2022

Analyzing impact of economic indicators on Vietnam stock market with machine learning techniques

Nuttawan Sangsawai
Faculty of Engineering

Analyzing Impact of Economic Indicators on Vietnam Stock Market with Machine Learning Techniques

Miss Nuttawan Sangsawai

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Industrial Engineering
Department of Industrial Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2022

การวิเคราะห์ผลของตัวบ่งชี้ทางเศรษฐศาสตร์ต่อตลาดหลักทรัพย์เวียดนาม ด้วยเทคนิคการเรียนรู้ของเครื่อง

น.ส.ณัฐวรรณ แสงไสว

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมอุตสาหการ ภาควิชาวิศวกรรมอุตสาหการ
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

Thesis Title	Analyzing Impact of Economic Indicators on Vietnam Stock Market with Machine Learning Techniques
By	Miss Nuttawan Sangsawai
Field of Study	Industrial Engineering
Thesis Advisor	Associate Professor Daricha Sutivong, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Engineering

..... Dean of the FACULTY OF ENGINEERING
(Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

THESIS COMMITTEE

..... Chairman
(Associate Professor NARAGAIN PHUMCHUSRI, Ph.D.)
..... Thesis Advisor
(Associate Professor Daricha Sutivong, Ph.D.)
..... Examiner
(Assistant Professor NANTACHAI KANTANANTHA, Ph.D.)
..... External Examiner
(Associate Professor Chansiri Singhtaun, D.Eng.)

ณัฐวรรณ แสงไสว : การวิเคราะห์ผลของตัวบ่งชี้ทางเศรษฐศาสตร์ต่อตลาดหลักทรัพย์เวียดนามด้วยเทคนิคการเรียนรู้ของเครื่อง. (Analyzing Impact of Economic Indicators on Vietnam Stock Market with Machine Learning Techniques) อ. ที่ปรึกษาหลัก : รศ. ดร.ดาริชา สุธีวงศ์

การศึกษานี้วิเคราะห์ตลาดหลักทรัพย์เวียดนามโดยใช้ตัวแบบทางสถิติและการเรียนรู้ของเครื่อง ชุดข้อมูลแสดงให้เห็นว่าคุณลักษณะของข้อมูลทั้งหมดมีความสัมพันธ์เชิงเส้นในทางบวกกับดัชนี VN-Index และชุดข้อมูลทั้งหมดมีขนาด หน่วย และระดับความเบ้ที่แตกต่างกัน จากการทดสอบรากหน่วย Augmented Dickey-Fuller (ADF) เพื่อดูว่าตัวแปรมีลักษณะนิ่งหรือไม่ พบว่าตัวแปรส่วนใหญ่ถูกแปลงเป็นข้อมูลลักษณะนิ่งหลังผ่านความแตกต่างครั้งแรก จึงเลือกใช้วิธี OLS เพื่อสร้างโมเดลหาความสัมพันธ์ระยะสั้นและผลการวิเคราะห์แสดงว่ามีเพียงสามตัวแปรคือ ดัชนีราคาผู้บริโภค (CPI), อัตราแลกเปลี่ยน และดัชนี S&P500 ที่มีนัยยะสำคัญทางสถิติ นอกจากนี้ยังได้ดำเนินการทดสอบ ARDL Bound Test และผลการวิเคราะห์แสดงให้เห็นว่ามีความสัมพันธ์ระยะยาวระหว่างตัวแปรที่พิจารณาและแสดงว่ามีเพียงสามตัวแปรคือ ดัชนีราคาผู้บริโภค (CPI), ผลิตภัณฑ์มวลรวมในประเทศ (GDP) และดัชนี S&P500 ที่มีนัยยะสำคัญทางสถิติ ในส่วนของตัวแบบจากการเรียนรู้ของเครื่อง โดยใช้ตัวแบบ Decision tree, Random forest และ XGBoost เพื่อประเมินความสัมพันธ์ระยะสั้นและระยะยาว ผลการวิเคราะห์แสดงให้เห็นว่าตัวแบบ Random forest มีความถูกต้องมากที่สุดในตัวแบบระยะสั้น ในขณะที่ตัวแบบ XGBoost มีความถูกต้องที่สุดในตัวแบบระยะยาว โดยตัวแปรที่มีนัยสำคัญทางสถิติทั้งสามของทั้งจากวิธี OLS และ ARDL ได้รับการจัดอันดับให้เป็นตัวแปรที่มีอิทธิพลสูงสุดสามอันดับแรกใน Random Forest และ XGBoost ตามลำดับ

สาขาวิชา วิศวกรรมอุตสาหการ ลายมือชื่อนิสิต .....

ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6470339921 : MAJOR INDUSTRIAL ENGINEERING

KEYWORD: Vietnam, Stock market, Data Analytics, Machine Learning, Statistics

Nuttawan Sangsawai : Analyzing Impact of Economic Indicators on Vietnam Stock Market with Machine Learning Techniques. Advisor: Assoc. Prof. Daricha Sutivong, Ph.D.

The study provides an analysis of the Vietnamese stock market using statistical and machine learning models. The dataset shows that all features have a positive linear relationship with the VN Index, but exhibit different scales and degrees of skewness. The Augmented Dickey-Fuller (ADF) unit root test was conducted to identify whether the variables were stationary or non-stationary, and most variables were transformed into stationary data through first differencing. The OLS method was used to construct a short run model, and the results indicated that only three variables, namely CPI, exchange rate, and S&P500 index, exhibited statistical significance. The ARDL Bound test was conducted, and the results indicated that there is a long-run relationship between the variables under consideration, and the results indicated that only three variables, namely CPI, GDP, and S&P500 index, exhibited statistical significance. The decision tree, random forest, and XGBoost models were used to study short and long run relationships. The findings suggest that the random forest model performed the best in the short run, while the XGBoost model performed the best in the long run. The three statistically significant variables of both OLS and ARDL were ranked as the top-three influential variables on Random Forest and XGBoost, respectively.

Field of Study: Industrial Engineering Student's Signature .....

Academic Year: 2022 Advisor's Signature .....

ACKNOWLEDGEMENTS

I would like to express my deep gratitude and appreciation to my advisor, Assoc. Prof. Daricha Sutivong, Ph.D., for her guidance, support, and patience throughout my research. Her expertise, insight, and encouragement were invaluable to me and helped me to develop as a researcher and a scholar.

I am also grateful to the members of my thesis committee, Assoc. Prof. Naragain Phumchusri, Ph.D., Asst. Prof. Nantachai Kantanantha, Ph.D. and Assoc. Prof. Chansiri Singhtaun, D.Eng., for their valuable feedback, constructive criticism, and insights that helped me to shape and refine my research. Their willingness to take the time to read and comment on my work was greatly appreciated.

I am indebted to my colleagues and friends, who provided me with valuable support, encouragement, and a sense of community throughout my studies. Their insights, ideas, and friendship made my academic journey more enjoyable and meaningful.

I would also like to thank my family for their unwavering support, love, and encouragement throughout my studies. Their faith in me and my abilities was a constant source of inspiration and motivation.

Finally, I want to express my gratitude to all those who contributed in any way to the completion of this thesis. Your support, whether big or small, was greatly appreciated and did not go unnoticed.

Thank you all for your help and support.

Nuttawan Sangsawai

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1

Introduction

1.1 Background of Research

With a rapid gross domestic product (GDP) growth of about 6-7% over the past ten years and being one of the few nations to experience positive economic 2.9 % growth during the COVID-19 pandemic in 2020 as shown in Figure 1, Vietnam has become one of South East Asia's most outstanding countries. Things that make Vietnam's economy expeditiously develop are its strong fundamentals:

1. Large population size; The number of Vietnam's population is 98.51 million at the end of 2021, and the age of 69% of the population is between 15 and 64 and only 10% of the population is elderly people, which means Vietnam has enough labor and power to enhance the economy.
2. Vietnam is a unitary socialist state under the Communist Party of Vietnam (CPV), which means its policies and fundamental investments have continuity.
3. High export values; Export is a star of Vietnam since it is a production base of the world industries, besides the government has trade agreements with many other countries such as the Comprehensive and Progressive Agreement for Trans-Pacific Partnership (CPTPP), the European Union–Vietnam Free Trade Agreement (EVFTA) which make exporters do not have to pay import duty when exporting their goods to these countries.
4. Increasing of purchasing power; Due to becoming a production base of the world industries, employment increased, which means Vietnamese people have higher incomes. Although Vietnam was rated as a lower-middle-income country in 2020, the Vietnamese government forecasts that by 2035, the nation would be upper-middle-income, with a US$1 trillion GDP.

5. Incoming foreign cash flows: Foreign Direct Investment (FDI) of Vietnam continually increased and hit a new all-time high over 10 years at US$ 16.12 billion in 2019 as shown in Figure 2.

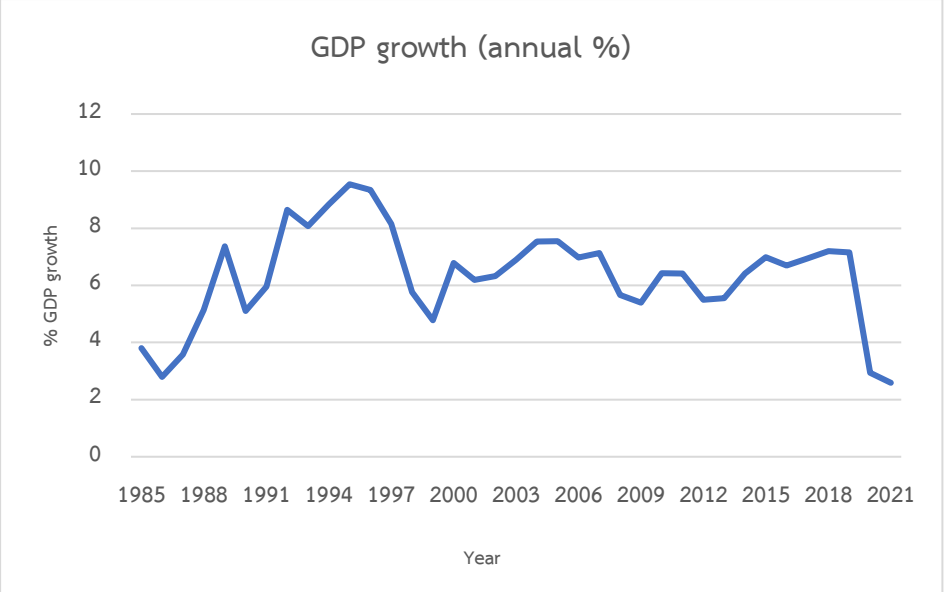


Figure 1 Vietnam GDP growth (annual %)

(Source: www.data.worldbank.org)

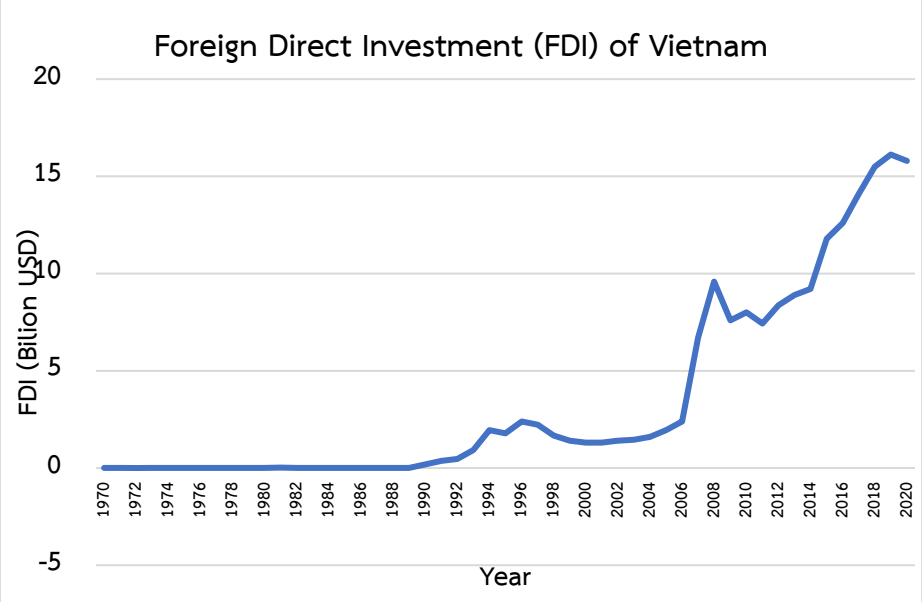


Figure 2 Foreign Direct Investment (FDI) of Vietnam from 1970 to 2020

(Source: www.data.worldbank.org)

Due to its strong fundamentals, foreign cash flows are continuously going to Vietnam in both industry segments, about 50% of GDP, and financial segments e.g., the stock market.

1. Vietnamese stock market

Vietnamese stock market includes Ho Chi Minh Stock Exchange (HOSE), and Hanoi Stock Exchange (HNX).

1.1 Ho Chi Minh Stock Exchange (HOSE)

Ho Chi Minh Stock Exchange (HOSE), originally the HCM Securities Trading Center, was founded in 1988 in Ho Chi Minh City, Vietnam. It had its first trading session on July 28, 2000, the day after it had its official opening on July 20, 2000. Ho Chi Minh Stock Exchange is the first and the largest stock exchange in Vietnam, and its main index is the VN Index (VNI) which indexed all registered stocks in this stock exchange.

1.2 Hanoi Stock Exchange (HNX)

Hanoi Stock Exchange (HNX), originally the Hanoi Securities Trading Center (Hanoi STC), was founded in 2005 in Hanoi, Vietnam. Hanoi Stock Exchange is the second stock exchange in Vietnam after Ho Chi Minh Stock Exchange (HOSE), and its main index is the HNX 30 Index (HNX30) which indexed the largest 30 market capitalization registered stocks in this stock exchange and HNX 30 Total Return Index (HNX30TRI) which indexed the return of HNX30.

The Vietnamese stock market has had continuous growth over 10 years and is now one of the most attractive markets in the world, with the VN-INDEX stock index growing by 36% in 2021 as shown in Figure 3. Despite its remarkable expansion since its founding, Vietnamese stock market is still regarded as a frontier market with a small market size (Nguyen et al., 2019), so studies about this market are not popular.

Any studies about stock markets, so studying in a small market and still rated as a frontier market but has strong fundamentals for growing as Vietnam is compelling.

However, as with any investment, there are also risks to consider. The Vietnamese economy is still developing, and there may be volatility and uncertainty in the market. Political and regulatory risks, such as changes in government policy or restrictions on foreign investment, could also impact the market. So, Understanding the fundamentals of each stock market is crucial as each has its own unique traits and driving factors.

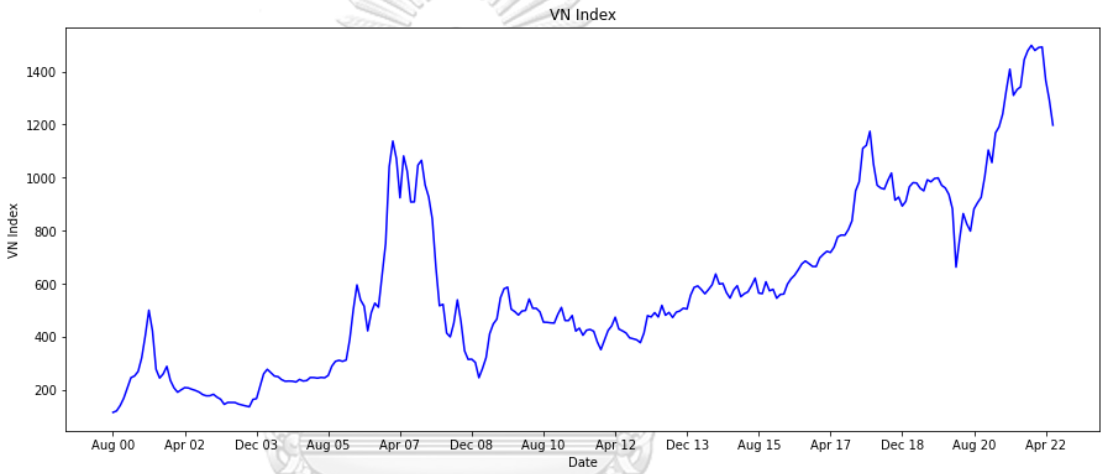


Figure 3 VN-Index from established to June 2022
(Source: www.investing.com)

By knowing how certain factors impact on stock market, investors and financial field people can understand the fundamentals of each stock market which be useful for investment planning. Macroeconomic indicators have a considerable impact on stock prices, according to (Fama, 1981), who analyzed the relationship between stock prices and macroeconomic indicators in the US. Numerous academics have attempted to examine the same topic using different macroeconomic variables and other countries, but the majority of studies used different statistical approaches and were conducted in developed countries (Hussainey & Khanh Ngoc, 2009).

The capacity of computers and technology has increased recently to deal with more complex issues facing humanity. The use of machine learning algorithms is expanding. Numerous studies demonstrate that when it comes to stock price forecasting, machine learning algorithms beat more conventional statistical methods (Bhattacharjee & Bhattacharja, 2019). Additionally, it has been claimed that other influences, such as commodity prices and other countries' economies, affect stock markets.

1.2 Problem Statement

Currently, studies in stock markets have been majority in developed countries and analyzing a relationship or an impact between some factors and stock markets has been done by statistical methods but the results are still uncertain and contrast among those studies. As a result, the study's primary goal is to investigate the variables, such as local and global factors, that affect the Vietnamese stock market. Different machine learning methodologies are used to construct models. To clarify how these indicators affect the stock market, they will also be examined according to their feature importance. Finally, the resulting economic indicators will be compared between traditional statistical techniques and machine learning techniques.

1.3 Research Objectives

1. To compare the impacts of economic indicators on the Vietnamese stock market in short and long run.
2. To compare between traditional statistical techniques and machine learning techniques.

1.4 Scopes of Research

1. The index used for representing Vietnamese stock market is the VN Index (VNI).
2. The data used in this study is monthly data for the period August 2000 to June 2022, published on Investing and Tradingeconomics website.
3. The dependent variable is VN Index (VNI) and the independent variables are Exports, Imports, CPI (Consumer price index), Unemployment rate, GDP, Exchange rate, Brent oil price and S&P500 index.
4. Machine learning techniques are applied in this study including Decision Tree, Random Forest, and XGBoost.
5. Performance metric to indicate model performance in this study is the root mean squared error (RMSE).

1.5 Expected Outcomes

1. Models for explaining how factors impact Vietnam stock market.
2. Factors that contribute and are important to Vietnam stock market.

1.6 Expected Benefits

1. Achieving the data insights or driven factors for the Vietnamese stock market, which can be used to explain the market fundamentals in order to preliminary predict the market and plan investment.
2. Understand the comparison between analyzing economic indicators using the traditional statistical techniques and the machine learning techniques.

1.7 Timeline

The timeline of this research is shown in Table 1.

Table 1 Timeline of this study

No.	Task	2022																				2023															
		AUG				SEP				OCT				NOV				DEC				JAN				FEB				MAR				APR				MAY	
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2
1	Interested topics finding																																						
2	Data exploration and collection																																						
3	Identify problems and scope of the thesis																																						
4	Literature review																																						
5	Data preparation																																						
6	Explanatory data analysis (EDA)																																						
7	Proposal Preparation																																						
8	Proposal Exam																																						
9	Machine learning model construction																																						
10	Machine learning model interpretation																																						
11	Statistical model construction																																						
12	Conclusion																																						
13	Conference preparation																																						
14	Defense preparation																																						
15	Conference participation																																						
16	Defense preparation																																						
17	Defense Exam																																						

Chapter 2

Literature review

2.1 Related Theory

1. Regression analysis

1.1) Linear regression

For regression problems in any field of study, one of the most widely used techniques which easy to create and interpret is linear regression.

Linear regression is a statistical technique that employs at least two independent variables to predict the value of the dependent variables. A simple equation can be used to illustrate the relationship, as shown in (1).

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad \text{.....(1)}$$

where,

Y_i is dependent variables

$x_{i1}, x_{i2}, \ldots, x_{ip}$ are independent variables

β_0 is intercept

$\beta_1, \beta_2, \ldots, \beta_p$ are regression coefficients

ε_i is prediction errors

The regression model and all regression coefficients are examined by minimization of the sum of squares of the errors, which is also known as the ordinary least squares (OLS) method.

The assumptions underlying the multiple linear regression model are as follows:

1) Linearity: The relationship between the dependent variable and the independent variables is linear.
2) Independence: The observations in the dataset are independent of each other. There should be no systematic patterns or correlations among the residuals or errors.
3) Homoscedasticity: The variance of the residuals or errors is constant across all levels of the independent variables.
4) Normality: The residuals follow a normal distribution.
5) No multicollinearity: The independent variables should not be highly correlated with each other.

In the context of time series analysis, linear regression can be used to model the relationship between a dependent variable and one or more independent variables that vary over time. However, applying linear regression directly to time series data can be problematic because the data may violate the assumptions of linear regression, such as independence, homoscedasticity, and normality. Time series data may also exhibit autocorrelation, where the values of the variable at different time points are correlated with each other and may lead to spurious regression, a high coefficient of determination (R^2) but low value for the Durbin-Watson statistic regression (Granger & Newbold, 1974).

1.1.1) Spurious regression

Spurious regression occurs when two or more time series variables are not cointegrated, but they exhibit a strong correlation over time. In other words, the relationship between the variables is purely coincidental and does not reflect a true economic or statistical relationship.

The consequences of spurious regression are as follows:

1) Regression coefficient estimates are ineffective.
2) Regression-based forecasts are not as accurate as they could be.
3) The regular significance tests are disabled on the coefficients.

Spurious regression can lead to misleading or erroneous conclusions and may result in incorrect policy decisions or investment strategies.

One way to fix spurious regression is to test for cointegration between the variables using appropriate methods such as the Engle-Granger test or the Johansen test. If the variables are not cointegrated, it suggests that they do not share a long-term relationship, and any short-term correlation between them is likely to be spurious.

Overall, spurious regression is a common problem in time series analysis, and it is important to use appropriate methods and techniques to avoid it and obtain reliable and meaningful results.

1.1.2) Cointegration

Cointegration is a statistical concept that measures the long-term relationship between two or more non-stationary time series variables that are not directly causally related. Two or more non-stationary time series variables are said to be cointegrated if they share a common trend or move together in the long run called "Random walk together". This means that there is a long-run equilibrium relationship between the variables, which implies that they are cointegrated.

Cointegration is a property of non-stationary time series that allows them to be used in regression analysis even though they are

not stationary. It is important because it allows us to model the relationship between non-stationary variables in a way that avoids spurious or misleading regression results. For example, if two non-stationary variables are regressed against each other, they may show a strong relationship even if there is no real underlying connection between them.

The presence of cointegration can be tested using statistical tests such as the Johansen test or the Engle-Granger test. If the null hypothesis of no cointegration is rejected, it suggests that the variables are cointegrated.

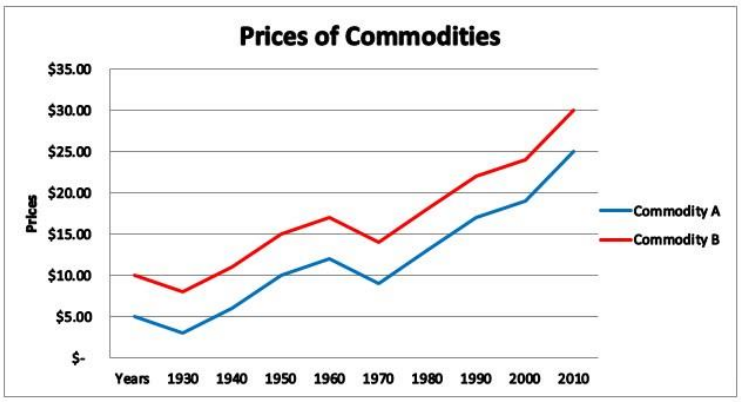


Figure 4 Example the cointegration of two non-stationary time series (Source: https://www.wallstreetmojo.com/cointegration/)

1.2) Statical regression methods for time series data

(Shrestha & Bhatta, 2018) aggregated all statistical regression techniques for time series data as shown in Figure 5.

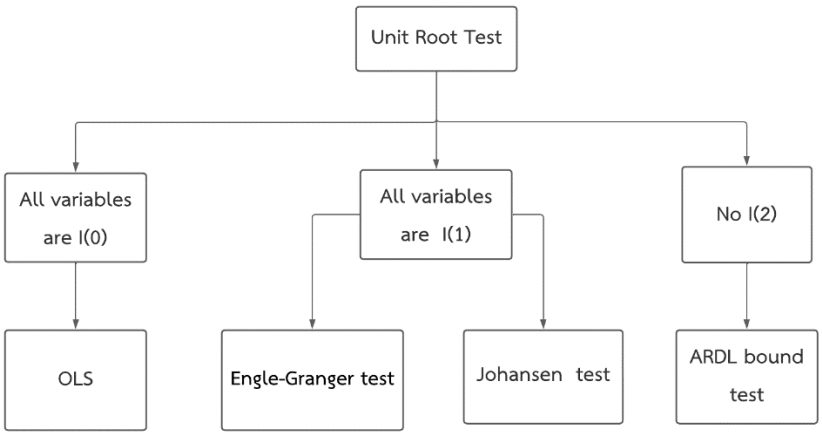


Figure 5 Method selection for time series data

1.2.1) Unit root test method

A unit root test is a statistical test used to determine whether a time series variable has a unit root. A unit root is a type of non-stationarity in a time series, where the mean and variance of the series increase over time.

A unit root test is important in time series analysis because it can affect the results of statistical analysis. If a time series has a unit root, it can lead to spurious regression results, where a strong relationship between two variables is observed, even though they are not actually related.

There are several types of unit root tests, including the Augmented Dickey-Fuller (ADF) test, the Phillips-Perron (PP) test, and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. These tests differ in their assumptions and the null hypothesis they test.

The Augmented Dickey-Fuller (ADF) test is perhaps the most commonly used method, the test is as follows:

$H_0 : \delta = 0$

$H_1 : \delta < 0$

$$\Delta y_t = \mu + \delta y_{t-1} + \sum_{i=1}^{k} \beta_i \Delta y_{t-i} + e_t \quad \text{.....(2)}$$

The null hypothesis can be rejected if ADF value is lower than the critical values, which means the variable is stationary.

The results of the unit root test, which identify the stationarity of the variable, and their order of differencing, which identify the minimum number of differencing required for that variable to be stationary, are used to select the method for time series analysis.

1.2.2) Order of integration

The order of integration is a concept in time series analysis that measures the degree to which a time series variable is non-stationary. A time series is said to be integrated of order d, denoted as I(d), if it can be made stationary by differencing it d times.

The order of integration can be determined through visual inspection of the time series plot or through statistical tests such as the Augmented Dickey-Fuller (ADF) test, which tests for the presence of a unit root in the time series. If the null hypothesis of the ADF test is rejected, it suggests that the time series is integrated and the order of integration can be determined by examining the number of times the series needs to be differenced to become stationary.

1.2.3) Ordinary least squares (OLS)

Ordinary least squares (OLS) is a method commonly used in linear regression analysis to estimate the parameters of a linear equation that describes the relationship between a dependent variable and one or more independent variables. The goal of OLS is to find the values of the parameters that minimize the sum of the squared errors between the predicted values and the actual values of the dependent variable. In other words, OLS finds the "best-fit" line that describes the relationship between the independent and dependent variables.

The OLS method assumes that the errors or residuals (the differences between the predicted values and actual values) are normally distributed, have a constant variance (homoscedasticity), and are uncorrelated with the independent variables (independence assumption). Violations of these assumptions can affect the accuracy of the OLS estimates and lead to biased or inefficient results.

A non-stationary time series can be made stationary by differencing; however, it may appear to be simpler to evaluate the relationship to take the difference of a non-stationary time series and use the OLS method after making all the variables stationary. However, the difference only accurately captures the short-term change in the time series and completely ignores the long-term data. It highlights the immediate effects or fluctuations in the data without considering long-term trends or seasonal variations. The analysis of non-stationary variables is therefore not advised to use this method (Shrestha & Bhatta, 2018).

To see why this happens consider a simple model. Suppose that the long-run relationship between Y and X is

$$Y_t = BX_t + u_t \quad \text{.....(3)}$$

The differenced equation is

$$\Delta Y_t = B\Delta X_t + v_t \quad \text{.....(4)}$$

where v=Δu

Since $E[v] = 0$,

$$\Delta Y_t = B\Delta X_t \quad \text{.....(5)}$$

If we use equation 5, we will predict that future changes in Y will be B times the changes in X.

1.2.4) Cointegration test

The results of cointegration tests reveal situations in which multiple non-stationary time series are merged in a manner that prevents them from deviating from equilibrium over an extended period. These tests aid in assessing the level of sensitivity of two variables to the same average price during a specific time frame.

There are several methods for testing cointegration, but two commonly used methods in case of all variables are I(1) are the Engle-Granger test and the Johansen test.

1). Engle-Granger test

The Engle-Granger test is a two-step procedure for testing cointegration between two non-stationary time series variables. In the first step, the variables are regressed against each other to obtain the residuals. In the second step, the

residuals are tested for stationarity using a unit root test such as the Augmented Dickey-Fuller (ADF) test. If the residuals are stationary, it suggests that the variables are cointegrated (Engle & Granger, 1987).

2). Johansen test

The Johansen test is a more general method for testing cointegration between two or more non-stationary time series variables. It is a multivariate test that estimates the number of cointegrating relationships and their coefficients. The Johansen test involves estimating a Vector Error Correction Model (VECM) model using Maximum Likelihood Estimation (MLE) and testing the eigenvalues of the resulting matrix. If the eigenvalues are significant, it suggests the presence of cointegration (Johansen, 1988) (Johansen & Juselius, 1990).

Both methods have their advantages and limitations. The Engle-Granger test is a simple and computationally efficient method, but it assumes that there is only one cointegrating relationship between the variables. The Johansen test is a more flexible and powerful method that can handle multiple cointegrating relationships, but it can be computationally intensive and may require more data to obtain reliable results.

1.2.5) Autoregressive distributed lag (ARDL) approach

1). Autoregressive distributed lag (ARDL) model

Autoregressive distributed lag (ARDL) model is an ordinary least square (OLS) based models that may be used for both non-stationary and mixed order of integration time series. The model is a type of regression model that includes lags of the dependent variable, lags of

the independent variables, and possibly differenced variables. The order of the ARDL model is usually denoted as (p, q), where p is the number of lags of the dependent variable and q is the number of lags of the independent variable(s).

Determining the appropriate order of an ARDL model is an important step in building the model. One approach is to use statistical tests, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), to select the optimal order. These tests compare the goodness-of-fit of different ARDL models with varying orders, and select the model with the lowest AIC or BIC value as the best fit.The general form of the ARDL model is:

$$Y_t(p,q) = \alpha + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=0}^{q} \delta_i X_{t-i} + \varepsilon_t \quad \text{.....(6)}$$

where:

Y_t is the dependent variable at time t

α is the constant term

Y_{t-i} are lags of the dependent variable, with i = 1,2,...,p

X_{t-i} are lags of the independent variable(s), with i = 0,1,2,...,q

ε_t is the error term

p is the number of lags of the dependent variable

q is the number of lags of the independent variable(s).

2). ARDL bound test

ARDL bound test, also known as the bounds testing approach, is a statistical test used to determine whether there is a long-run

relationship between two or more variables in an ARDL model. The test is based on the concept of cointegration, which suggests that if two or more variables are non-stationary, but a linear combination of them is stationary, then there is a long-run relationship between the variables (Pesaran & Shin, 1999) (Pesaran et al., 2001).

The ARDL bound test is a useful tool in econometric analysis because it allows for the estimation of long-run relationships between variables without requiring them to be integrated of the same order. This means that the test can be applied to a wide range of economic variables with different levels of integration. Specifically, the test is designed for cases where the dependent variable may be integrated of order zero (I(0)), order one (I(1)), or some combination of the two (I(0)/I(1)).

However, the ARDL bound test is not commonly used for variables that are integrated of order two (I(2)) or higher, because the test is not designed to handle such cases. When the variables are integrated of higher orders.

The ARDL bound test can indicate cointegration relationship as follows:

The error correction version of the ARDL model is given by:

$$\Delta y_t = \beta_0 + \sum_{i=1}^{p} \beta_i \Delta Y_{t-i} + \sum_{i=0}^{q} \delta_i \Delta X_{1,t-i} + \sum_{i=0}^{q2} \varepsilon_i \Delta X_{2,t-i} + \varphi z_{t-1} + e_t \quad \text{.....(7)}$$

Cointegration / Long-run equation is given by:

$$y_t = \alpha_0 + \alpha_1 X_{1,t} + \alpha_2 X_{2,t} + \varepsilon_t \quad \text{.....(8)}$$

Thus, the OLS residuals series from the long-run "cointegrating regression" serves as the "error-correction term" (z).

$$Z_{t-1} = \varepsilon_{t-1} \quad \text{.....(9)}$$

Equation (7) can be transformed to Equation (10) as follows:

$$\Delta y_t = \beta_0 + \sum_{i=1}^{p} \beta_i \Delta Y_{t-i} + \sum_{i=0}^{q} \delta_i \Delta X_{1,t-i} + \sum_{i=0}^{q2} \varepsilon_i \Delta X_{2,t-i} + \lambda_0 y_{t-1} + \lambda_1 x_{1,t-1} + \lambda_2 x_{2,t-1} + e_t \quad \text{.....(10)}$$

where:

Y_t is the dependent variable at time t

X_t are the independent variable(s) at time t

α_0 , β_0 are the constant term(s)

Y_{t-i} are lags of the dependent variable, with i=1,2,...,p

X_{t-i} are lags of the independent variable(s), with i=1,2,...,q

$\boldsymbol{\varepsilon}_t$ is the error term

$\beta_i, \delta_i, \varepsilon_i$ are short run dynamics of the model

λ_i are s long run relationship

ARDL Bound test is as follows:

H_0 : $\lambda_0 = \lambda_1 = \lambda_2 = 0$ (There is no long-run relationship between)

H_1 : At least one $\lambda_i \neq 0$ (There is a long-run relationship between the variables)

If the F-test value is greater than the critical value, then the null hypothesis of no long-run relationship is rejected, and there is evidence of a long-run relationship between the variables.

If the F-test value is less than the critical value, then the null hypothesis is not rejected, and there is no evidence of a long-run relationship between the variables.

2. Pearson correlation coefficient

The Pearson correlation coefficient, often known as Pearson's r, is used to quantify the linear relationship between two sets of data. Its values might range from -1 to 1. A direct and completely positive or negative association is indicated by values of 1 and -1, respectively. When the correlation coefficient is 0, there is no linear relationship.

$$r_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad \text{.....(11)}$$

where,

cov(X,Y) is the covariance of variables X and Y

σ_X is the standard deviation of X

σ_Y is the standard deviation of Y

Interpretation of Pearson correlation coefficient are as follows:

$0.8 \leq r \leq 1$ means very high correlation

$0.6 \leq r \leq 0.79$ means high correlation

$0.4 \leq r \leq 0.59$ means moderate correlation

$0.2 \leq r \leq 0.39$ means low correlation

$0 \leq r \leq 0.19$ means very low correlation

3. Machine learning

Machine learning is an application of artificial intelligence (AI) that aims to replicate how the human brain functions. The type of machine learning is shown in Figure 6.

Based on their algorithms, machine learning is divided into three types:

1. Supervised Machine Learning

Supervised machine learning models are taught using datasets that consist of inputs without labels and corresponding outputs with labels. These models then learn to associate the inputs with the correct outputs. The algorithm identifies patterns within the data, learns from observations, and generates predictions. For instance, an algorithm could be trained using images of dogs and various other objects that have been previously labeled by humans. Subsequently, the algorithm learns to independently recognize images of dogs. Supervised machine learning is frequently employed in two categories of problems.

- Classification problem: to categorize the discrete output variable, e.g., spam filtering for emails.

- Regression problem: to understand the relationships among variables when output variable is a continuous value.

2. Unsupervised Machine Learning

Unsupervised machine learning models are trained exclusively using unlabeled data to detect patterns or trends that might not be consciously sought by individuals. For example, through the analysis of online sales data, an unsupervised machine learning algorithm can identify different customer groups based on their purchasing behavior. Unsupervised machine learning can be used for two types of problems.

- Clustering: To divide data into several groups based on their similarity. For example, clothing companies may collect customers' weight, height, and body size data to identify how many sizes of their products should be produced.
- Association: To find the relationships between variables in the dataset by determining the co-occurrence situation, such as customer recommendations, people who buy diapers also tend to buy powdered milk.

3. Reinforcement Machine Learning

Reinforcement machine learning teaches the machine through trial and error to explore several options and select the best path to maximize the reward in a situation.

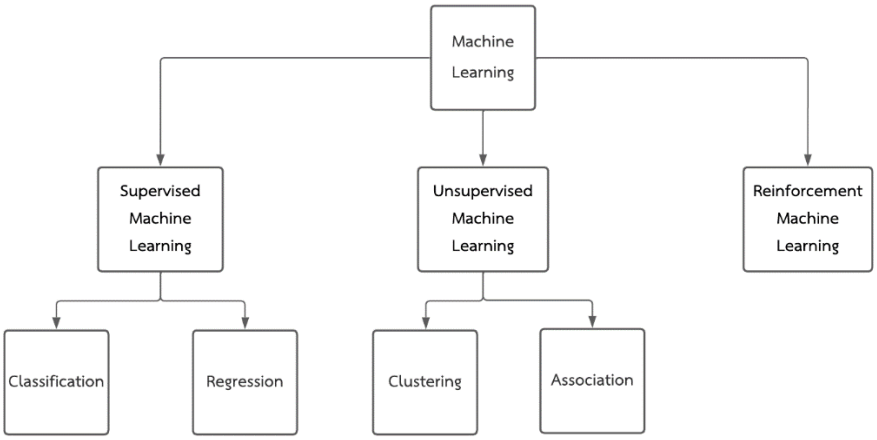


Figure 6 Type of machine learning

3.1) Decision Tree

A decision tree (also known as a CART) is a supervised learning algorithm that can be used for both classification and regression. It creates a tree-like structure by dividing the data into branches that each represent a possible decision or outcome and that demonstrate how one decision leads to the next while also demonstrating that each option is mutually exclusive. The tree's root node, which denotes the start of the process, decision nodes, which denote the division of the data, and leaf nodes, which denote the outcomes, can all be understood. This splitting process begins at the root node and continues until a leaf node is reached. It is impossible to further separate the leaf node. Figure 7 demonstrates an example of decision tree.

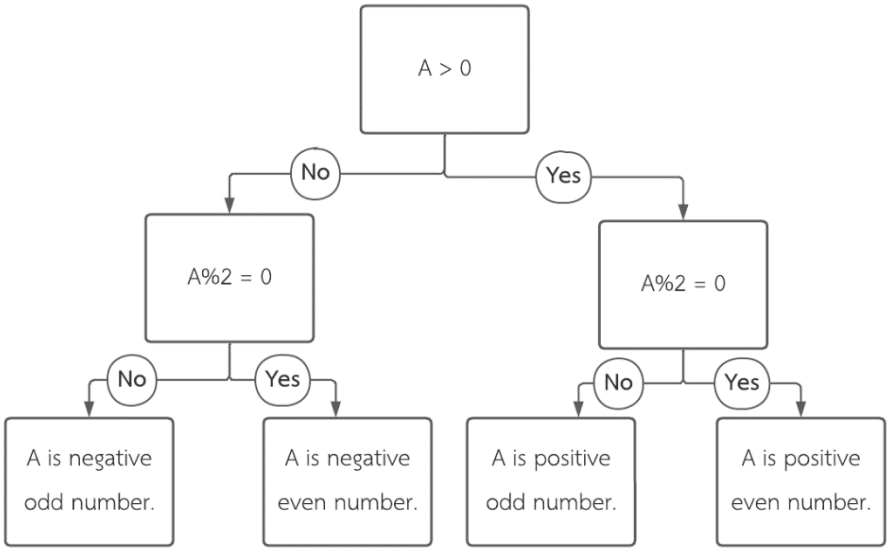


Figure 7 Example of decision tree

3.1.1) Decision Tree for classification

Within a decision tree, the algorithm commences at the root node and progresses towards predicting the class of the provided dataset. This algorithm proceeds by traversing branches and moving to the next node by comparing the values of the root attribute with the attribute of the record in the dataset. At each subsequent node, the algorithm once again examines the attribute value with respect to the remaining sub-nodes before proceeding further. This process is repeated until the algorithm arrives at a leaf node.

The primary challenge in constructing a decision tree is determining the optimal attribute selection for the root node and subsequent sub-nodes. To address this challenge, a technique called attribute selection measure (ASM) is employed. ASM provides a solution for choosing the most suitable attribute for the tree nodes in a straightforward manner. There are two commonly utilized ASM approaches, outlined below:

1). Information Gain

Following the segmentation of a dataset based on an attribute, information gain is the measurement of changes in entropy before and after the split.

By measuring the size of uncertainty, disorder, or impurity in general, Information Gain is used to determine which feature provides the most information about the classification based on the concept of entropy, with the goal of reducing the amount of entropy starting from the top (root node) to bottom (leaves nodes).

$$Information\ Gain\ =\ Entropy(before)\ -\sum_{j=1}^{K} Entropy(j, after) \quad \text{.....(12)}$$

In this case, "before" refers to the dataset before the split, "after" refers to the subset j created after the split, and K is the number of subsets produced by the split.

In a decision tree algorithm, the priority for splitting a node or attribute is based on the one that yields the highest information gain. This approach ensures that the algorithm consistently aims to maximize the value of information gain.

1.1). Entropy

Entropy is a term used to measure the amount of unpredictability or impurity in the data points. A high degree of disorder indicates a low degree of impurity. Entropy is measured between 0 and 1, though it may exceed 1 depending on the number of groups or

classes present in the data set. Either way, it denotes a higher level of disorder.

$$Entropy(p) = -\sum_{i=1}^{N} p_i log_2 p_i \quad \text{.....(13)}$$

Where “p” refers the probability that it is a function of entropy.

2). Gini Index

The Gini Index, commonly referred to as Gini impurity, determines the probability that a certain feature would be inaccurately classified when chosen at random. It can be said to as pure if every element is connected to a single class. Consider the Gini Index as a measurement standard. Similar to the entropy qualities, the Gini Index ranges from 0 to 1. An attribute with a lower Gini index is more desirable than one with a higher Gini value.

$$Gini\ Index = 1 - \sum_{i=1}^{n} {p_i}^2 \quad \text{.....(14)}$$

3.1.2) Decision Tree for regression

Decision Tree for regression is also known as regression tree. A regression tree, which is used to predict continuous valued outputs rather than discrete outputs, is essentially a decision tree that is employed for the regression problem.

Gini Impurity is used by CART in classification cases to divide the dataset into decision trees. Contrarily, CART employs least squares in regression scenarios, where splits are logically chosen to reduce the

residual sum of squares between the observation and the mean in each node. In mathematics, residual can be expressed as follows.

In mathematics, RSS (residual sum of squares) can be expressed as follows:

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad \text{.....(15)}$$

where,

n is the number of observations

Y_i is actual values

$\hat{Y}_i$ is predicted values

3.2) Random Forest

A supervised learning approach called random forest is created by individually creating each decision tree and then combining the trees using the bagging method or bootstrap aggregation. The random forest uses its majority vote for classification and average in the case of regression at the end of the procedure rather than relying just on one decision tree as shown in Figure 8. The larger number of trees in the forest inhibits higher accuracy and overfitting.

3.2.1) Bagging

Random forest utilizes the ensemble method known as bagging, also referred to as Bootstrap Aggregation, to construct multiple decision trees simultaneously. Bagging involves selecting random samples from the dataset, where each tree is built using these samples known as Bootstrap Samples. The sampling process

involves selecting instances with replacement, a technique called row sampling. The term "Bootstrap" pertains to this phase of sampling with replacement. Each decision tree is trained independently, generating individual results. Subsequently, the outputs of all the models are combined, and the final decision is made based on a majority vote in classification problems or an average value in regression problems.

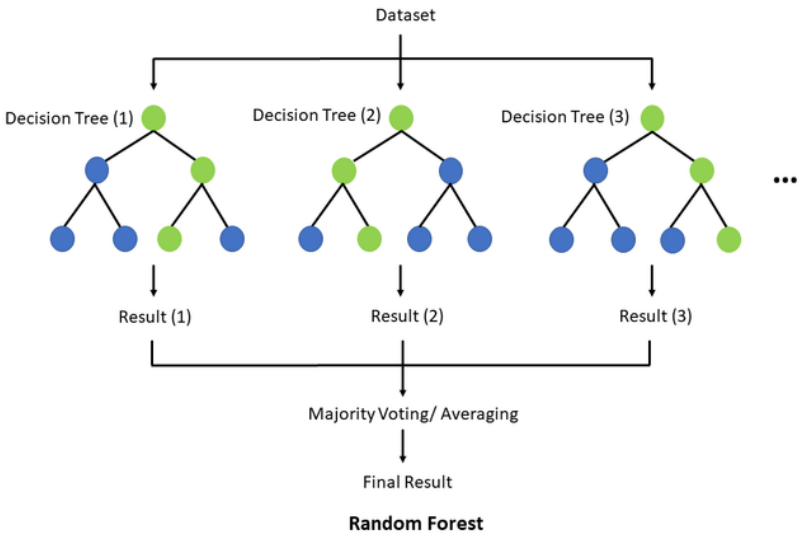


Figure 8 Random forest

(Source: https://en.m.wikipedia.org/wiki/File:Random_forest_explain.png)

3.3) XGBoost

Gradient Boosting Decision Trees (GBDTs), a decision tree ensemble learning approach that is comparable to the random forest, are implemented by XGBoost. The XGBoost technique uses gradient boosting to iteratively combine many decision trees learning models into a stronger learner. The objective loss function is optimized by using the residual at each iteration to increase the previous predictor's prediction accuracy. To avoid overfitting, a regularization term is also included in the objective function.

Gradient boosting refers to the process of enhancing a weak model by combining it with multiple other weak models, resulting in a stronger overall model. This technique, known as boosting, involves iteratively creating weak models and adding them together using a gradient descent method over an objective function. The aim is to minimize errors by establishing desired outcomes for the subsequent models based on the gradient of the error with respect to the predictions. This approach is referred to as gradient boosting.

In the case of gradient boosting decision trees (GBDTs), a set of shallow decision trees is trained iteratively. Each iteration uses the error residuals of the previous model to fit the new model. The final prediction is obtained by taking the weighted average of the predictions from all the trees. This process of boosting helps reduce bias and underfitting in the model. In contrast, random forest employs a "bagging" approach to minimize variance and overfitting.

XGBoost is a scalable and highly accurate gradient boosting algorithm designed to enhance the performance and computational speed of machine learning models. It surpasses the computational capabilities of traditional boosted tree algorithms. XGBoost constructs trees in parallel, in contrast to the sequential method used in GBDT. It employs a level-wise approach, evaluating the quality of potential splits in the training set based on the partial sums of gradient values.

4. K-fold cross validation

K-fold cross-validation is a commonly used technique for assessing the performance of a machine learning model. It involves dividing the dataset into k equal-sized subsets, or "folds," and then training and evaluating the model k times, each time using a different fold as the validation set and the remaining k-1 folds as the training set as shown in Figure 9.

In each iteration, the model is trained on the training set and evaluated on the validation set. The performance metric (such as accuracy, precision, recall, or F1 score) is then computed for each iteration, and the average metric value across all k iterations is used as an estimate of the model's generalization performance.

K-fold cross-validation is particularly useful when the dataset is small or when the model has many parameters that need to be tuned. By using multiple splits of the data, it provides a more robust estimate of the model's performance than a single train-test split, and helps to reduce overfitting.

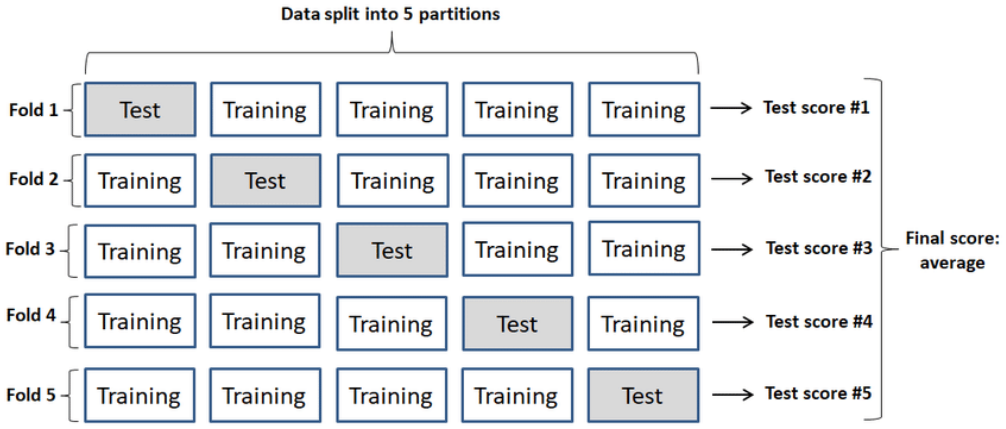


Figure 9 5-fold cross validation

(Source: (Phung & Rhee, 2019))

5. Performance Metrics

5.1) Coefficient of determination or R-Squared (R^2)

The coefficient of determination (R^2) is an important statistical indicator of regression models that examines how well a model predicts or explains the outcomes. The value increases with the model's credibility.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2} \quad \text{.....(16)}$$

where,

n is the number of observations

$\mathrm{Y_i}$ is actual values

$\widehat{\mathrm{Y}}_\mathrm{i}$ is predicted values

$\overline{\mathrm{Y_i}}$ is an average of the actual values.

There are several weaknesses or limitations associated with the R-squared (R^2) statistic, which measures the proportion of the variance in the dependent variable that is explained by the independent variables in a linear regression model. Some of these weaknesses include:

1. R-squared does not indicate the causal relationship between the independent and dependent variables. It only measures the goodness of fit of the model.
2. R-squared does not tell us how well the model predicts new observations. In other words, it does not indicate the predictive accuracy of the model.

3. R-squared is sensitive to the number of independent variables in the model. Adding more independent variables to the model will typically increase R-squared, even if the additional variables are not actually related to the dependent variable.
4. R-squared can be misleading in the presence of multicollinearity, where two or more independent variables are highly correlated with each other. In such cases, R-squared may suggest a stronger relationship between the independent variables and the dependent variable than actually exists.
5. R-squared can be affected by outliers or influential observations that have a strong influence on the regression line.
6. R-squared is not an appropriate measure of model performance and invalid for non-linear regression models.

The practice of utilizing R-squared in nonlinear regression is not commonly adopted. While R-squared is useful in linear regression, as it compares the best fit regression line with a basic horizontal line that has a slope of 0.0 and intercept of mean, the same approach may not be appropriate for nonlinear regression. This is because, in most models used for nonlinear regression, a horizontal line is not a straightforward case and cannot be created using any set of parameters from the model. Therefore, comparing the goodness of fit of the chosen model with that of a horizontal line is not a valid comparison method for nonlinear regression models (Spiess & Neumeyer, 2010).

According to the weakness for non-linear regression models of the R-squared, it is generally not appropriate to use R-squared as an evaluation metric for tree-based models such as decision trees, random forests, and XGBoost. This is because the nature of these

models is fundamentally different from linear regression models. Instead, metrics such as mean absolute error (MAE), mean squared error (MSE), or root mean squared error (RMSE) are more appropriate for evaluating the performance of regression tree models. These metrics quantify the average difference between the predicted and actual values of the dependent variable, which is more relevant for evaluating the accuracy of the model's predictions.

5.2) Root Mean Squared Error (RMSE)

One metric used to gauge a model's errors is the Root Mean Squared Error (RMSE). Its value is the square root of the average of the squares of the errors, the difference between the expected and actual values. More precision is indicated by a lower score.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \quad \text{.....(17)}$$

where,

n is the number of observations

Y_i is actual values

$\hat{Y}_i$ is predicted values

6. Feature importance

6.1) SHAP value (SHapley Additive exPlanations)

In order to break down and explain machine learning models, SHAP value was offered in this study.

SHAP (SHapley Additive exPlanations) values are a method for explaining the output of machine learning models. They provide a way to understand how different features contribute to the final prediction of the model.

The SHAP value of a feature represents the contribution of that feature to the difference between the predicted value and the average predicted value for a given set of data. It is calculated using game theory, specifically the concept of Shapley values, which provides a fair way to allocate the contribution of each feature in a cooperative game.

SHAP values provide a unified and consistent way to explain the predictions of any model, including linear regression, decision trees, neural networks, and ensemble models. They can be used to identify which features are the most important for a given prediction, and how the values of those features affect the output.

In addition to feature importance, SHAP values can also be used to explain the behavior of the model for specific instances of data. This allows for more detailed and nuanced explanations of the model's behavior, and can help to build trust and transparency in machine learning models. Figure 10 illustrates how SHAP explain the features.

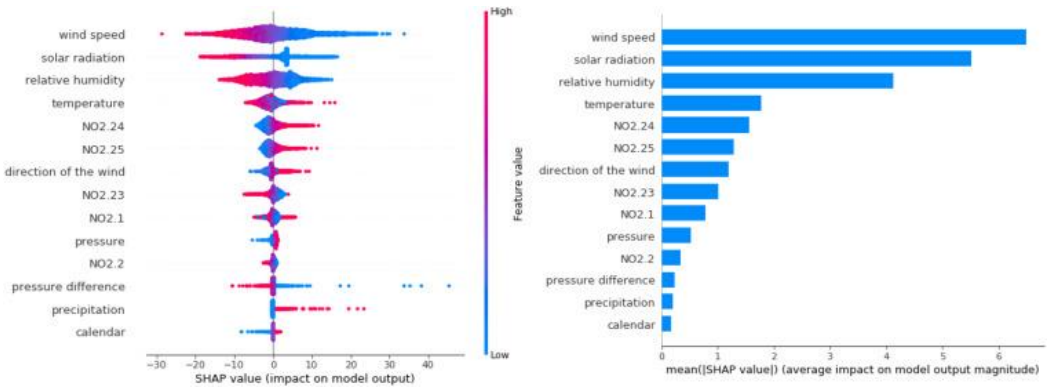


Figure 10 Example of SHAP feature importance

(Source: (Vega García & Aznarte, 2020))

2.2 Related Research

1. Analyzing impact of factors on stock market in various countries

Numerous studies have investigated the effects of macroeconomic indicators on the stock market. As a result, the outcomes are still ambiguous. High inflation, for instance, will increase the cost of commodities and lower purchasing power. As a result, it affects the cost of living for the typical person and could have an adverse effect on business profits because of greater input costs. Inflation is therefore anticipated to have an adverse effect on stock markets, despite contrasting findings from earlier research. The relationship between the inflation rate and the U.S. stock market was reported to be negative by (Fama, 1981) in contrast to (Boudoukh & Richardson, 1993) finding that it was positively correlated.

Besides the U.S., many stock markets have been examined. (Pražák, 2018) studied the effects of macroeconomic and microeconomic factors on company stock prices in the Czech Republic from 2006 to 2016 using the Johansen and Juselius cointegration test, the Hansen cointegration test, and the VECM model to examine the long-run relationships between selected indicators and stock prices. The research results showed that industrial

production, gross domestic product, and profitability ratios significantly influence stock prices in the long run. (Adam & Tweneboah, 2008) showed that interest rates and inflation significantly influence stock prices in Ghana by using Johansen's cointegration test and the VECM model. The OLS method was employed by (Rjoub et al., 2009) to determine how macroeconomic factors affected various portfolios in the Istanbul stock market. They discovered that inflation has positive effects on all portfolios, but only seven out of thirteen of these effects are significant. They also found that the exchange rate has a variety of effects and the unemployment rate has positive effects, but none of them are significant, and only two of the ten portfolios are differently significantly impacted by the money supply (M1).

In Asia, the long-run and short-run relationships between macroeconomic factors and the Malaysian stock market indices for the years 1977 to 2011 were examined by (Bekhet & Mugableh, 2012) using the ARDL bounds test approach. The study demonstrated that, over the long run, GDP has a positive impact on stock indexes while exchange rates, inflation, money supply (M3), and producer price index have a negative impact. However, it experienced positive short-term effects from inflation and the money supply M3, while experiencing negative short-term effects from exchange rates, GDP, and the producer price index. Regarding the relationship between exchange rates and stock prices, (Narayan et al., 2012) found that in eight Asian countries, namely Hong Kong, Indonesia, Japan, Korea, Malaysia, the Philippines, Singapore, and Thailand, the exchange rates and stock prices had only a short-run relationship from 1991 to 2005.

In Vietnam, many studies have been conducted using various approaches and different economic indicators to examine the impact of those indicators on the stock market. There, (DAO et al., 2022) examined the impact of macroeconomic factors on stock prices in the short and long run through

Vietnam's stock market, the VN-Index, using the ARDL model. The research results showed that in the long run, there's no relationship between world oil prices and interest rates on the VN-Index, the money supply, and the exchange rate have a positive and negative effect on the stock market, respectively. In contrast, interest rates and exchange rates have a negative effect in the short term. (Le et al., 2019) used the OLS, ECM model and the Granger causality test to examine the long- and short-term relationships between stock price and macroeconomic variables. In the long-term, the oil price, money supply, and interest rate have a positive impact on the VN-Index, while the exchange rate and SJC gold price have a negative impact, and the consumer price index showed insignificantly. In the short term, the index has a positive relationship with oil price volatility and a negative relationship with interest rate variability, and the Granger causality test also revealed that oil prices, money supply, and interest rates have a causal relationship with the VNI stock price index.

By using the VECM model, (Duy, 2016) studied the impact of macroeconomic factors on the stock price index through Vietnam's stock market, the VN-Index stock price. Six indicators, including the consumer price index (CPI), industrial production indices, interest rates, the VND/USD exchange rate, retail oil prices, and gold prices, were examined. The study showed that the VN-Index stock price has a positive relationship with industrial production indices and retail oil prices. On the contrary, the rest of the indicators, along with the consumer price index, interest rates, exchange rate, and gold prices, negatively affected the VN-Index stock price. According to (Duy & Hau, 2017), the money supply had a positive relationship with the VN-Index. On the other hand, the index had negative relationships with exchange rates and inflation.

Other than economic indicators, numerous pieces of research have reported that the Vietnamese stock market has been influenced by global economies and commodity prices. Macroeconomic news updates from the U.S. had a substantial impact on the Vietnamese stock market, as demonstrated by (Mai, 2016). According to (Hussainey & Khanh Ngoc, 2009), there is a strong positive correlation between Vietnamese stock prices and the US industrial sector and the US money market, and a positive relationship between VNI stock prices and the S&P 500 (Hussainey & Khanh Ngoc, 2009) (Tien, 2021). Regarding the relationship between commodity prices and the Vietnamese stock market, (Tien, 2021) (Nguyen et al., 2020) found that Vietnamese stock prices had a positive relationship with the oil price. Besides, the VN-Index showed a negative relationship with the gold price. All of research about analyzing impact of any factors are aggerated in Table 2.

Table 2 Related research

No.	Author	Data			Methods	Findings
		Countries /Target (Y)	Periods	Variables (X)		
1	(Fama, 1981)	U.S.stock returns	post-1953	Inflation	Economic regression model	There's a negative relationship between stock returns and inflation (-).
2	(Boudoukh & Richardson, 1993)	U.S.stock returns	1802-1990	Inflation	Economic regression model	There's a positive relationship between nominal stock returns and inflation at long horizons (+).
3	(Pražák, 2018)	Czech Republic stock prices of companies listed on the Prague Stock Exchange	2006 to 2016	Macroeconomic and microeconomic factors	-Johansen and Juselius (1990) and Hansen (1982) - Vector Error Correction Model (VECM)	**In long run**, industrial production, the gross domestic product and profitability ratios have significant impact on stock prices.

No.	Author	Data			Methods	Findings
		Countries /Target (Y)	Periods	Variables (X)		
4	(Adam & Tweneboah, 2008)	Ghana Databank Stock Index	January 1991 to April 2006	- Net foreign direct investments - Treasury bill rate - Consumer price index - Exchange rate	- Augmented Dickey-Fuller (ADF) and Phillips-Perron -Johansen's multivariate cointegration test - Vector Error Correction Model (VECM)	-**In long run,** interest rate and inflation have very significant impacts on stock price. - **In short run**, inflation and exchange rates matter on share price movements.
5	(Rjoub et al., 2009)	Istanbul Stock Exchange (ISE)	January 2001 to September 2005 (monthly)	- Interest rate - Unanticipated inflation - Risk premium - Exchange rate - Money supply (M1)	OLS linear model as suggested by CR&R (1986)	- Inflation has significant positive effects on seven out of thirteen portfolios (+). - Exchange rate has a variety of effects on all portfolios. - Unemployment rate has insignificant positive effects. - Two out of the ten portfolios are

No.	Author	Data			Methods	Findings
		Countries /Target (Y)	Periods	Variables (X)		
				- Unemployment rate		differently significantly impacted by the money supply (M1).
6	(Bekhet & Mugableh, 2012)	Malaysian stock market	1977-2011	- Money supply (M3) (RM billion) - Exchange Rate (RM/US$) - GDP - PPI - CPI	-ARDL Bound test	**In long run,** - PPI, CPI, Exchange Rate, and M3 have a negative impact (-). - GDP has a positive effect (+). **In short run,** - GDP, PPI, and Exchange Rate have a negative impact (-). - CPI and M3 have a positive effect (+).
7	(Narayan et al., 2012)	Hong Kong, Indonesia, Japan, Korea, Malaysia, the Philippines, Singapore and Thailand	1991-2005	- Exchange rates	-Cointegration test -Granger causality tests	- Exchange rates and stock prices are not cointegrated. - For the eight countries, exchange rates and stock prices had only a short-run relationship.

No.	Author	Data			Methods	Findings
		Countries /Target (Y)	Periods	Variables (X)		
8	(DAO et al., 2022)	VN-Index (VNI)	2010-2021	- World oil prices - Interest rates - Money supply M2 - Exchange rates	- ADF stationary test - Autoregressive distributed lag model (ARDL)	**In long run** - Money supply and exchange rate have positive (+) and negative (-) impacts on the stock market, respectively. - No relationship between world oil price and interest rates on the stock market. **In short run** - Interest rates and exchange rates have a negative impact (-).
9	(Le et al., 2019)	VN-Index (VNI)	January 2009 to December 2018	- Oil price - Money supply - Interest rate - SJC gold price - Consumer price index (CPI)	- OLS - Vector Error Correction Model (VECM)	**In long run** - SJC gold price and exchange rate have a negative impact (-). - Oil prices, money supply, and interest rates have a positive impact (+). - CPI is not statistically important.

No.	Author	Data			Methods	Findings
		Countries /Target (Y)	Periods	Variables (X)		
				- Exchange rate.		**In short run** - The volatility of oil prices is positively related (+). - The variability of interest rates has negative impacts (-).
10	(Duy, 2016)	Vietnam's stock market	2005-2014	- Consumer price index - Industrial production indices - Interest rates - Exchange rate, - Retail oil prices, - Gold price	- ADF and Phillip-Person (PP) stationary test -Vector Error Correction Model (VECM)	**In long run** - Consumer price index, interest rates, exchange rate, and gold price have a negative impact (-). - Retail oil prices and industrial production indices have a positive impact (+).
11	(Duy & Hau, 2017)	VN-Index (VNI)	January 2006 to December 2015	- Consumer price index - Exchange rate	- ADF Test - Vector Error Correction Model	-Market price index and money supply have a positive impact (+). - CPI and exchange rate have a negative

No.	Author	Data			Methods	Findings
		Countries /Target (Y)	Periods	Variables (X)		
				- Money supply (M2)	(VECM)	impact (-).
12	(Mai, 2016)	Vietnamese market	Jan 1, 2010 - Dec 1, 2015	- Macroeconomic announcements from U.S. and Vietnam Central Bank. - Interest rate changing announcement from Vietnam.	-GARCH	-The Vietnamese stock market is influenced by U.S.'s - Interest rate changing announcement has no impact on the market.

No.	Author	Data			Methods	Findings
		Countries /Target (Y)	Periods	Variables (X)		
13	(Hussainey & Khanh Ngoc, 2009)	Vietnamese stock prices.	January 2001 to April 2008 (monthly)	Domestic indicators: interest rate, industrial production, consumer prices index US stock price - S&P 500	- Monthly time series introduced by Nasseh and Strauss (2000) and Canova and de Nicolo (1995) (Regression based model)	-There is a positive relationship between VNI stock prices and S&P500 - U.S. industrial sector and the money market have significant positive relations with the Vietnamese stock market.
14	(Tien, 2021)	VN-Index (VNI)	January 2020 to June 2021	- S&P500 - Oil price - Gold price - Total number of Covid-19 infections	- Autoregressive distributed lag model (ARDL)	**In long run** - S&P500 and oil price have a positive relationship (+). - Gold price has a negative relationship (-). - The number of Covid-19 infections does not affect the market.

No.	Author	Data			Methods	Findings
		Countries /Target (Y)	Periods	Variables (X)		
15	(Nguyen et al., 2020)	VN-Index (VNI) and HXN index	August 2000 to October 2019	- Brent oil price - WTI oil price, - USD/VND - EUR/VND	- GARCH (1,1)	-Oil price has a significant positive impact on the stock market indices (+). -Oil price has a significant negative effect on stock market volatility (-). -The exchange rates inconsistently influenced Vietnamese stock market indices. -USD/VND significantly impacts the market return and volatility.

According to studies above, the summary results of related research about the Vietnamese stock market is shown in Table 3.

Table 3 Summary results of related research about Vietnamese stock market

Indicators	long run sign	Studies	short run sign	Studies
Interest rate	+	(Le et al., 2019)	+	-
	-	-	-	(DAO et al., 2022), (Le et al., 2019)
	No	(DAO et al., 2022), (Duy, 2016)	No	-
Money supply	+	(DAO et al., 2022), (Le et al., 2019)	+	-
	-	-	-	-
	No	-	No	-
Oil price	+	(Le et al., 2019) (Duy, 2016), (Duy & Hau, 2017), (Tien, 2021), (Nguyen et al., 2020)	+	(Le et al., 2019)
	-	-	-	(Nguyen et al., 2020)
	No	(DAO et al., 2022)	No	-
Exchange rate	+	-	+	-
	-	(DAO et al., 2022)), (Le et al., 2019), (Duy, 2016), (Duy & Hau, 2017)	-	(DAO et al., 2022)
	No	(Nguyen et al., 2020)	No	-
Gold price	+	-	+	-
	-	(Le et al., 2019), (Duy, 2016), (Tien, 2021)	-	-
	No	-	No	-
Consumer	+	-	+	-

Indicators	long run sign	Studies	short run sign	Studies
price index (CPI)	-	(Duy & Hau, 2017)	-	-
	No	(Le et al., 2019), (Duy, 2016)	No	-
Industrial production index	+	(Duy, 2016)	+	-
	-	-	-	-
	No	-	No	-
S&P 500	+	(Hussainey & Khanh Ngoc, 2009)	+	-
	-	-	-	-
	No	-	No	-

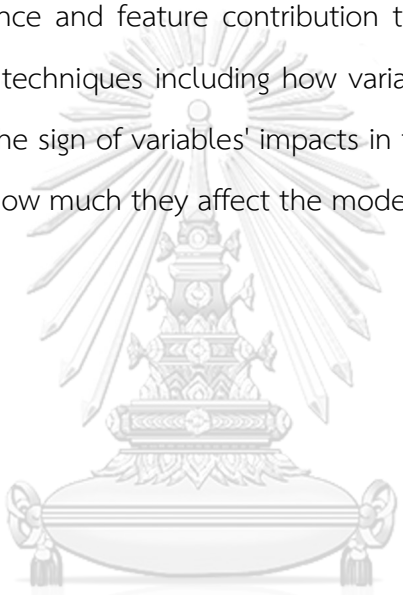
2. Utilizing machine learning techniques to analyze any impacts on time series data

Most studies that have explored impact of certain factors on time series data used statistical regression based models like those mentioned above, here is a study that has utilized a different method, namely machine learning.

(Al-Maadid et al., 2022) examined the impact of coronavirus pandemic on the stock markets in five GCC countries using machine learning techniques including Decision Tree, Random Forest and XGBoost. They conducted the study using economic indicators, namely interest rate, GDP surprises, U.S. stock volatility index (VIX), oil price, and daily returns of other GCC markets, to build based models, then adding Covid-19 features such as the number of local cases, global cases, local deaths, global deaths and local recoveries to see how those models improved from RMSE. Afterward, feature importance is applied to draw feature importance ranking of features. They found that XGBoost is outstanding, and four out of five stock markets are influenced by

Covid-19 since the model's performance improved after adding Covid-19 features, and each market is mainly affected by different features.

Instead of applying traditional statistical techniques as in previous research on the same topic which only determine statistical significance of variables and the sign of those variables' impacts, three machine learning models will be constructed including Decision Tree, Random Forest and XGBoost as (Al-Maadid et al., 2022) have done in their research, then explore feature importance and feature contribution to gain more beneficial results than traditional techniques including how variables affect the models, which is the same as the sign of variables' impacts in traditional techniques, and the more insight is how much they affect the models.



Chapter 3

Research Methodology

3.1 Data

The data used in this study is monthly data for the period August 2000 to June 2022, with 263 observations (21 years and 11 months). The VN-Index, S&P 500 index, Brent oil price, and exchange rate (Vietnamese Dongs to US Dollars) were obtained from the Investing.com website, and other economic indicators, such as the consumer price index (which measures inflation), imports, exports, GDP, and the unemployment rate, were obtained from the tradingeconomics.com website. It notes that the unemployment rate and GDP values are provided on a quarterly and annual basis, respectively. The variables' definition and unit are explained in Table 4.

Table 4 Variables' definition

Variables	Definition
Price	VN-Index, the main stock market index of the Ho Chi Minh Stock Exchange (HOSE), largest stock market in Vietnam.
Brent	Brent crude, which serves as one of the three main benchmarks for crude oil prices per barrel, is produced by blending crude oil from 19 different North Sea oil sources. This blending process combines various crudes to create a relatively light type of crude oil known as Brent.
CPI	Consumer price index (which measures inflation) of Vietnam.
Exchange rate	Exchange rate of Vietnamese Dongs to US Dollars.
Exports	Vietnam's total export value in Billion USD
GDP	The economic value of all the end products and services manufactured and sold (excluding resale) within Vietnam during a specific timeframe is quantified in dollars through the concept of "gross domestic product" (GDP) of Vietnam, measured in billions of US dollars.

Variables	Definition
Imports	Vietnam's total import value in Billion USD
S&P 500	The S&P 500, also known as the Standard and Poor's 500, is a stock market index that gauges the performance of 500 prominent companies listed on American stock exchanges. It is widely recognized as one of the primary equity indices that receives significant attention and scrutiny.
Unemployment rate	The percentage of the Vietnamese labor force without a job in %YoY

3.2 Methodology

1. Data Collection

The data used in this study is monthly data for the period August 2000 to June 2022, with 263 observations.

- The VN-Index, S&P 500 index, Brent oil price, and exchange rate (Vietnamese Dongs to US Dollars) were obtained from the Investing.com website.
- Other economic indicators, such as the consumer price index, imports, exports, GDP, and the unemployment rate, were obtained from the tradingeconomics.com website.

2. Data Cleaning

All variables were obtained on a monthly basis except the unemployment rate and GDP values are provided on a quarterly and annual basis, respectively. The missing data were filled by interpolated values among those periods.

1. Exploratory Data Analysis (EDA)

   Exploratory Data Analysis (EDA) is an approach to analyzing and summarizing datasets to uncover insights and identify patterns,

relationships, and anomalies in the data. It is an important step in the data analysis process, as it helps to understand the underlying structure and distribution of the data, identify outliers, and check for missing or inconsistent values including:

- Descriptive Statistics
- Pearson correlation coefficient

3. Data Standardization

Although the machine learning techniques such as Decision Tree, Random Forest, and XGBoost do not require any data transformation, standardizing data before doing statistic models may have some benefits. The benefit of standardizing the data is that it makes it easier to compare the magnitudes of the coefficients for the different independent variables in the model. When the data are not standardized, the coefficients are affected by the scale of the variable, which can make it difficult to compare their relative importance.

To avoid any doubts about the results, standardized data were used for both statistics and machine learning models.

$$z = \frac{x - \mu}{\sigma} \quad \text{.....(15)}$$

where μ is the mean and σ is the standard deviation.

4. Statistical models

Statistical models construction for both long and short run relationship has procedures as follows:

a) Perform the Augmented Dickey-Fuller (ADF) test to determine the stationarity of the time series variables and to determine the order of integration (I(0), I(1), I(2), etc.) for each variable.

b) Selected an appropriate statistical method for the long and short run based on the results of the ADF test.

- If all variables were I(0), the Ordinary Least Squares (OLS) method was employed.
- If all variables were I(1), either the Johansen cointegration test or the Engle-Granger test was conducted.
- If variables were a mix of I(0) and I(1), an ARDL bound test was executed.

c) Divided data into 2 sets : an 80% training set and a 20% test set.

d) The selected models were built using the training set data to develop long and short run models.

e) The performance of the models was evaluated by using the Root Mean Squared Error (RMSE) on the test set.

5. Machine Learning Models for short run relationship

Machine learning models construction for short run relationship has procedures as follows:

a) Employed differencing techniques to transform the data, thereby capturing exclusively the short-run relationships among the variables.

b) Divided data into 2 sets : an 80% training set and a 20% test set.

c) Constructed machine learning models utilizing the training set data, incorporating a 10-fold cross-validation methodology to fine-tune the hyperparameters and systematically assess the performance of each model across multiple validation sets.

d) Conducted model performance evaluation by employing the hyperparameters associated with the highest average accuracy (or lowest RMSE) observed during the cross-validation process,

subsequently applying these optimal settings to the independent test set.

e) Computed the Root Mean Squared Error (RMSE) metric for the best-performing model, serving as a quantitative assessment of the model's predictive accuracy when applied to the test set.

f) Applied the SHAP (SHapley Additive exPlanations) value, which is a method for interpreting the impact of each feature on the model's predictions to gain insights into the relationship between the features and the target variable.

6. Machine Learning Models for lond run relationship

Machine learning models construction for long run relationship has procedures as follows:

a) Divided data into 2 sets : an 80% training set and a 20% test set.

b) Constructed machine learning models utilizing the training set data, incorporating a 10-fold cross-validation methodology to fine-tune the hyperparameters and systematically assess the performance of each model across multiple validation sets.

c) Conducted model performance evaluation by employing the hyperparameters associated with the highest average accuracy (or lowest RMSE) observed during the cross-validation process, subsequently applying these optimal settings to the independent test set.

d) Computed the Root Mean Squared Error (RMSE) metric for the best-performing model, serving as a quantitative assessment of the model's predictive accuracy when applied to the test set.

e) Applied the SHAP (SHapley Additive exPlanations) value, which is a method for interpreting the impact of each feature on the model's

predictions to gain insights into the relationship between the features and the target variable.

.

7. Results comparison

Compared the long run and short run results from both statistical and machine learning model in terms of model performance and their relationship results.

The research methodology is illustrated in Figure 11 and Figure 12.

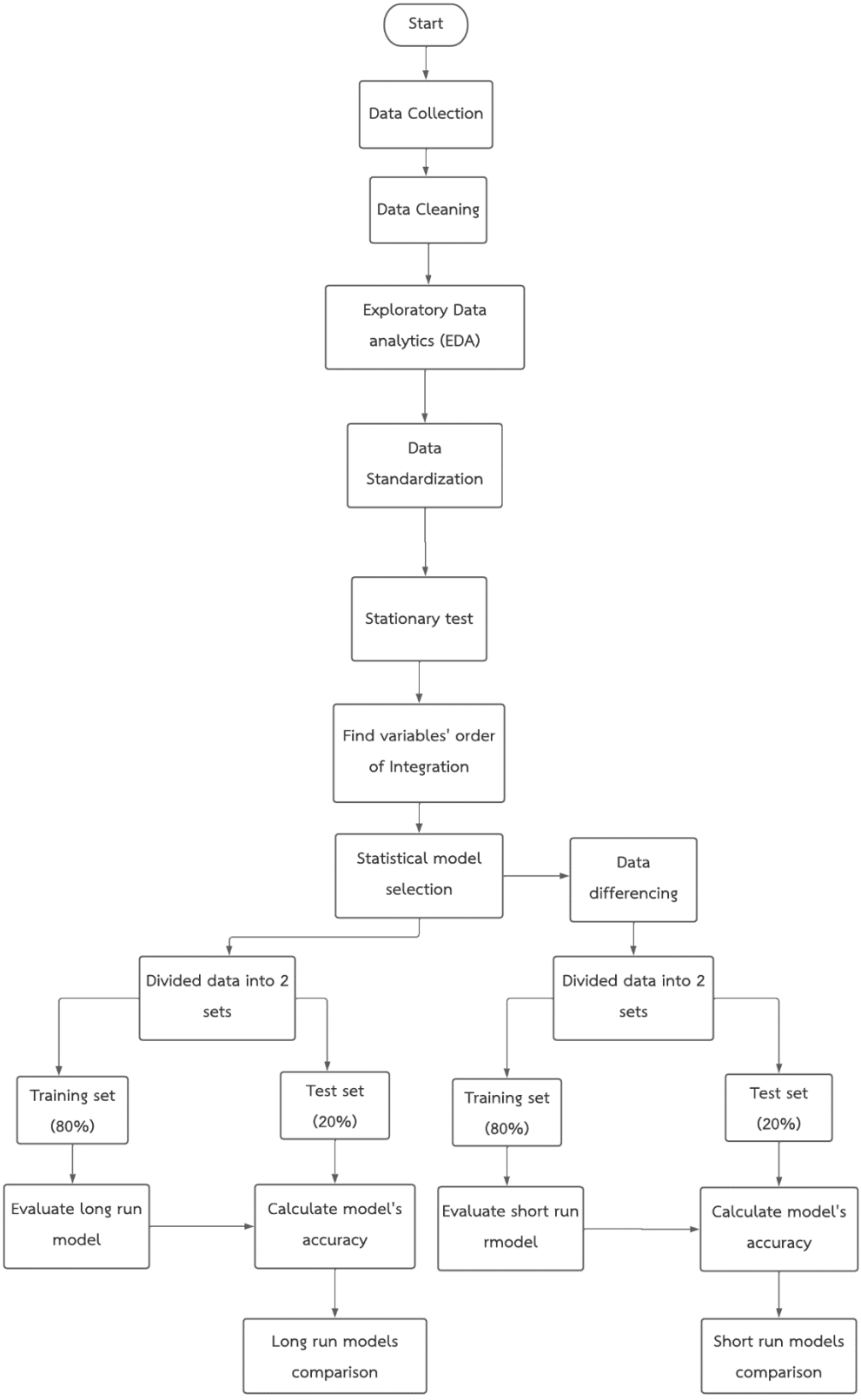


Figure 11 Methodology for statistical techniques

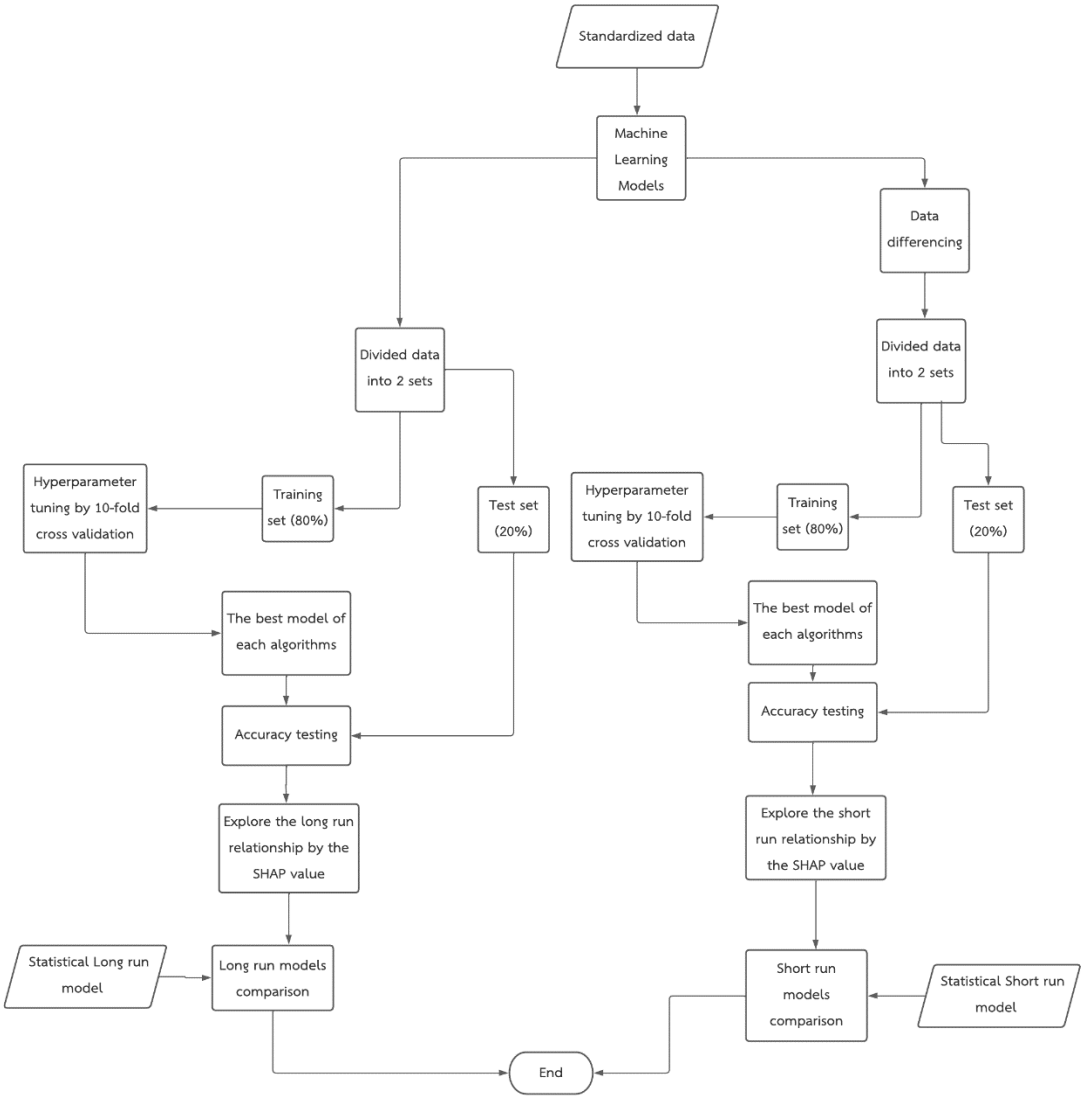


Figure 12 Methodology for machine learning techniques

Chapter 4

Results and discussion

4.1 Exploratory Data Analysis (EDA)

1. Descriptive statistics

The descriptive statistics of variables are demonstrated in Table 5.

Table 5 Descriptive statistics of variables

	Price	Brent	CPI	EXCH	Exports	GDP	Imports	S&P500	UNEM
Mean	590.317	66.374	67.591	19323	10.450	168.908	10.578	1840.673	2.349
Standard Error	20.592	1.788	1.701	195.917	0.537	6.948	0.507	57.817	0.020
Median	510.600	64.140	74.140	20810	7.498	162.017	8.654	1416.180	2.300
Mode	#N/A	55.580	29.130	22750	1.400	#N/A	1.380	1130.200	2.300
Standard Deviation	333.950	29.000	27.581	3177.235	8.703	112.685	8.229	937.627	0.332
Sample Variance	111522.6	841.009	760.698	10094822	75.739	12697.981	67.715	879144	0.110
Kurtosis	-0.070	-0.856	-1.585	-1.662	-0.295	-1.326	-0.256	1.028	6.970
Skewness	0.784	0.324	-0.104	-0.145	0.847	0.287	0.808	1.324	2.109
Range	1383.13	120.690	79.340	9496	33.760	345.310	31.608	4031.09	2.170
Minimum	115.150	19.140	29.080	14120	0.950	30.340	1.052	735.090	1.810
Maximum	1498.28	139.830	108.420	23616.	34.710	375.650	32.660	4766.18	3.980
Sum	155253	17456	17777	5081922	2748.37	44423	2782.010	484097	617.80
Count	263	263	263	263	263	263	263	263	263

The descriptive statistics presented in Table 5 offer valuable insights into the data set being analyzed, indicating that each of the variables is on a different scale. Moreover, the skewness of the data is an important characteristic to consider. Specifically, the skewness values reveal that the Price (VN Index), Brent, exports, GDP, Imports, S&P500, and unemployment rate variables all exhibit positive skewness, indicating that their distributions

are right-skewed. On the other hand, the CPI and exchange rate exhibit a little left-skewed distribution. While this information is useful, a more comprehensive view of the data's distribution is provided by the histograms presented in Figure 13, which not only reinforce the skewness information but also highlight that none of the variables exhibit normal and symmetric distributions.

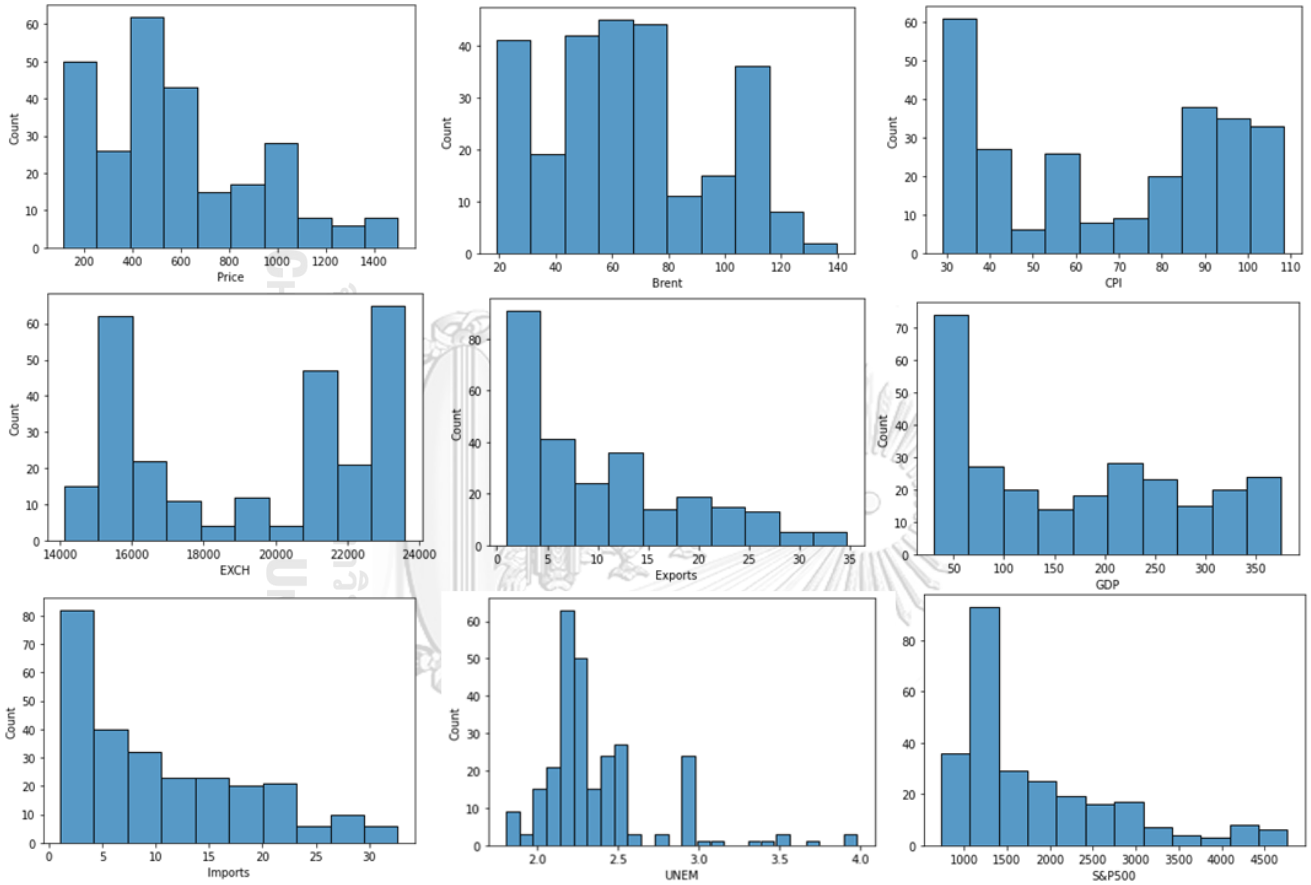


Figure 13 The variables' distribution

2. Pearson correlation coefficient

In Table 6, we can see that the Pearson correlation coefficients reveal some interesting insights about the relationship between the stock price and the various features under consideration. Among these features, the Brent oil price and unemployment rate exhibit a low linear relationship with the stock price, indicating that changes in these variables are not strongly associated with changes in the stock price.

However, it's worth noting that there are also strong linear correlations among all eight features, which raises concerns about the appropriateness of using a conventional multiple linear regression model. Specifically, this may violate several key assumptions of linear regression, such as independence, homoscedasticity, and normality, which can lead to the issue of spurious regression.

Additionally, if there are strong linear relationships among the features in the dataset, stepwise regression might face some difficulties and give unreliable results. When there is high multicollinearity, stepwise regression might struggle to accurately pick the most important predictors and provide stable coefficient estimates. To determine the extent of correlation between the features, we can calculate correlation coefficients or use other methods like variance inflation factors (VIF). If the correlations are indeed strong, it indicates a high level of multicollinearity. In such cases, these features will be removed from the model, and the final model will only include a few selected features.

Despite these concerns, one interesting finding is that all of the features under consideration exhibit a positive linear relationship with the VN Index, as evidenced by the positive signs of the correlation coefficients.

Table 6 Pearson correlation coefficients

	Price	Brent	CPI	EXCH	Exports	GDP	Imports	S&P500	UNEM
Price	1								
Brent	0.2715	1							
CPI	0.7192	0.3921	1						
EXCH	0.6676	0.3581	0.9885	1					
Exports	0.8283	0.2191	0.9167	0.8876	1				
GDP	0.7836	0.2835	0.9818	0.9652	0.9669	1			
Imports	0.8421	0.2527	0.9210	0.8903	0.9940	0.9674	1		
S&P500	0.8744	0.1130	0.8051	0.7578	0.9466	0.8929	0.9463	1	
UNEM	0.2295	-0.0224	0.0909	0.0880	0.1708	0.1327	0.1865	0.2462	1

4.2. Statistical models

1. Augmented Dickey-Fuller (ADF) unit root test

In order to accurately analyze the time series data under consideration, it is important to first identify whether the variables are stationary or non-stationary. To accomplish this, the Augmented Dickey-Fuller (ADF) unit root test was conducted and the results are presented in Table 7.

Upon examining the ADF values, it became clear that all of the variables under consideration, with the exception of the unemployment rate (UNEM), exhibited non-stationary data. Specifically, the ADF values for these variables were greater than the critical values of 1%, 5%, and 10%, or their corresponding p-values were greater than 0.05, indicating that they did not meet the criteria for stationarity.

However, it's worth noting that the unemployment rate (UNEM) was found to have an order of integration of zero, meaning that it is a stationary variable. This suggests that the behavior of the unemployment rate over time is relatively stable and does not exhibit significant fluctuations or trends.

Table 7 ADF values

	ADF Statistics	P-value	1% Critical Values	5% Critical Values	10% Critical Values
Price	-1.6111	0.4755	-3.4553	-2.8724	-2.5726
Brent	-2.2863	0.1772	-3.4553	-2.8724	-2.5726
CPI	-0.0311	0.9540	-3.4553	-2.8724	-2.5726
Exchange	-1.1547	0.6943	-3.4552	-2.8724	-2.5726
Exports	5.2999	1.0000	-3.4564	-2.8729	-2.5729
GDP	0.3267	0.9793	-3.4565	-2.8730	-2.5729
Log(GDP)	-2.2045	0.2053	-3.4565	-2.8730	-2.5729
Imports	3.6029	1.0000	-3.4564	-2.8729	-2.5729
S&P500	0.5554	0.9883	-3.4552	-2.8724	-2.5726
UNEM	-3.7730	0.0036	-3.4552	-2.8724	-2.5726

Upon conducting further analysis, it was discovered that while most of the variables under consideration exhibited non-stationary data in their original form, they could be transformed into stationary data through first differencing. As Table 8 illustrates, this was true for all variables with the exception of GDP.

In the case of GDP, the integration order was found to be larger than one, meaning that it was not possible to use any statistical models for this analysis. However, after taking the logarithm of GDP, the results revealed that the logarithm of GDP (Log(GDP)) became stationary data after the first differencing, indicating that it had an integration order of one.

As a result of these findings, the logarithm of GDP (Log(GDP)) was used for the statistical analysis instead of GDP. The conclusions regarding the order of integration for all variables under consideration are presented in Table 9.

Table 8 ADF values after differencing

	ADF Statistics	P-value	1% Critical Values	5% Critical Values	10% Critical Values
D(Price)	-12.282	0.0000	-3.455	-2.872	-2.573
D(Brent)	-12.155	0.0000	-3.455	-2.872	-2.573
D(CPI)	-8.554	0.0000	-3.455	-2.872	-2.573
D(Exchange)	-16.092	0.0000	-3.455	-2.872	-2.573
D(Exports)	-4.764	0.0000	-3.457	-2.873	-2.573
D(GDP)	-2.074	0.2554	-3.457	-2.873	-2.573
D(D(GDP))	-7.984	0.0000	-3.457	-2.873	-2.573
D(Log(GDP))	-3.058	0.0310	-3.457	-2.873	-2.573
D(Imports)	-4.295	0.0006	-3.457	-2.873	-2.573
D(S&P500)	-16.482	0.0000	-3.455	-2.872	-2.573
D(Unem)	-16.093	0.0000	-3.455	-2.872	-2.573

Table 9 Orders of integration of variables

	Order of Integration
Price	I(1)
Brent	I(1)
CPI	I(1)
Exchange	I(1)
Exports	I(1)
GDP	I(2)
Log(GDP)	I(1)
Imports	I(1)
S&P500	I(1)
Unem	I(0)

Given that the variables under consideration have different orders of integration, it is essential to employ appropriate methods to determine whether there exists a long run relationship among them. In order to accomplish this, the autoregressive distributed lag (ARDL) bound test will be used, which can help to identify whether the variables are co-integrated and exhibit a long run relationship.

In addition, it is important to examine the short run relationship among the variables, as the differenced data only captures short run information. Since all variables become stationary after the first differencing, the ordinary least squares (OLS) method will be used to indicate the short run relationship among the variables. By utilizing both the ARDL bound test and OLS method, it is possible to obtain a more comprehensive understanding of the relationships among the variables under consideration over both the long and short term.

2. Statistical model for short run relationship

2.1) Short run model from OLS

Upon constructing a short run model using the OLS method and training data set, the results were analyzed and the findings are presented in Figure 14.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(BRENT)	0.064556	0.051916	1.243462	0.2152
D(CPI)	-1.466529	0.688237	-2.130847	0.0343
D(EXCH)	-0.697362	0.208576	-3.343438	0.0010
D(EXPORTS)	-0.079137	0.089041	-0.888769	0.3752
D(LOG_GDP)	0.012634	0.099496	0.126975	0.8991
D(IMPORTS)	0.222554	1.434599	0.155133	0.8769
D(SP500)	0.732246	0.118866	6.160251	0.0000
D(UNEM)	-0.014158	0.021537	-0.657386	0.5117
C	0.027658	0.018902	1.463229	0.1450

R-squared	0.269126	Mean dependent var	0.014036
Adjusted R-squared	0.239892	S.D. dependent var	0.167367
S.E. of regression	0.145917	Akaike info criterion	-0.969444
Sum squared resid	4.258380	Schwarz criterion	-0.825516
Log likelihood	110.3069	Hannan-Quinn criter.	-0.911253
F-statistic	9.205644	Durbin-Watson stat	1.663507
Prob(F-statistic)	0.000000		

Figure 14 Short run model from OLS

The analysis of the short run model constructed using the OLS method and training data set revealed that, out of all the variables under consideration, only three exhibited statistical significance within the model. These three variables were identified as the CPI, exchange rate, and S&P500 index, and they were found to have varying impacts on the model depending on their respective signs of coefficients.

Specifically, the CPI and exchange rate were observed to have negative impacts on the model, while the S&P500 index exhibited a positive impact. Furthermore, upon examining the standardized coefficients of each variable, it was found that the CPI had the largest impact on the model, followed by the S&P500 index and exchange rate, respectively.

2.2) Short run statistical model performance analysis

Upon applying the constructed model to the test set, the results obtained were analyzed and compared to the training set values. The findings from this analysis are presented in Table 10, which reveals that the root mean squared error (RMSE) values in both the training and test sets are relatively close, with values of 0.1427 and 0.1396, respectively.

These findings suggest that the model is not overfitting or underfitting, and that the predictions generated by the model are reasonably accurate. By achieving a low level of RMSE in both the training and test sets, it is clear that the model has been constructed in a robust and reliable manner, and that it is capable of generating accurate predictions even when applied to new and previously unseen data.

Table 10 Performance analysis for short run statistical model

	Training set	Test set
RMSE	0.1427	0.1396

3. Statistical model for long run relationship

3.1) Order selection from ARDL model

In the process of building an ARDL Bound test, one of the key steps is to determine the appropriate order of the ARDL model that will be used to perform the analysis. This is a crucial step, as selecting the wrong order can result in inaccurate and unreliable results.

To address this issue, a rigorous process was employed to select the optimal order of the ARDL model. Specifically, a range of different order specifications were tested, and the Akaike Information

Criterion (AIC) was used as the primary metric for selecting the best performing model.

After careful consideration of the various order specifications, it was determined that the ARDL(2, 0, 0, 1, 0, 0, 1, 1, 0) model exhibited the lowest AIC value as shown in Figure 15, indicating that it was the optimal order specification for the ARDL Bound test. This finding is significant, as it provides a strong basis for conducting further analysis using the ARDL Bound test, and ensures that the results generated by the test will be accurate and reliable.

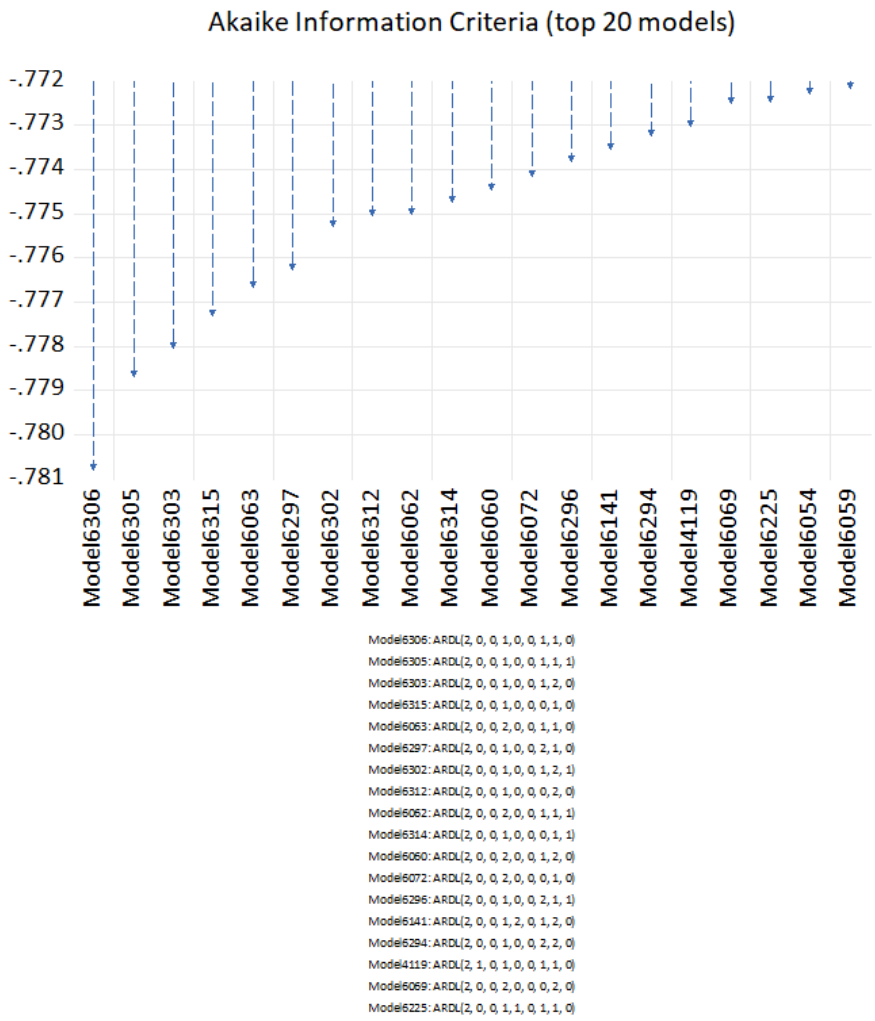


Figure 15 AIC values of top 20 models

3.2) ARDL Bound test

The ARDL Bound test is a widely used approach for examining the presence of cointegration and determining the long-run relationship between variables. In order to apply this test to the current data set, a series of rigorous analyses were conducted, culminating in the examination of the results shown in Figure 16.

Upon reviewing the results of the ARDL Bound test, it was found that the F-test value exceeded the critical value, which provides strong evidence that the null hypothesis of no long-run relationship can be rejected. This result is highly significant, as it indicates that there is indeed a long-run relationship between the variables under consideration.

F-Bounds Test		Null Hypothesis: No levels relationship		
Test Statistic	Value	Signif.	I(0)	I(1)
F-statistic	4.072500	10%	1.85	2.85
k	8	5%	2.11	3.15
		2.5%	2.33	3.42
		1%	2.62	3.77

Figure 16 ARDL Bound test

3.3) Statistical model for long run relationship

After conducting the ARDL Bound test to determine the presence of cointegration and the long-run relationship between variables, the next step was to derive the long-run equation that describes the relationship between the variables of interest. The long-run equation, presented in Figure 17, provides crucial insights into the underlying dynamics of the data.

Upon analyzing the results presented in the long-run equation, it was found that only three variables, namely the CPI, Log GDP, and S&P500 index, were found to be statistically significant to the model. According to the signs of the coefficients, the CPI has a negative impact on the model, while Log GDP and S&P500 index show a positive impact. Furthermore, the CPI was found to have the greatest impact on the model, followed by Log GDP and S&P500 index, respectively, according to their magnitude of standardized coefficients.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
BRENT	0.033177	0.097071	0.341783	0.7329
CPI	-3.595774	0.936790	-3.838399	0.0002
EXCH	0.388695	0.557786	0.696853	0.4867
EXPORTS	0.235097	0.670991	0.350372	0.7264
IMPORTS	-1.205797	0.788491	-1.529247	0.1278
LOG_GDP	3.242697	0.648523	5.000127	0.0000
SP500	1.860604	0.330973	5.621623	0.0000
UNEM	-0.078605	0.072653	-1.081928	0.2806
C	0.237225	0.128083	1.852116	0.0655

EC = PRICE - (0.0332*BRENT -3.5958*CPI + 0.3887*EXCH + 0.2351 *EXPORTS -1.2058*IMPORTS + 3.2427*LOG_GDP + 1.8606*SP500 -0.0786*UNEM + 0.2372)

Figure 17 Statistical model for long run relationship

3.4) Long run statistical model performance analysis

Upon evaluating the performance of the long-run model using the test set, it was observed from Table 11 that the root mean squared error (RMSE) values for the training and test sets are slightly different at 0.4598 and 0.5035, respectively. The difference between these values indicates that the model's performance might be slightly worse when predicting the test set, but the overall RMSE values

suggest that the model is not overfitting or underfitting and the predictions are reasonably accurate.

Table 11 Performance analysis for long run statistical model

	Training set	Test set
RMSE	0.4598	0.5035

4.3 Machine learning models

1. Machine learning models for short run relationship

1.1) Short run model construction and hyperparameter tuning

To ensure the robustness of machine learning models, the process of constructing these models typically involves training set data with a 10-fold cross validation approach, which enables the tuning of hyperparameters and the evaluation of each model's performance. To select the most appropriate hyperparameters, various values of hyperparameters are tested, and the selected hyperparameters are those that achieve the highest accuracy or lowest RMSE of each model. These selected hyperparameters are then presented in Table 12.

Table 12 Selected hyperparameters of each short run machine learning model

Hyperparameters	Decision Tree		Random Forest		XGBoost	
	Setting	Used (10- fold CV)	Setting	Used (10- fold CV)	Setting	Used (10- fold CV)
Impurity function	Squared Error	Squared Error	Squared Error	Squared Error	Squared Error	Squared Error
Tree depth	2,3,5,8,10	2	2,3,5,8,10	3	2,3,5,8,10	5
Minimum number of samples required to split	2,5,10,20,40	2	2,5,10,20,40	40	NA	N/A
Minimum number of samples required to be at leaf nodes	2,5,10,20,40	40	2,5,10,20,40	2	NA	N/A
Maximum number of leaf nodes	3,5,10	5	3,5,10	10	NA	N/A
The number of estimators	1	1	5,10,15,20,40	40	5,10,15,20,40	40
learning rate	N/A	N/A	N/A	N/A	0.01,0.03,0.1, 0.3	0.01
Minimum sum of weights of all observations required in a child	N/A	N/A	N/A	N/A	1,2,3	1
Pseudoregularization (Gamma)	N/A	N/A	N/A	N/A	0,0.1,0.2,0.3	0.1
L1 regularization (alpha)	N/A	N/A	N/A	N/A	0	0
L2 regularization (lambda)	N/A	N/A	N/A	N/A	1	1

1.2) Short run machine learning models' performance

After the construction of the machine learning models and the evaluation of their performance through the 10-fold cross-validation approach, the outcomes depicted in Table 13 indicate that among the various models, the random forest algorithm achieves the highest level of accuracy or the lowest RMSE value, as evidenced by the obtained average value of 0.1382. Furthermore, upon applying the random forest model to the test set, the RMSE value of 0.1374 was obtained, which is notably similar to the RMSE values derived from both the training and cross-validation data. This observation implies that the model exhibits consistency and robustness with regards to both seen and unseen data, and that it is free from overfitting or underfitting issues.

Table 13 Short run machine learning models' performance

	RMSE Train	RMSE CV	RMSE Test
Decision Tree	0.1498	0.1504	
Random Forest	0.1328	0.1382	0.1374
XGBoost	0.1252	0.1427	

1.3) Feature importance of short run machine learning model

After evaluating the machine learning models, it was determined that the random forest model achieved the highest accuracy or the lowest RMSE value at 0.1382, based on the average cross-validation data from a 10-fold cross-validation approach. To better understand the important features and their impact on the model predictions, the SHAP (SHapley Additive exPlanations) value was obtained from this model. The SHAP value is a method used to explain the output of any machine learning model and allows us to

identify the contribution of each feature to the prediction, providing a deeper insight into the model. The SHAP value was analyzed and the results are displayed in both Figure 18 and Figure 19.

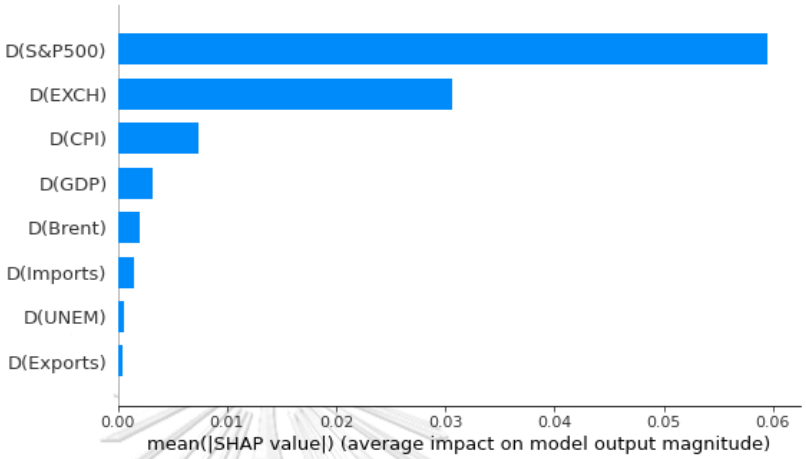


Figure 18 Feature importance from SHAP for short run model

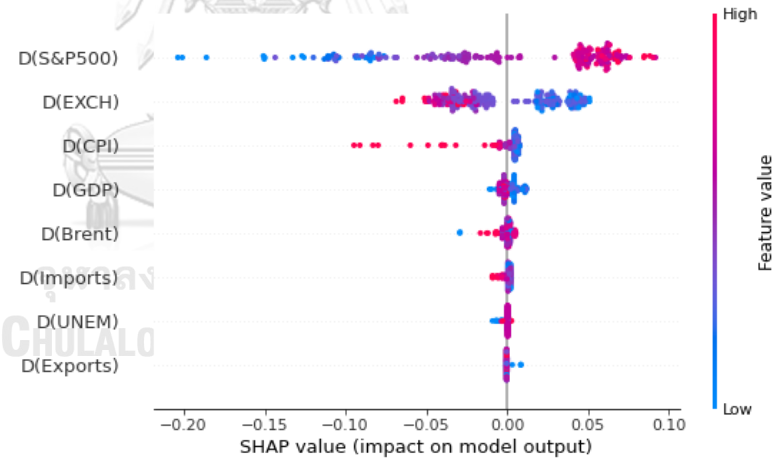


Figure 19 Feature contribution from SHAP for short run model

After analyzing the feature importance in Figure 18, it can be concluded that the S&P 500 index has the greatest impact on the model, implying that it has the most significant influence on the VN-Index compared to the other features. Furthermore, the exchange rate, consumer price index (CPI), and GDP have a noticeable impact on the model, respectively. Conversely, the impacts of Brent oil price,

imports, unemployment rate, and exports appear to have a minor impact on the model, indicating they have a smaller effect on the VN-Index change than the other four features.

To further understand the relationship between the features and their impact on the model, Figure 19 displays the feature contribution from the SHAP value. The findings reveal that higher values of the S&P 500 index have a greater positive impact on the model, or the pink dots representing high feature values are on the right side, indicating a positive impact on the VN-Index. Therefore, there is a clear positive relationship between the S&P 500 index and the VN-Index. However, the exchange rate, consumer price index (CPI), and GDP show a negative impact on the model. The contributions of changes in Brent oil price, imports, unemployment rate, and exports are still undetermined.

2. Machine learning models for long run relationship

2.1) long run model construction and hyperparameter tuning

To ensure the robustness of machine learning models, the process of constructing these models typically involves training set data with a 10-fold cross validation approach, which enables the tuning of hyperparameters and the evaluation of each model's performance. To select the most appropriate hyperparameters, various values of hyperparameters are tested, and the selected hyperparameters are those that achieve the highest accuracy or lowest RMSE of each model. These selected hyperparameters are then presented in Table 14.

Table 14 Selected hyperparameters of each long run machine learning model

Hyperparameters	Decision Tree		Random Forest		XGBoost	
	Setting	Used (10- fold CV)	Setting	Used (10- fold CV)	Setting	Used (10- fold CV)
Impurity function	Squared Error	Squared Error	Squared Error	Squared Error	Squared Error	Squared Error
Tree depth	2,3,5,8,10	5	2,3,5,8,10	5	2,3,5,8,10	10
Minimum number of samples required to split	2,5,10,20,40	10	2,5,10,20,40	5	NA	NA
Minimum number of samples required to be at leaf nodes	2,5,10,20,40	2	2,5,10,20,40	2	NA	NA
Maximum number of leaf nodes	3,5,10	10	3,5,10	10	NA	NA
The number of estimators	1	1	5,10,15,20,40	40	5,10,15,20,40	40
learning rate	NA	NA	NA	NA	0.01,0.03,0.1, 0.3	0.01
Minimum sum of weights of all observations required in a child	NA	NA	NA	NA	1,2,3	2
Pseudoregularization (Gamma)	NA	NA	NA	NA	0,0.1,0.2,0.3	0
L1 regularization (alpha)	NA	NA	NA	NA	0	0
L2 regularization (lambda)	NA	NA	NA	NA	1	1

2.2) Long run machine learning models' performance

After constructing machine learning models using a 10-fold cross-validation approach to tune the hyperparameters and evaluate their performance, the findings in Table 15 reveal that XGBoost delivers the highest accuracy or the lowest RMSE value at 0.2462 on average cross-validation data. Moreover, to assess the consistency of the model's performance on unseen data, XGBoost is applied to the test set, and the resulting RMSE value is 0.1682, which is comparable to the RMSE values from both the training and cross-validation data. These results suggest that the model is not prone to overfitting or underfitting and is capable of making reasonably accurate predictions for new data. Therefore, XGBoost can be considered a reliable model for this particular dataset.

Table 15 Long run machine learning models' performance

	RMSE Train	RMSE CV	RMSE Test
Decision Tree	0.2071	0.3014	
Random Forest	0.1816	0.2613	
XGBoost	0.1395	0.2462	0.1682

2.3) Feature importance of long run machine learning model

After constructing multiple machine learning models and evaluating their performance based on average cross-validation data from a 10-fold cross-validation approach, it was determined that the XGBoost model achieved the highest accuracy or the lowest RMSE value, as illustrated in Table 15. Subsequently, the XGBoost model was applied to the test set, and the resulting RMSE value of 0.1682 was found to be consistent with the values obtained from both the training and cross-validation data, suggesting that the model is reliable

for predicting both seen and unseen data without any overfitting or underfitting. Therefore, the XGBoost model is the most accurate model for predicting the VN-Index based on the given features. Furthermore, the SHAP value obtained from this model was analyzed and visualized in Figure 20 and Figure 21 to identify the feature importance and contribution to the model's performance.

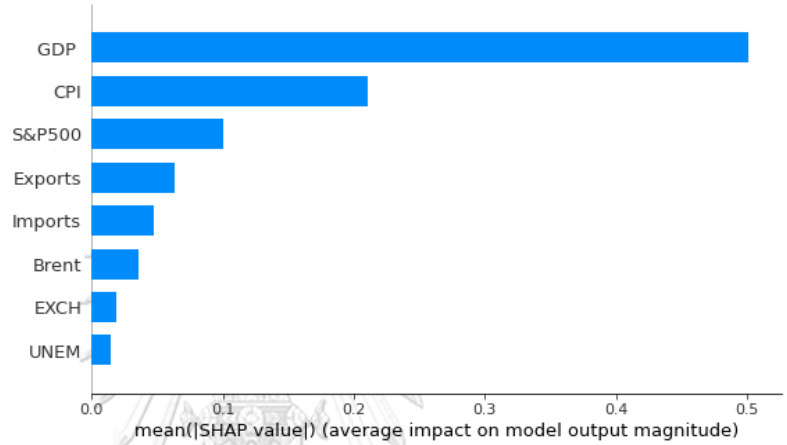


Figure 20 Feature importance from SHAP for long run model

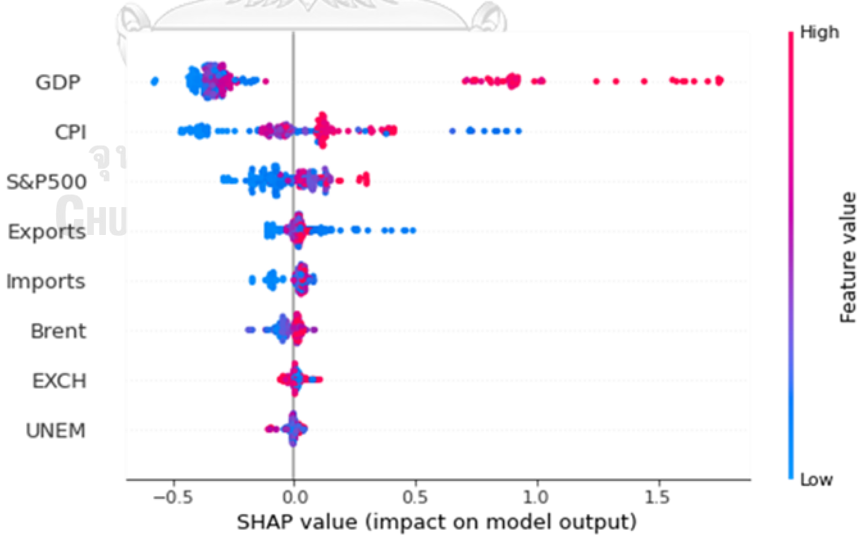


Figure 21 Feature contribution from SHAP for long run model

As XGBoost has been shown to be the most accurate model, the feature importance in Figure 20 was studied. It was found that the GDP feature had the most significant impact on the model, suggesting that it had the greatest influence on the VN-Index compared to other features. Additionally, the consumer price index (CPI), S&P 500 index, exports, imports, and Brent oil price each exhibited a noticeable effect on the model. However, the exchange rate and unemployment rate's impacts appeared to be negligible in comparison to the other six features.

Figure 21 displays the feature contribution from the SHAP value. The findings suggest that higher values of the GDP, S&P 500 index, imports, and Brent oil price have a greater positive impact on the model, as the pink dots representing high values of these features are on the right side of the plot and the blue dots representing low values of these features are on the left side of the plot, signifying a positive influence on the VN Index. Therefore, there appears to be a clear positive relationship between the GDP, S&P 500 index, imports, Brent oil price, and the VN-Index. Furthermore, the consumer price index (CPI) is also likely to have a positive impact on the model even the blue dots representing its low values distribute along the plot but the pink dots representing high values of these features are on the right side of the plot. However, the contributions of the exports, exchange rate, and unemployment rate are still undetermined since their pink dots are located in the middle.

4.4 Model comparison

1. Short run models comparison

1.1) Short run model performance comparison

The short run model performance of both statistical model from OLS method and random forest, the most accurate machine learning model for short run relationship, are shown in Table 16.

Table 16 Short run models' performance

	RMSE Train	RMSE CV	RMSE Test
OLS	0.1427	-	0.1396
Random Forest	0.1328	0.1382	0.1374

Upon analyzing the root mean squared error (RMSE) values of both the test set predictions of the random forest model and the model from ordinary least squares (OLS), it can be inferred that the performance of the random forest model is slightly better. The RMSE values obtained from the test set of the random forest model are slightly lower than those of the OLS method. This indicates that the random forest model has better predictive capabilities, as it is able to more accurately predict the outcomes of the test set. Therefore, it can be concluded that the random forest model may be a more suitable model for the given data set.

1.2) Short run model relationship comparison

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(BRENT)	0.064556	0.051916	1.243462	0.2152
D(CPI)	-1.466529	0.688237	-2.130847	0.0343
D(EXCH)	-0.697362	0.208576	-3.343438	0.0010
D(EXPORTS)	-0.079137	0.089041	-0.888769	0.3752
D(LOG_GDP)	0.012634	0.099496	0.126975	0.8991
D(IMPORTS)	0.222554	1.434599	0.155133	0.8769
D(SP500)	0.732246	0.118866	6.160251	0.0000
D(UNEM)	-0.014158	0.021537	-0.657386	0.5117
C	0.027658	0.018902	1.463229	0.1450

R-squared	0.269126	Mean dependent var	0.014036
Adjusted R-squared	0.239892	S.D. dependent var	0.167367
S.E. of regression	0.145917	Akaike info criterion	-0.969444
Sum squared resid	4.258380	Schwarz criterion	-0.825516
Log likelihood	110.3069	Hannan-Quinn criter.	-0.911253
F-statistic	9.205644	Durbin-Watson stat	1.663507
Prob(F-statistic)	0.000000		

Figure 22 Short run model from OLS

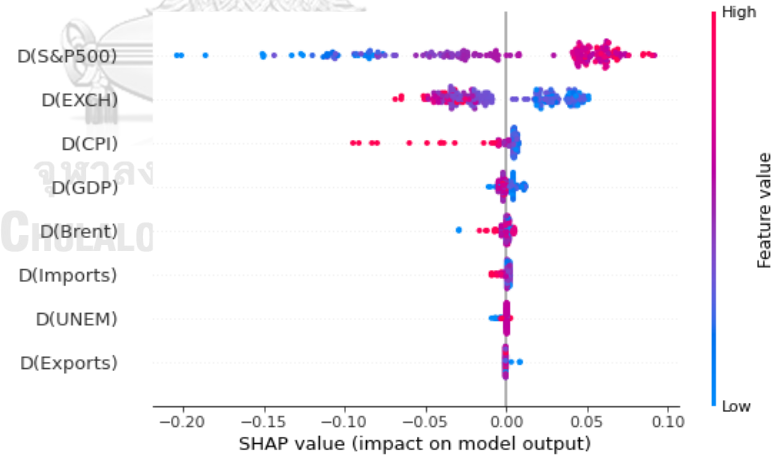


Figure 23 Feature contribution from SHAP for short run model

Table 17 Summary of feature contribution from short run models

Features	OLS	Random Forest
CPI	-	-
S&P500	+	+
Exchange Rate (VND/USD)	-	-

After analyzing the relationship results of both the OLS method and random forest, it was found that the three statistically significant variables, namely the Consumer Price Index (CPI), exchange rate, and S&P500 index, were among the top three factors that had the most impact on the random forest model. However, the rankings of their importance differed between the two models. The standardized coefficients obtained from the OLS method showed that the CPI had the most important impact, followed by the S&P500 index and exchange rate, while the feature importance derived from SHAP values indicated that the S&P500 index had the most impact, followed by the exchange rate and CPI.

Although the contributions of these variables to both models and their relationship with the VN Index in the short run were found to be the same, it is worth noting that the CPI and exchange rate had a negative relationship with the VN Index, whereas the S&P500 index showed a positive relationship in the short run as shown in Table 17.

These findings underscore the importance of utilizing different models and techniques to gain a comprehensive understanding of the relationship between variables and their impact on the VN Index, as different models may yield different rankings of feature importance and highlight varying factors that contribute to the overall performance of the model.

The findings from the analysis suggest a potential relationship between the fluctuations in the S&P500 index and the corresponding movements in the VN Index. Specifically, when there is an increase in the S&P500 index, it can be inferred that the VN Index will also experience an upward trend. This association can be attributed to the significant reliance of Vietnam on exports, with the United States serving as one of its major trading partners. Consequently, any shifts in the US economy have the potential to exert an impact on Vietnam's economic performance, thereby influencing the movements of the VN Index. Additionally, the results indicate that changes in two key economic indicators, namely the Consumer Price Index (CPI) and the exchange rate (VND/USD), can have discernible effects on the behavior of the VN Index. In instances where the CPI, which serves as a proxy for inflation, and the exchange rate (VND/USD) experience significant spikes, the VN Index is likely to exhibit a decline. This can be attributed to the adverse consequences associated with higher inflation, such as a potential decrease in purchasing power and domestic consumption. Moreover, an increase in the exchange rate (VND/USD) implies a decrease in the demand for Vietnamese currency or a decline in foreign investment, both of which can contribute to a decrease in the VN Index.

2. Long run models comparison

2.1) Long run model performance comparison

The long run model performance of both statistical model (ARDL) and XGBoost, the most accurate machine learning model for long run relationship, are shown in Table 18.

Table 18 Long run models' performance

	RMSE train	RMSE CV	RMSE Test
ARDL	0.4598	-	0.5035
XGBoost	0.1395	0.2462	0.1682

Upon analyzing the root mean squared error (RMSE) values of both the test set predictions of the XGBoost model and the ARDL model, it can be inferred that the performance of the XGBoost model is obviously better. The RMSE values obtained from the test set of the XGBoost model are apparently lower than those of the ARDL model. This indicates that the XGBoost model has better predictive capabilities, as it is able to more accurately predict the outcomes of the test set since XGBoost model can capture non-linear relationships between features and the target variable and can handle complex interactions and non-linearities more effectively. Therefore, it can be concluded that the XGBoost model may be a more suitable model for the given data set.

2.2) Long run model relationship comparison

Variable	Coefficient	Std. Error	t-Statistic	Prob.
BRENT	0.033177	0.097071	0.341783	0.7329
CPI	-3.595774	0.936790	-3.838399	0.0002
EXCH	0.388695	0.557786	0.696853	0.4867
EXPORTS	0.235097	0.670991	0.350372	0.7264
IMPORTS	-1.205797	0.788491	-1.529247	0.1278
LOG_GDP	3.242697	0.648523	5.000127	0.0000
SP500	1.860604	0.330973	5.621623	0.0000
UNEM	-0.078605	0.072653	-1.081928	0.2806
C	0.237225	0.128083	1.852116	0.0655

EC = PRICE - (0.0332*BRENT -3.5958*CPI + 0.3887*EXCH + 0.2351 *EXPORTS -1.2058*IMPORTS + 3.2427*LOG_GDP + 1.8606*SP500 -0.0786*UNEM + 0.2372)

Figure 24 Statistical model for long run relationship

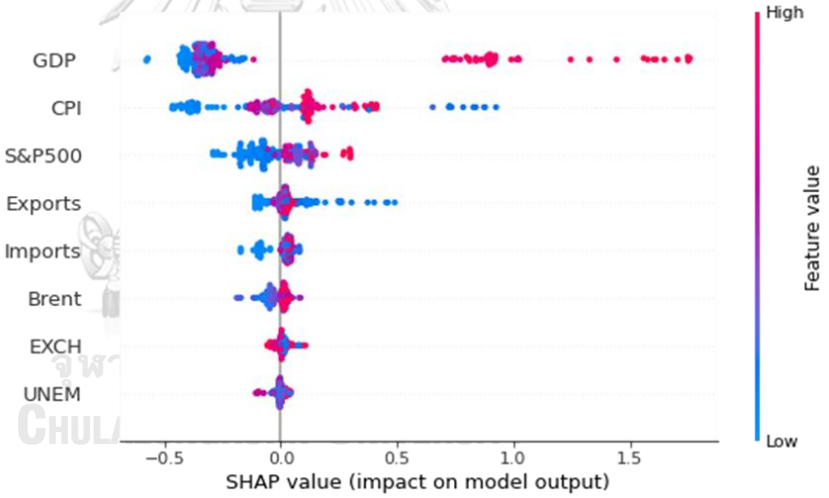


Figure 25 Feature contribution from SHAP for long run model

Table 19 Summary of feature contribution from long run models

Features	ARDL	XGBoost
CPI	-	+
GDP	+	+
S&P500	+	+

Upon analyzing the relationship results of both the ARDL and XGBoost models, it can be inferred that the three statistically significant variables, namely the Consumer Price Index (CPI), GDP, and S&P500 index, had the most impact on the XGBoost model, as they were ranked as the top-three influential variables. However, despite their similar rankings, their importance rankings varied between the two models. The magnitude of standardized coefficients derived from the ARDL model indicated that the Consumer Price Index (CPI) had the most significant impact, followed by GDP and the S&P500 index, whereas the feature importance from SHAP revealed that the GDP had the most impact on the model, followed by the Consumer Price Index (CPI) and the S&P500 index. It's noteworthy that their contributions to both models or their relationship with the VN Index in the long run were not exactly the same. There was a significant difference in the direction of the relationship in case of the consumer price index (CPI). While the Consumer Price Index (CPI) had a negative relationship with the VN Index in ARDL approach, it showed a positive relationship in SHAP value from XGBoost. Additionally, the GDP and S&P500 index showed a positive relationship with the VN Index in the long run from both models as shown in Table 19.

The results suggest that both the S&P500 index and GDP have a positive relationship with the VN Index. This can be attributed to the

significant role played by the US economy in global financial markets. Developments in the US, such as changes in interest rates, monetary policy, or investor sentiment, have the potential to impact global financial conditions. Consequently, these conditions can spill over and influence Vietnam's financial markets. Moreover, Vietnam's GDP heavily relies on exports, with the US serving as a major trading partner. As a result, changes in the US economy can have significant implications for Vietnam's economy.

The distinct contribution of the Consumer Price Index (CPI) can be explained by considering the various causes of inflation. Inflation can arise from different factors, including demand-pull inflation and cost-push inflation. Demand-pull inflation occurs when the aggregate demand surpasses the available supply of goods and services. This leads to price increases as consumers compete for limited resources. Factors contributing to demand-pull inflation encompass robust consumer spending, increased government expenditure, low interest rates encouraging borrowing and spending, and expansionary monetary policies. This type of inflation stems from the growth of purchasing power and can be regarded as a favorable inflationary environment. On the other hand, cost-push inflation arises from an increase in production costs for goods and services. Factors contributing to cost-push inflation include escalating wages, higher costs of raw materials or energy, increased taxes or regulations, and disruptions in the supply chain. When businesses face higher production costs, they often pass on these costs to consumers through higher prices. Consequently, cost-push inflation reduces purchasing power and is commonly perceived as an unfavorable inflationary condition.

Chapter 5

Conclusion

The descriptive statistics and histograms provide valuable insights into the dataset, revealing that the variables are on different scales and exhibit various degrees of skewness. The Pearson correlation coefficients suggest that there are strong linear correlations among all eight features, which could violate key assumptions of linear regression, such as independence, homoscedasticity, and normality. However, the findings also indicate that all features exhibit a positive linear relationship with the VN Index.

The Augmented Dickey-Fuller (ADF) unit root test was conducted to identify whether the variables in the time series data were stationary or non-stationary. The results showed that all variables, except for unemployment rate, were non-stationary. However, most of these variables were transformed into stationary data through first differencing. The integration order of GDP was found to be larger than one, but it became stationary after taking the logarithm and first differencing. As a result of these findings, the logarithm of GDP (Log(GDP)) was used for the statistical analysis instead of GDP and the ARDL bound test will be used to identify co-integration and long-run relationships, while the OLS method will be used to examine short-run relationships.

The OLS method was used to construct a short run model to analyze the statistical significance of the variables. The results indicated that only three variables, namely CPI, exchange rate, and S&P500 index, exhibited statistical significance within the model. These variables had varying impacts on the model depending on their respective signs of coefficients, with the CPI and exchange rate having negative impacts and the S&P500 index having a positive impact. The CPI had the largest impact on the model, followed by the S&P500 index and exchange rate, respectively. The constructed model was applied to a test set and the results were compared to

the training set values. The root mean squared error (RMSE) values in both the training and test sets are relatively close, with values of 0.1427 and 0.1396, respectively. The findings suggested that the model is not overfitting or underfitting and that it is capable of generating accurate predictions even when applied to new and previously unseen data.

The selection of the optimal order of the ARDL model is a crucial step in building an ARDL Bound test, as selecting the wrong order can result in inaccurate results. A rigorous process was used to select the optimal order of the ARDL model, and it was determined that the ARDL(2, 0, 0, 1, 0, 0, 1, 1, 0) model was the best performing. The ARDL Bound test was then conducted, and the results indicated that there is a long-run relationship between the variables under consideration. The long-run equation was derived, revealing that only three variables (CPI, Log GDP, and S&P500 index) were statistically significant to the model, and their coefficients showed that CPI has a negative impact on the model, while Log GDP and S&P500 index have a positive impact. The long-run model's performance was evaluated using the test set, and the root mean squared error (RMSE) values for the training and test sets are slightly different at 0.4598 and 0.5035, respectively. they suggest that the model is not overfitting or underfitting and the predictions are reasonably accurate.

To evaluate short and long run relationship with machine learning models, the Decision tree, Random Forest, and XGBoost are used. For short run models, the models were trained with differenced data using a 10-fold cross-validation approach to tune hyperparameters and evaluate performance. The random forest algorithm achieved the highest accuracy or lowest RMSE value as evidenced by the obtained average value of 0.1382. Furthermore, upon applying the random forest model to the test set, the RMSE value of 0.1374 was obtained, indicating its robustness and lack of overfitting or underfitting issues. The SHAP value method was used to analyze feature importance and the results revealed that the S&P 500 index had the greatest impact on the model, followed by the exchange rate, consumer price index (CPI),

and GDP. The impacts of Brent oil price, imports, unemployment rate, and exports were relatively minor. The SHAP value also showed a clear positive relationship between the S&P 500 index and the VN-Index, while the exchange rate, CPI, and GDP had a negative impact on the model.

For long run models, the models were trained using a 10-fold cross-validation approach to tune hyperparameters and evaluate performance. The results show that the XGBoost model achieved the highest accuracy or the lowest RMSE value on average cross-validation data and is reliable for predicting both seen and unseen data without overfitting or underfitting due to its ability to comprehend the non-linear connections between features and the target variable, as well as its effectiveness in managing intricate interactions and non-linearities. The SHAP value obtained from the XGBoost model was analyzed to identify the feature importance and contribution to the model's performance. The GDP feature had the most significant impact on the model, followed by the CPI, S&P 500 index, exports, imports, and Brent oil price. The feature contribution analysis showed a positive relationship between the GDP, CPI, S&P 500 index, imports, Brent oil price, and the VN-Index, and the contributions of the exports, exchange rate, and unemployment rate are still undetermined.

In comparison between both short run models, the performance of a random forest model and an ordinary least squares (OLS) method were compared using root mean squared error (RMSE) values, and it was found that the random forest model performed slightly better than the OLS method. The Consumer Price Index (CPI), exchange rate, and S&P500 index were identified as the top three variables with significant impact on the random forest model, although their rankings differed from the OLS method. While the contribution of these variables to both models and their relationship with the VN Index were found to be the same, the CPI and exchange rate showed a negative relationship with the VN Index in the short run, whereas the S&P500 index showed a positive relationship.

In comparison between both long run models, the performance of the XGBoost model and the ARDL model were compared using root mean squared error (RMSE) values. The root mean squared error (RMSE) values show that the XGBoost model has better predictive capabilities than the ARDL model. The three statistically significant variables in both models, namely the CPI, GDP, and S&P500 index, have the most impact on the XGBoost model. However, the rankings of their importance differ between the two models. While the CPI has the most significant impact according to the ARDL model, the GDP has the most impact according to the SHAP feature importance. The impact of the three significant variables on both models and their long run relationship with the VN Index were not identical. The relationship between the Consumer Price Index (CPI) and the VN Index differed significantly between the two models. In the ARDL approach, the CPI had a negative relationship with the VN Index, but in the SHAP value from XGBoost, it had a positive relationship. Moreover, the GDP and S&P500 index had a positive relationship with the VN Index in the long run in both models. Based on the performance results, it can be concluded that the XGBoost model is a more suitable model for the given data set.

REFERENCES

Adam, A., & Tweneboah, G. (2008). Macroeconomic Factors and Stock Market Movement: Evidence from Ghana. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1289842

Al-Maadid, A., Alhazbi, S., & Al-Thelaya, K. (2022). Using machine learning to analyze the impact of coronavirus pandemic news on the stock markets in GCC countries. *Res Int Bus Finance*, *61*, 101667. https://doi.org/10.1016/j.ribaf.2022.101667

Bekhet, H. A., & Mugableh, M. I. (2012). Investigating Equilibrium Relationship between Macroeconomic Variables and Malaysian Stock Market Index through Bounds Tests Approach. *International Journal of Economics and Finance*, *4*(10). https://doi.org/10.5539/ijef.v4n10p69

Bhattacharjee, I., & Bhattacharja, P. (2019). *Stock Price Prediction: A Comparative Study between Traditional Statistical Approach and Machine Learning Approach* 2019 4th International Conference on Electrical Information and Communication Technology (EICT),

Boudoukh, J., & Richardson, M. (1993). Stock Returns and Inflation: A Long-Horizon Perspective. *The American Economic Review*, *83*(5), 1346-1355.

DAO, H. T., VU, L. H., PHAM, T. L., & NGUYEN, K. T. (2022). Macro-Economic Factors Affecting the Vietnam Stock Price Index: An Application of the ARDL Model. *Journal of Asian Finance, Economics and Business*, *9*(5), 285-294. https://doi.org/10.13106/jafeb.2022.vol9.no5.0285

Duy, V. Q. (2016). The Impact of Macroeconomic Factors on Stock Price Index, VN-Index. *International Journal of Innovative Science, Engineering & Technology*, *3*(7), 69-84.

Duy, V. Q., & Hau, L. L. (2017). Impact of Macroeconomic Factors on Share Price Index in Vietnam's Stock Market. *The International Journal of Engineering and Science*, *6*(01), 52-59. https://doi.org/10.9790/1813-0601025259

Engle, R. F., & Granger, C. W. J. (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, *55*(2), 251-276.

https://doi.org/10.2307/1913236

Fama, E. F. (1981). Stock Returns, Real Activity, Inflation, and Money. *The American Economic Review*, *71*(4), 545-565.

Granger, C. W. J., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, *2*(2), 111-120. https://doi.org/https://doi.org/10.1016/0304-4076(74)90034-7

Hussainey, K., & Khanh Ngoc, L. (2009). The impact of macroeconomic indicators on Vietnamese stock prices. *The Journal of Risk Finance*, *10*(4), 321-332. https://doi.org/10.1108/15265940910980632

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, *12*(2-3), 231-254. https://EconPapers.repec.org/RePEc:eee:dyncon:v:12:y:1988:i:2-3:p:231-254

Johansen, S., & Juselius, K. (1990). Maximum Likelihood Estimation and Inference on Cointegration--With Applications to the Demand for Money. *Oxford Bulletin of Economics and Statistics*, *52*(2), 169-210. https://EconPapers.repec.org/RePEc:bla:obuest:v:52:y:1990:i:2:p:169-210

Le, T. M. H., Zhihong, J., & Zhu, Z. (2019). Impact of Macroeconomic Variables on Stock Price Index: Evidence from Vietnam Stock Market. *Research Journal of Finance and Accounting*, *10*(12). https://doi.org/10.7176/RJFA

Mai, N. C. (2016). The Influence of Macroeconomic Announcements into Vietnamese Stock Market Volatility. https://doi.org/10.31219/osf.io/ydmhx

Narayan, P., Smyth, R., & Lean, H. H. (2012). Exchange Rate and Stock Price Interaction in Major Asian Markets: Evidence for Individual Countries and Panels Allowing for Structural Breaks. *The Singapore Economic Review (SER)*, *56*, 255-277. https://doi.org/10.1142/S0217590811004250

Nguyen, A. P., Nguyen, H. A., Ho, T. H. M., Ngo, P. T., & McMillan, D. (2019). Risk and returns of different foreign ownership portfolios: Evidence from Vietnam stock market. *Cogent Economics & Finance*, *7*(1). https://doi.org/10.1080/23322039.2019.1589412

Nguyen, T. N., Nguyen, D. T., & Nguyen, V. N. (2020). The Impacts of Oil Price and Exchange Rate on Vietnamese Stock Market. *The Journal of Asian Finance,*

Economics and Business, *7*(8), 143-150. https://doi.org/10.13106/jafeb.2020.vol7.no8.143

Pesaran, H., & Shin, Y. (1999). An autoregressive distributed-lag modelling approach to cointegration analysis. *Econometric Society Monographs*, *31*, 371-413.

Pesaran, M. H., Shin, Y., & Smith, R. J. (2001). Bounds Testing Approaches to the Analysis of Level Relationships. *Journal of Applied Econometrics*, *16*(3), 289-326. http://www.jstor.org/stable/2678547

Phung, & Rhee. (2019). A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets. *Applied Sciences*, *9*, 4500. https://doi.org/10.3390/app9214500

Pražák, T. (2018). The Effect of Economic Factors on Performance of the Stock Market in the Czech Republic. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *66*(6), 1613-1626. https://doi.org/10.11118/actaun201866061613

Rjoub, H., Türsoy, T., & Günsel, N. (2009). The effects of macroeconomic factors on stock returns: Istanbul Stock Market. *Studies in Economics and Finance*, *26*(1), 36-45. https://doi.org/10.1108/10867370910946315

Shrestha, M. B., & Bhatta, G. R. (2018). Selecting appropriate methodological framework for time series data analysis. *The Journal of Finance and Data Science*, *4*(2), 71-89. https://doi.org/10.1016/j.jfds.2017.11.001

Spiess, A.-N., & Neumeyer, N. (2010). An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: A Monte Carlo approach. *BMC pharmacology*, *10*, 6. https://doi.org/10.1186/1471-2210-10-6

Tien, N. H. (2021). The Impact of World Market on Ho Chi Minh City Stock Exchange in Context of Covid-19 Pandemic. *Turkish Journal of Computer and Mathematics Education*, *12*(14), 4252-4264.

Vega García, M., & Aznarte, J. L. (2020). Shapley additive explanations for NO2 forecasting. *Ecological Informatics*, *56*. https://doi.org/10.1016/j.ecoinf.2019.101039

VITA

NAME	Nuttawan Sangsawai
DATE OF BIRTH	April 1997
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Chulalongkorn University
HOME ADDRESS	Bangkok, Thailand.