UNIVERSITY OF INFORMATION TECHNOLOGY
FACULTY OF INFORMATION SYSTEMS

Application of ARIMA, SARIMA, FB PROPHET, LSTM and regression to predict the petroleum price

• ĐỖ CÔNG TRÌNH - 20522058
• NGUYỄN BẢO LÂM - 20521515
• VÕ TRỌNG HUÂN - 20520524

Application of ARIMA, SARIMA, FB PROPHET, LSTM and regression to predict the petroleum price.

1st Trinh Do Cong
Undergraduate Student of University of Information Technology - VNUHCM
Ho Chi Minh city, Vietnam

2nd Lam Nguyen Bao
Undergraduate Student of University of Information Technology - VNUHCM
Ho Chi Minh city, Vietnam

3rd Huan Vo Trong
Undergraduate Student of University of Information Technology - VNUHCM
Ho Chi Minh city, Vietnam

***Abstract*— Petroleum is a special commodity indispensable in any country, it is an important factor in promoting economic development. In fact, all economic sectors are greatly influenced by the petroleum industry. Because this is an essential item, serving daily life such as commuting, working, studying, transporting goods of people and businesses. Once the price of gasoline fluctuates, other economic sectors as well as the prices of goods also fluctuate. This study will be about forecasting the price of petroleum using linear regression, nonlinear regression, ARIMA, SARIMA, FP Prophet and LSTM model.**

Keywords—Petroleum Price forecasting, linear regression, nonlinear regression, ARIMA model, SARIMA model, FB Prophet, LSTM..

I. INTRODUCTION

Petroleum is a strategic commodity of each country, essential for social life, and has a direct impact on the economic development and national security and defense. Petroleum is also one of the main energy sources that are balanced in the energy balance policy and is one of the important commodities stored in the National Reserve. On the other hand, the oil and gas industry itself and the trading of petroleum products are also one of the key economic sectors of the country. The reality of recent development has proved that the development of this industry contributes greatly to GDP growth as well as to the industrialization and modernization of Vietnam.

In this study, a forecasting model to predict the petroleum price, specifically gasoline E95 will be chosen by comparing the following models: linear regression, nonlinear regression, Auto Regressive Integrated Moving Average (ARIMA), Seasonal Auto Regressive Integrated Moving Average (SARIMA), FB Prophet and Long Short-Term Memory.

II. RELATED WORK

The study of Ozturk and Ozturk (2018) about forecasting energy consumption in Turkey from 1970 to 2015 using ARIMA models [1]. In the paper, they apply different models for each type of energy and predict that energy consumption will continue to increase for the next 25 years.

Another ARIMA model was one of the models Godfred Kwame and Agbodah Kobima (2012) [2] used to forecast and analyze the impact of oil price fluctuations in their research. Using Hyndman Khandakar and Haslett and Raftery algorithms to select p, q and d, ARIMA (1,1,0) was the best model which was used to predict the oil price stated that the research results could serve decision-makers in responding to oil price shocks.

T.Jai et al. (2022) applied Seasonal Autoregressive Integrated Moving Average (SARIMA) in the study about predicting and forecasting analysis LPG prices. The study showed that SARIMA (0,1,1) (1,1,0) is the most suitable model to predict LPG and provide proof of future gas prices.

Lin Yao et al. (2021) used FB Prophet and LSTM to predict oil price from OPEC (Organization of the Petroleum Exporting Countries) [4]. In the paper, both models achieved good result but the Prophet model got smaller RMSE and MAE, which means the model worked better.

Anna Manowska and Anna Bluszcz (2022) suggested a new model based on artificial neural networks that includes long-term memory (LSTM) [5]. As a result, the model predicted that the demand for crude oil will increase in Poland until 2030.

III. METHODOLOGY

A. Linear Regression

Linear regression is a form of regression analysis and used to predict the value of dependent variable based on the value of independent variable. Linear regression calculates the coefficients of the linear equation, including one or more independent variables and attempts to plot a line graph between variable x and y. With x is independent variable and y is dependent variable.

There are two types of linear regression, simple linear regression with one independent and one dependent variable, multivariable linear regression with more than one independent and one dependent variable. Simple linear regression can be

used to model the relationship between age and height or food consumption, exercise to health for multivariable linear regression.
Simple linear regression is defined by the linear function:

$$Y_i = \beta 0 + \beta 1 X_i + \epsilon_i$$

Where $\beta 0$ and $\beta 1$ are unknown constants representing the regression rate, ϵ_i is the error term.
In this study, simple linear regression will be applied to predict the E95 gas price.

B. Nonlinear Regression

Nonlinear regression is a regression analysis that the regression model describes a nonlinear relationship between a dependent variable and independent variables. In other words, the relationship between the predictor and response variable follows a nonlinear pattern.
The goal of the model is to make the sum of square as small as possible. The sum of square is a measure that tracks how far the Y observations vary from the nonlinear (curved) function that is used to predict Y.
It is computed by first finding the difference between the fitted nonlinear function and every Y point of data in the set. Then, each of those differences is squared. Lastly, all of the squared figures are added together. The smaller the sum of these squared figures, the better the function fits the data points in the set. Nonlinear regression uses logarithmic functions, trigonometric functions, exponential functions, power functions, Lorenz curves, Gaussian functions, and other fitting methods.
A simple nonlinear regression formula:

$$Y = f(X, \beta) + \epsilon$$

Where X is a vector of P predictor, β is a vector of k parameters and $f(X, \beta)$ is the know regression function.

C. ARIMA

An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.

A statistical model is autoregressive if it predicts future values based on past values. For example, an ARIMA model might seek to predict a stock's future prices based on its past performance or forecast a company's earnings based on past periods.

An ARIMA model can be understood by outlining each of its components as follows:

- Autoregression (AR): refers to a model that shows a changing variable that regresses on its own lagged, or prior, values. A typical linear regression model will have this equation:

$$y = m + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_m x_m$$

  - y is predicted variable.
  - m is the intercept
  - $(x_1, x_{2,\ldots,} x_n)$ are feature or independent variable.
  - $\theta_1, \theta_2, .., \theta_n$ are the coefficient for each of the independent variable.

  Compare that with the autoregressive equation and you will see the similarities:

$$AR(p) = y_t = m + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \ldots + \theta_p y_{t-p} + \epsilon_t$$

  - p is order of AR model
  - y is current time
  - $\alpha, y_{t-1}, y_{t-2}, y_{t-p}$ are lagged version of y

- Integrated (I): represents the differencing of raw observations to allow the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).
- Moving average (MA): incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations. If ARIMA model have order p = 0, the moving average or MA(q) uses past errors to make a prediction. This is equation of MA(q) model.[6]

$$MA(q) = y_t = \beta + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \theta_p \epsilon_{t-q} + \epsilon_t$$

  - p is order of AR model
  - y is current time
  - $\beta, \epsilon_{t-1}, \epsilon_{t-2,} \epsilon_{t-q}$ are lagged version of y

The difference between Autoregressive and Moving average models is that ARIMA combines autoregressive features with those of moving averages. An AR autoregressive process, for instance, is one in which the current value is based on the immediately preceding value, while an AR process is one in which the current value is based on the previous two values. A moving average is a calculation used to analyze data points by creating a series of averages of different subsets of the full data set to smooth out the influence of outliers. As a result of this combination of techniques, ARIMA models can take into account trends, cycles, seasonality, and other non-static types of data when making forecasts.

The parameters can be defined as:

p: the number of lag observations in the model, also known as the lag order.

d: the number of times the raw observations are differenced; also known as the degree of differencing.

q: the size of the moving average window, also known as the order of the moving average.

For example, a linear regression model includes the number and type of terms. A value of zero (0), which can be used as a parameter, would mean that particular component should not be used in the model. This way, the ARIMA model can be constructed to perform the function of an ARMA model, or even simple AR, I, or MA models.

To begin building an ARIMA model for an investment, you download as much of the price data as you can. Once you've identified the trends for the data, you identify the lowest order of differencing (d) by observing the autocorrelations. If the lag-1 autocorrelation is zero or negative, the series is already differenced. You may need to difference the series more if the lag-1 is higher than zero.

Next, determine the order of regression (p) and order of moving average (q) by comparing autocorrelations and partial autocorrelations. Once you have the information you need, you can choose the model you'll use.

The ARIMA model is used as a forecasting tool to predict how something will act in the future based on past performance. It is used in technical analysis to predict an asset's future performance.

ARIMA modeling is generally inadequate for long-term forecasting, such as more than six months ahead, because it uses past data and parameters that are influenced by human thinking. For this reason, it is best used with other technical analysis tools to get a clearer picture of an asset's performance.

B. SARIMA

SARIMA stands for Seasonal-ARIMA and it includes seasonality contribution to the forecast. The importance of seasonality is quite evident and ARIMA fails to encapsulate that information implicitly.
The Autoregressive (AR), Integrated (I), and Moving Average (MA) parts of the model remain as that of ARIMA. The addition of Seasonality adds robustness to the SARIMA model. It's represented as

$$\text{ARIMA}\quad \underbrace{(p,d,q)}_{\text{Non-seasonal part of the model}} \quad \underbrace{(P,D,Q)_m}_{\text{Seasonal part of of the model}}$$

where m is the number of observations per year. We use the uppercase notation for the seasonal parts of the model, and lowercase notation for the non-seasonal parts of the model.
Similar to ARIMA, the P, D, Q values for seasonal parts of the model can be deduced from the ACF and PACF plots of the data. Let's implement SARIMA for the same Catfish sales model
Together, the notation for an SARIMA model is specified as:

SARIMA(p,d,q)(P,D,Q)m

Where the specifically chosen hyperparameters for a model are specified, for example:

SARIMA(3,1,0)(1,1,0)12

Importantly, the *m* parameter influences the *P*, *D*, and *Q* parameters. For example, an m of 12 for monthly data suggests a yearly seasonal cycle.
A *P*=1 would make use of the first seasonally offset observation in the model, e.g . t-(m*1) or t-12. A *P*=2, would use the last two seasonally offset observations t-(m * 1), t-(m * 2).
Similarly, a *D* of 1 would calculate a first order seasonal difference and a *Q*=1 would use a first order errors in the model (e.g. . moving average).

A seasonal ARIMA model uses differencing at a lag equal to the number of seasons (s) to remove additive seasonal effects. As with lag 1 differencing to remove a trend, the lag s differencing introduces a moving average term. The seasonal ARIMA model includes autoregressive and moving average terms at lag s.

C. FB prophet

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.
The Prophet algorithm in an additive regression model can handle non-linear trend and work well with strong seasonal effects. The algorithm decomposes a time series into three main components: *trend*, *seasonality*, *holiday*. The model can write as follows:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

- S_{t-1} is the trend function
- $s(t)$ represent the periodic seasonality function

where y(t) is the prediction, g(t) is the trend, s(t) is the seasonality, h(t) is the holiday component ϵ_t is a white noise term.

Accurate and fast.
Prophet is used in many applications across Facebook for producing reliable forecasts for planning and goal setting.

Fully automatic.

Get a reasonable forecast on messy data with no manual effort. Prophet is robust to outliers, missing data, and dramatic changes in time series.

Tunable forecasts.
The Prophet procedure includes many possibilities to tweak and adjust forecasts. Human-interpretable parameters can be used to improve forecasting by adding domain knowledge.

D. LSTM

Neural Networks

An artificial neural network is a layered structure of connected neurons, inspired by biological neural networks. It is not one algorithm but combinations of various algorithms which allows us to do complex operations on data.

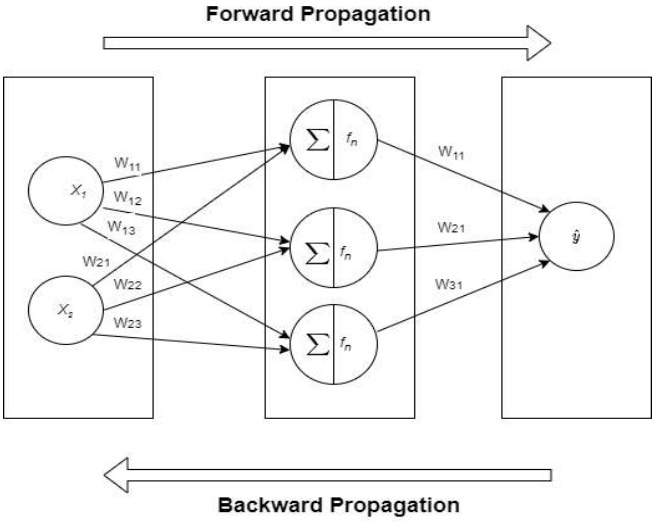

Figure1: a single-layer neural network

A network with more than one hidden is consider a deep neural network. In figure 1, there are three layers input layer, a hidden layer, and an output layer. The hidden layer represents a layer of connected neural that perform a mathematical function. Improving how neuron pass information from one layer to another by done is adding an activation function. Some common activation for hidden layer nodes includes Sigmoid, Tanh and so on…. The output layer will depend on the type of problem you are solving. For example, Linear activation for regression, sigmoid activation for binary.[6]

LSTM

An LSTM unit very similar to an RNN but with additional enhancements. An RNN only has a hidden state (*h)*, while an LSTM adds another state – the cell state (Y).

In an LSTM unit, there are four main gates (*the input, input modulation, forget,* and *output gates*) that determine how the cell state gets updated. The following diagram shows how the cell and hidden state from a previous node (C_{t-1}, h_{t-1}) gets fed with the input (X_t) to produce a new cell state and hidden state (C_t , h_t).[7]

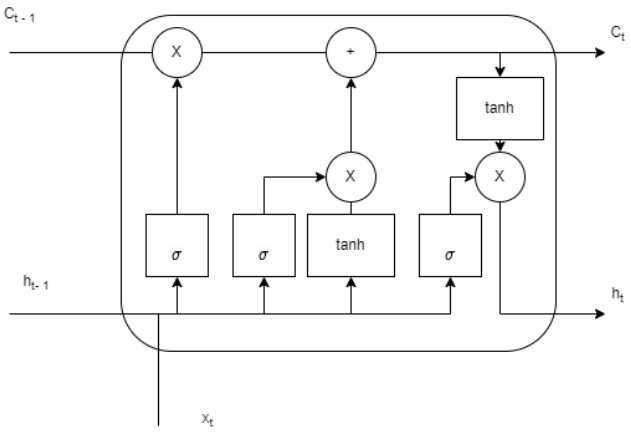

Figure 2: an LSTM cell.

Gates are mathematical computations:

- $i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi})$
- $f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf})$
- $g_t = tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg})$
- $o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho})$
- $c_t = f_t \odot c_{t-1} + i_t \odot g_t$
- $h_t = o_t \odot tanh(c_t)$

Where g_t is the input modulation gate, i_t β is the input gate, is the forget gate, $\hat{x}_t$ is the cell state, and both and represent the output gate, which determines the value of the hidden state that gets passed on. Finally, σ is the sigmoid activation function. In many works of literature, Y and β_0 are referred to as input gates (thus describing an LSTM with three gates and not four)[8]

E. ASSESSMENT METRIC

MSE:

Mean Square Error (MSE) is used to calculate the error number of statistical models. It evaluates the average squared difference between the observed and predicted values. The lower the model error, the more accurate the model. In this study, this method will be used in order to choose the best forecasting model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where y_i is the observations values, $\hat{y}_i$ is corresponding predicted values, and n is the number of observation available for analysis.[9]

MAPE:

Mean absolute percentage error (MAPE) is a metric that defines the accuracy of a forecasting method. It represents the average of the absolute percentage errors of each entry in a dataset to calculate how accurate the forecasted quantities were in comparison with the actual quantities.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Where y_i is the actual value and $\hat{y}_i$ is the forecast value.[10]

IV. DATA & RESEARCH RESULT

The data of gasoline E95 price using in this research is from Vietnam oil corporation PV OIL [7] from 2019 to 2022. Because data of Vietnam oil not suitable so that we have to scrapping from Petrolimex fanpage.

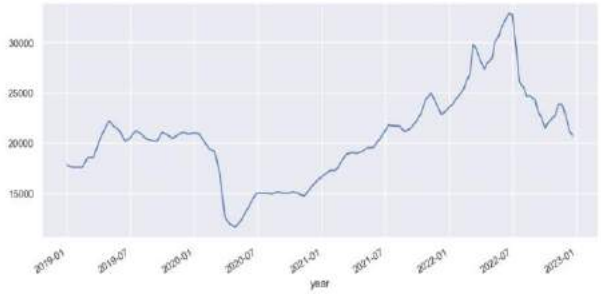

Figure 3. Visualizations of E95 price from 2019 to 2022

1) Results of MSE in 4 main models:

Model	Train: Test	MSE
ARIMA	7:3	15974147.71
	8:2	46602226.42
	9:1	3674378.9608
SARIMA	7:3	17412173.8255
	8:2	40091990.91
	9:1	3690367.35
FB PROPHET	7:3	17761234.88
	8:2	17027951.29
	9:1	13145836.91
LSTM	7:3	2268.1010
	8:2	2268.1010
	9:1	1280.2563

2) Results of MAPE in 4 main models:

Model	Train: Test	MAPE
ARIMA	7:3	0.1163
	8:2	0.1996
	9:1	0.0759
SARIMA	7:3	0.1163
	8:2	0.2418
	9:1	0.076
FB PROPHET	7:3	0.1748
	8:2	0.1629
	9:1	0.1461
LSTM	7:3	0.0895
	8:2	0.0895
	9:1	0.0573

3) Predictions of 6 model in test:

In all model, we splatted dataset follow size 8:2 and using python programming language to analysis data.

1. FB prophet:

Train: Test split 8:2

Figure 4: prediction E95 Price using FB prophet model.

2. SARIMA:

Train: Test split 8:2

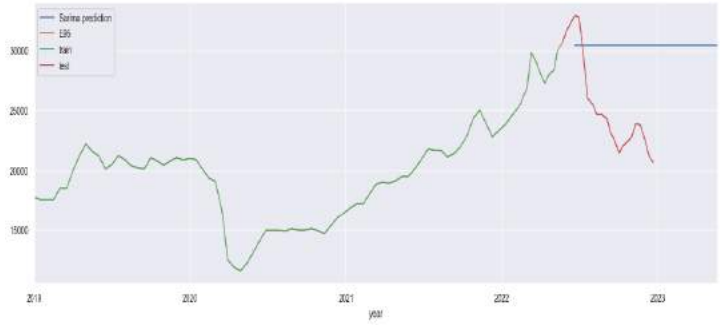

Figure 5: prediction E95 Price day using SARIMA model.

3. ARIMA:

Train: Test split 8:2

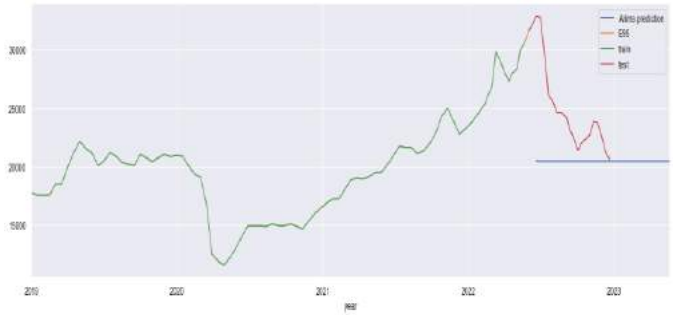

Figure 6: prediction E95 Price using ARIMA model.

4. LSTM:

Train: Test split 8:2

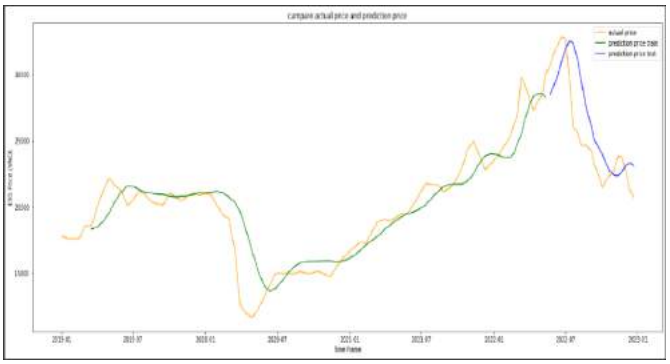

Figure 7: prediction E95 Price using SARIMA model.

5. LINEAR REGRESSION:

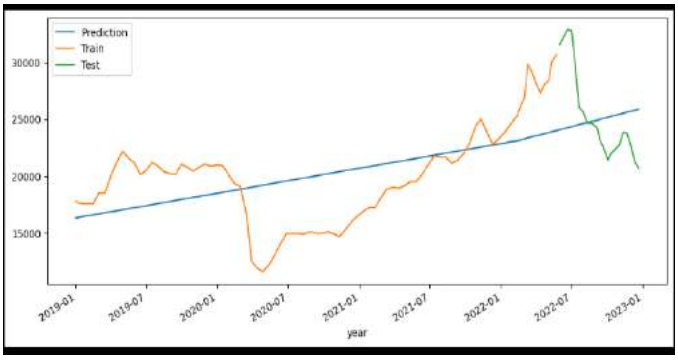

Figure 9: prediction E95 Price using Linear regression model.

6. NONLINEAR REGRESSION:

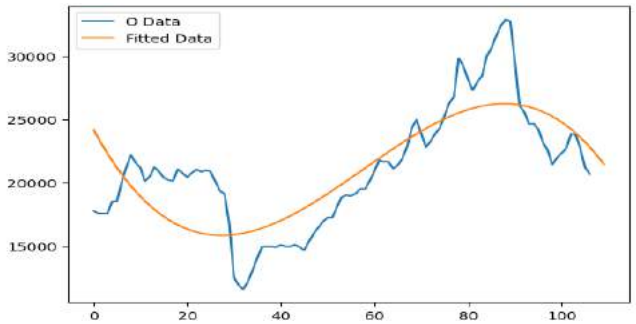

Figure 8: prediction E95 Price using Non-Linear regression model.

V. CONCLUSION

Base on the MAPE result in above. We easily to see the best model is LSTM in size 9:1. Sometimes the model still has some inadequacies that need to be solved for the model to predict better. To develop more better predictive models, we need to use some more optimal algorithms such as FFT, GRU ... to make the model more complete.

VI. REFERENCES

[1] Ozturk, S., & Ozturk, F. (2018). Forecasting energy consumption of Turkey by Arima model. *Journal of Asian Scientific Research*, *8*(2), 52-60.

[2] Abledu, G. K., & Agbodah, K. (2012). Stochastic forecasting and modeling of volatility of oil prices in Ghana using ARIMA time series model. *European Journal of Business and Management*, *4*(16), 122-131.

[3] Sankar, T. J., Mary, I. A. A., & Sameerabanu, P. (2022). Stochastic Modelling and Forecasting for LPG Prices: SARIMA Approach. *JOURNAL OF ALGEBRAIC STATISTICS*, *13*(2), 3362-3370.

[4] Yao, L., Pu, Y., & Qiu, B. (2021). Prediction of Oil Price Using LSTM and Prophet.

[5] Manowska, A., & Bluszcz, A. (2022). Forecasting Crude Oil Consumption in Poland Based on LSTM Recurrent Neural Network. *Energies*, *15*(13), 4885.

[6] Atwan, T. A. (2022). *Time Series Analysis with Python.* Packt Publising Ltd.

[7] Atwan, T. A. (2022). *Time Series Analysis with Python.* Packt Publising Ltd.

[8] Atwan, T. A. (2022). *Time Series Analysis with Python.* Packt Publising Ltd.

[9] Jim, S. B. (2022, 1 4). *Mean Squared Error* . Retrieved from Statistics By Jim: https://statisticsbyjim.com/regression/mean-squared-error-mse/

[10] To, S. H. (2022, 1 4). *mean absolute percentage error* . Retrieved from statisticshowto: https://www.statisticshowto.com/mean-absolute-percentage-error-mape/

i