 2.5% during the forecast period due to the outbreak of COVID-19 in Q1 of 2020 (Malaysia Oil and Gas Market Outlook (2022 - 27), 2022). Lack of physical demand for crude oil saw prices collapse quickly. The high capital investment required, coupled with a lack of financing due to a global economic slowdown are expected to hinder the market's growth. Meanwhile, Bernama (2022) argued in New Straits Times that, MIDF Amanah Investment Bank Bhd expects the government to record higher-than-expected petroleum-related revenue in 2022 at RM73.7 billion due to the average Brent crude oil price at US$66 per barrel this year. Fuel inflation is predicted to surge to 10.7 per cent year-on-year and headline inflation to rise to 3.6 per cent.

2.2 Forecasting Using Machine Learning

Machine learning and deep learning approaches are useful for precise forecasting. Deep neural networks are designed to model more complex functions by employing Multilayer Perceptrons (MLP) with more hidden layers (Ensafi et al., 2022; Alpaydin, 2009). Artificial Neural Networks (ANN)-based approaches have been used in several studies, such as stock price prediction (Zhuge et al., 2017) and short-term traffic forecasting (Zhao et al., 2017). To establish the mapping of inputs to outputs, artificial neural networks join several neurons (input/output units) in multiple layers. Each unit (neuron) will have a weight that will be adjusted throughout the training phase, and a weighted network with the fewest neurons can convert inputs into outputs with the least fitting error deviation.

Ensafi et al. (2022) and Kaneko and Yada (2016) presented a model for predicting sales of a retail store using a deep learning approach. Tkáč & Verner (2016) claimed that ANN models have been popular techniques for sales forecasting due to their flexibility for detecting patterns in data. Thivakaran et al. (2022) argued that data collection, data processing, feature development, model generation, and testing are utilized to predict sales using a random forest regressor XG-booster approach. Pavlyshenko & Bohdan M. (2019) proved that stacking approaches can improve the performance of predictive models for sales time series forecasting.

2.3 Autoregressive Integrated Moving Average

The ARIMA model is a type of regression analysis that evaluates the relative importance of one dependent variable to other varying variables. It consists of three components: Autoregressive (AR), Integrated (I) and Moving Average (MA). ARIMA (*p,d,q*) is generally written with *p* denoting the order of the autoregressive part, *d* as a measure of the degree of difference, and *q* as the order of the moving average part. In the autoregressive part of the process, the stationarity of the data is tested. If the data is stationary, a differencing process is required. The data is also tested for its moving average fit in part q of the analysis process. Initial analysis of the data prepares it for forecasting by determining parameters. The forecasting equation of ARIMA model can be expressed as:

$$y_t = c + \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{i=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{1}$$

The four steps needed to forecast using the ARIMA model are identification of models, estimation of parameters, diagnostic checking (tested for its adequacy) and model selection. Identification of models requires the data to be stationary and visual inspection of both autocorrelation function (ACF) and partial autocorrelation function (PACF). Upon model selection, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) criteria can be used to choose the best model. AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, while BIC is an estimate of the posterior probability of a model being true under a certain Bayesian setup. Ideally, the AIC and BIC value should be as small as possible. The formulation of AIC and BIC can be expressed as:

$$AIC = \ln \hat{\sigma}^2 + \frac{2}{n} r \tag{2}$$

$$BIC = \ln \hat{\sigma}^2 + \frac{\ln n}{n} r \tag{3}$$

$$\hat{\sigma}^2 = \frac{SSE}{n} \tag{4}$$

2.4 Exponential Smoothing

Exponential smoothing is a time series forecasting technique that can handle data with systematic trends or seasonal components. It can be divided into three types: seasonality support, an extension that explicitly handles trends, and a simple method that assumes no systematic structure.

2.4.1 Single Exponential Smoothing

Time series forecasting using single exponential smoothing (SES) is a method that works on univariate data without trend or seasonality. It requires only one parameter, alpha, which controls the exponential decay rate of the influence of previous observations. Formally, the SES model is:

$$F_{t+1} = \alpha y_t + (1-\alpha)F_t \tag{5}$$

2.4.2 Double Exponential Smoothing

Double Exponential Smoothing, also known as Holt's Exponential Smoothing, is a variation on Exponential Smoothing that supports trends in univariate time series. Beta is added to control the decay of the influence of the trend change. The approach promotes trends that evolve in several ways:

- **Additive Trend**: Double Exponential Smoothing with a linear trend.
- **Multiplicative Trend**: Double Exponential Smoothing with an exponential trend.
  The Holt's Smoothing model is

$$F_{t+m} = L_t + mb_t \tag{6}$$
$$L_t = \alpha y_t + (1-\alpha)(L_{t-1} + b_{t-1})$$
$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

2.4.3 Triple Exponential Smoothing

Triple Exponential Smoothing is an extension of Exponential Smoothing that supports seasonality in univariate time series by adding gamma parameters as smoothing factors. As with the trend the seasonality may be modeled as:

- Additive seasonality: Triple exponential smoothing with linear seasonality

$$F_{t+m} = L_t + mb_t + S_{t+m-s} \tag{7}$$
$$L_t = \alpha(y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1}$$
$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$
$$S_t = \gamma(y_t - L_t) + (1-\gamma)S_{t-s}$$

- Multiplicative Seasonality: Triple Exponential Smoothing with an exponential seasonality

$$F_{t+m} = (L_t + mb_t)S_{t+m-s} \tag{8}$$
$$L_t = \alpha\frac{y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1}$$
$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$
$$S_t = \gamma\frac{y_t}{L_t} + (1-\gamma)S_{t-s}$$

3. Model Comparison

The Model Comparison tool compares predictive models based on their performance with validation or test data. An error measurement and prediction table are generated, including RMSE, MAE, and MAPE. The model with the lowest error is the most suitable.

$$RMSE = \sqrt{\sum_{t=1}^{n}\frac{(y_t - F_t)^2}{n}} \tag{9}$$

$$MAE = \sum_{t=1}^{n} \frac{|y_t - F_t|}{n} \tag{10}$$

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \frac{|y_t - F_t|}{y_t} \tag{11}$$

4. Results and Discussion

This study collects data from Petronas Sdn Bhd Malaysia from October 2017 to October 2022, with 10% of the data used as an out sample. In sample data from October 2017 until April 2022 are used for the initial parameter estimation and model selection, while the out sample data from May 2021 until October 2022 are used to evaluate forecasting performance. The Exponential Smoothing and ARIMA model is constructed using software such as Microsoft Excel, RStudio and Minitab. The model is used to predict future petroleum sales for November 2022 and December 2022.

4.1 Exponential smoothing approach

The components of the time series were identified using various tests in order to identify whether Exponential Smoothing model is appropriate for usage. The results of the Mann-Kendall test in Rstudio show that the time series exhibits a pattern. The Autocorrelation Function plot is used to test for seasonality, and the findings show that there is no seasonality in the data. Since the data has trend and no seasonality, Holt's Exponential Smoothing method is suitable to fit the data.
The option of *Optimal ARIMA* in *Double Exp Smoothing* function in Minitab would calculate the first values of level, α and trend, γ. Figure 1 below suggests that the value of smoothing constant of level, $\alpha = 0.5359$ and trend, $\gamma = 0.0498$ is the most suitable values that would minimize the sum of square errors. Table 2 shows the results of RMSE, MAPE and MAE to assess the performance and accuracy of Holt's Exponential Smoothing.

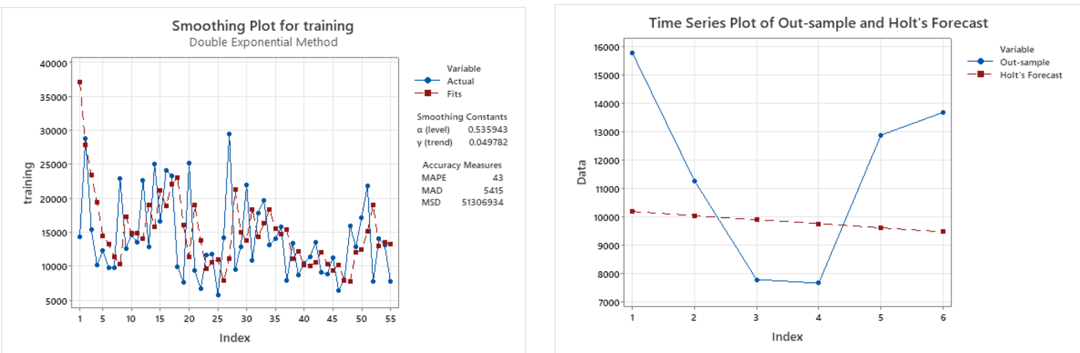

Figure 1: Time series plot of In-sample (left side) and Out-sample (right side) data and Holt's Smoothing forecast

Holt's Exponential Smoothing model was used to forecast for 6 months period and compared with out-sample data. The model showed a decreasing trend line compared to the out-sample data, which was assessed by RMSE, MAPE and MAE values.

4.2 ARIMA approach

ARIMA model is a time series forecasting method that uses historical data to predict future trends. Three stages are needed to select the best model for forecasting: model identification, selection and diagnostic checking.

Checking for stationarity of data is an important step before applying the ARIMA approach. There are various methods used to check for stationarity, such as the Augmented Dickey-Fuller (ADF) test, where the p-value is compared to the significance value of 0.05 and the stationarity of the data is determined. If the p-value is smaller than the significance value, then the null hypothesis is rejected and the time series data is stationary. If the p-value is greater than the significance value, then differencing needs to be carried out. The ADF test for petroleum sales using RStudio showed that the p-value was greater than the significance value of 0.05, indicating that the data was not stationary. Therefore, differencing is needed.

Differencing is a method of transforming a non-stationary time series into a stationary one. The first differencing value is the difference between the current and previous time periods. The plot of the first difference of petroleum sales suggests the series is stationary, and the three ARIMA(*p,d,q*) models to be used are ARIMA(2,1,1), ARIMA(1,1,1) and ARIMA(2,1,0).

The ACF plot of residuals is used to test for autocorrelation, and if most of the spikes in the ACF plot are within the confidence interval of 95%, then there is no autocorrelation in the residuals. The ACF plots of residuals for ARIMA(2,1,1), ARIMA(1,1,1) and ARIMA(2,1,0) show that most of the lags in each plot are within the threshold limit. This suggests that the residuals for all three models have the characteristics of white noise in which they have no autocorrelation.

The null hypothesis is that the term is not significantly different from 0, which indicates that no association exists between the term and the response. The residuals are tested for independence using Ljung-Box chi-square statistics. The overall findings show that ARIMA(2,1,1) and ARIMA(2,1,0) both have independent residuals and are regarded as adequate models. After computing AIC and BIC values, the model with the lowest AIC and BIC value is chosen as the best fit. The independent residuals assumption was also met by this model, making ARIMA(2,1,0) the best option for data fitting.

Table 1: ARIMA model selection using t-test, Q-test, AIC and BIC values

<table>
<tr><th>ARIMA</th><th>t-test</th><th>Q-test</th><th>AIC</th><th>BIC</th></tr>
<tr><td>(2,1,1)</td><td>Final Estimates of Parameters
Type Coef SE Coef T-Value P-Value
AR 1 -0.652 0.205 -3.18 0.002
AR 2 -0.527 0.150 -3.51 0.001
MA 1 0.154 0.238 0.65 0.520
Constant -328 697 -0.47 0.640
Mean Average (MA) term has p-value > 0.05, so the coefficient is not statistically significant.</td><td>Modified Box-Pierce (Ljung-Box)
Lag 12 24 36 48
Chi-Square 14.69 26.11 34.27 47.07
DF 8 20 32 44
P-Value 0.066 0.162 0.359 0.348
The p-values for the test are all greater than 0.05. The residuals are independent.</td><td>17.4880</td><td>17.6353</td></tr>
<tr><td>(1,1,1)</td><td>Final Estimates of Parameters
Type Coef SE Coef T-Value P-Value
AR 1 -0.053 0.145 -0.36 0.719
MA 1 0.9896 0.0741 13.36 0.000
Constant -81.2 38.2 -2.13 0.038
Autoregressive (AR) term has p-value > 0.05, so the coefficient is not statistically significant.</td><td>Modified Box-Pierce (Ljung-Box)
Lag 12 24 36 48
Chi-Square 23.42 48.07 64.21 75.11
DF 9 21 33 45
P-Value 0.005 0.001 0.001 0.003
All p-values for the test are smaller than 0.05. So, the residuals are not independent.</td><td>17.4051</td><td>17.5156</td></tr>
<tr><td>(2,1,0)</td><td>Final Estimates of Parameters
Type Coef SE Coef T-Value P-Value
AR 1 -0.760 0.115 -6.60 0.000
AR 2 -0.578 0.115 -5.05 0.000
Constant -369 818 -0.45 0.654
AR terms has p-value <0.05, so the coefficient is said to be statistically significant.</td><td>Modified Box-Pierce (Ljung-Box)
Lag 12 24 36 48
Chi-Square 15.55 26.18 33.40 44.62
DF 9 21 33 45
P-Value 0.077 0.200 0.448 0.488
The p-values for the test are all greater than 0.05. The residuals are independent.</td><td>17.4562</td><td>17.5667</td></tr>
</table>

The time series plot of the in-sample data and the fitted ARIMA(2,1,0) model in Figure 2 shows that the model fits the data well. RMSE, MAPE and MAE are used to measure forecasting accuracy for in-sample data. Out-sample data is used to measure how well the model forecasts on new data, and forecast accuracy metrics are calculated using the fitted value of the in-sample period (Table 2).

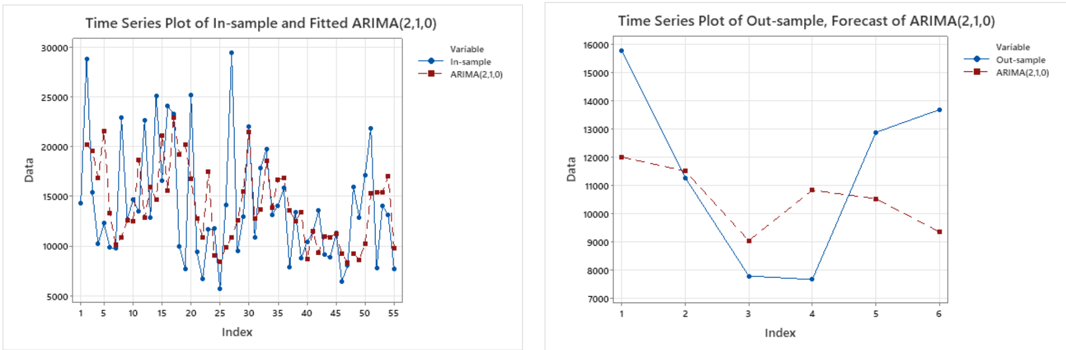

Figure 2: Time series plot of In-sample (left side) and Out-sample (right side) data and fitted values of ARIMA(2,1,0)

4.3 Model comparison

The performance of Holt's Smoothing and ARIMA(2,1,0) are evaluated by the whole sets. Table 2 shows the performance results of both models. Based on the results, ARIMA(2,1,0) produces more accurate forecasts than Holt's Exponential Smoothing due to smaller values of RMSE, MAPE and MAE. To visualize the individual point forecasting behavior, the actual data of in-sample and out-sample versus the forecasts from ARIMA(2,1,0) and Holt's Exponential Smoothing model are plotted in Figure 3.

Table 2: Performance analysis comparison of both ARIMA(2,1,0) and Holt's Exponential

	In-sample		**Out-sample**	
	ARIMA(2,1,0)	**Holt's Smoothing**	**ARIMA(2,1,0)**	**Holt's Smoothing**
RMSE	5840.2629	7162.8859	2889.3517	3412.8623
MAPE	33.8950	43.2913	22.1727	26.0960
MAE	4478.5697	5415.4661	2515.4974	3080.3144

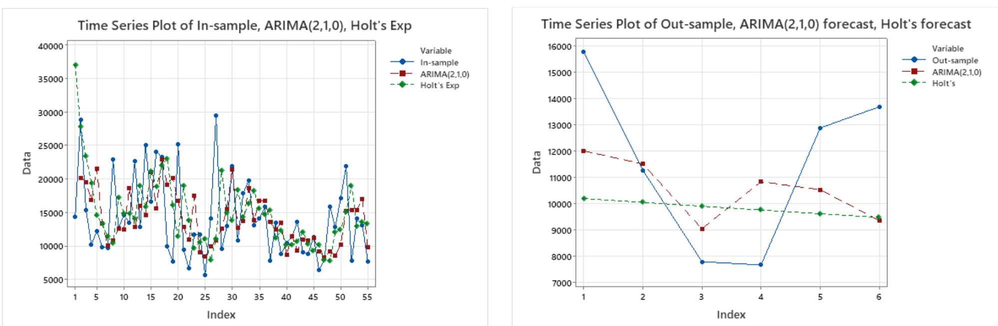

Figure 3: Time series plot of fitted values of in-sample (left side) and out-sample (right side) of ARIMA(2,1,0), Holt's Exponential Smoothing and actual data.

Conclusion

This study investigates the relative forecast performances of Exponential Smoothing and ARIMA modeling frameworks when they are applied to petroleum sales forecasting. The results showed that ARIMA(2,1,0) forecasts more accurately than Holt's Exponential Smoothing, regardless of the forecast error measure considered. The MAPE value of ARIMA(2,1,0) is 22.17%, which is lower than the MAPE of Holt's Exponential Smoothing, 26.1%. The value of RMSE and MAE in ARIMA(2,1,0) is 15% and 18% smaller, respectively as compared to Holt's Exponential Smoothing. These results indicate that both models are reasonable forecasting models since MAPE values are between 21%-50%.

References

[1] Bernama. (2022, June 17). MIDF expects Govt to record rm73.7bil petroleum-related revenue. New Straits Times. Retrieved November 16, 2022, from https://www.nst.com.my/business/2022/06/805718/midf-expects-govt-record-rm737bil-petroleum-related-revenue

[2] Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning – a comparative analysis. *International Journal of Information Management Data Insights*, *2*(1), 100058. https://doi.org/10.1016/j.jjimei.2022.100058

[3] Haiges, R., Wang, Y. D., Ghoshray, A., & Roskilly, A. P. (2017). Forecasting electricity generation capacity in Malaysia: An auto regressive integrated moving average approach. *Energy Procedia*, *105*, 3471–3478. https://doi.org/10.1016/j.egypro.2017.03.795

[4] Kashtanov, K., Kashevnik, A., & Shilov, N. (2022). Using triple exponential smoothing and autoregressive models to mining equipment details sales forecast. *Communications in Computer and Information Science*, 506–521. https://doi.org/10.1007/978-3-030-93715-7_36

[5] *Malaysia Oil and Gas Market Outlook (2022 - 27): Share, Trends*. Malaysia Oil and Gas Market Outlook (2022 - 27) | Share, Trends. (n.d.). Retrieved November 23, 2022, from https://www.mordorintelligence.com/industry-reports/malaysia-oil-and-gas-market

[6] *Malaysia oil and gas market report with covid-19 impact analysis (2020-2025)*. GlobalMonitor. (n.d.). Retrieved November 28, 2022, from https://www.globalmonitor.us//product/malaysia-oil-and-gas-market-report

[7] Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and Arima models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing*, *34*, 151–163. https://doi.org/10.1016/j.rcim.2014.12.015

[8] Thivakaran, T. K., & Ramesh, M. (2022). Exploratory Data Analysis and sales forecasting of bigmart dataset using supervised and Ann Algorithms. *Measurement: Sensors*, *23*, 100388. https://doi.org/10.1016/j.measen.2022.100388

[9] Tkáč, M., & Verner, R. (2016). Artificial Neural Networks in business: Two decades of research. *Applied Soft Computing*, *38*, 788–804. https://doi.org/10.1016/j.asoc.2015.09.040

[10] Udom, P. (2014). A comparison study between time series model and Arima model for sales forecasting of distributor in Plastic Industry. *IOSR Journal of Engineering*, *4*(2), 32–38. https://doi.org/10.9790/3021-04213238

[11] Zhuge, Q., Xu, L., & Zhang, G. (2017). LSTM Neural Network with Emotional Analysis for prediction of stock price. *Engineering letters*, *25*(2).