Shape, rectangle

Description automatically generated **VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**



**FINAL REPORT**

**Business Analysis**

**Instructor:** Nguyễn Đình Thuân

Nguyễn Minh Nhựt

Nguyễn Thị Viết Hương

**Class:** Business Analysis – IS403.M22.HTCL

**Team 1**

Lâm Ngọc Huy - 20521389

Dương Bảo Tâm - 20521865

Lê Thanh Giang - 20521260

Ho Chi Minh City, 18th June 2023

WORK ASSIGNMENT

|  |  |  |  |
| --- | --- | --- | --- |
| MEMBERS  TASK | Lâm Ngọc Huy  (Leader) | Dương Bảo Tâm | Lê Thanh Giang |
| Problem discussion | X | X | X |
| Template Building | X | X | X |
| ARIMA-ARIMAX-Random Forest |  |  | X |
| LN-LSTM-CNN |  | X |  |
| DLM-RNN-NNAR | X |  |  |
| Summarize and Edit Report | X |  |  |
| % Work involved | 34% | 33% | 33% |
| % Assigned work completed | 100% | 100% | 80% |

Contents

[**I.** **Introduction** 3](#_Toc137932066)

[**II.** **Related reasearch** 3](#_Toc137932067)

[**III.** **data and methodology**. 4](#_Toc137932068)

[A. Dataset source 4](#_Toc137932069)

[B. Measurement 4](#_Toc137932070)

[C. Modeling 4](#_Toc137932071)

[1) Linear Regression 4](#_Toc137932072)

[2) Random Forest 6](#_Toc137932073)

[3) ARIMA 8](#_Toc137932074)

[4) ARIMAX 10](#_Toc137932075)

[5) Dynamic Linear Model 11](#_Toc137932076)

[6) Recurrent Neural Network 12](#_Toc137932077)

[7) Long Short Term Memory Neural Network 12](#_Toc137932078)

[8) Neural Network Auto Regressive 14](#_Toc137932079)

[9) Convolutional Neural Network 14](#_Toc137932080)

[D. Result 15](#_Toc137932081)

[1) Linear Regression 15](#_Toc137932082)

[2) Random Forest 15](#_Toc137932083)

[3) ARIMA 16](#_Toc137932084)

[4) ARIMAX 16](#_Toc137932085)

[5) Dynamic Linear Model 16](#_Toc137932086)

[6) Recurrent Neural Network 16](#_Toc137932087)

[7) Long Short Term Meomory Neural Network 17](#_Toc137932088)

[8) Neural Network Auto Regressive 17](#_Toc137932089)

[9) Convolutional Neural Network 17](#_Toc137932090)

[**IV.** **Conclusion** 18](#_Toc137932091)

[1) MBB 18](#_Toc137932092)

[2) VCB 18](#_Toc137932093)

[3) STB 18](#_Toc137932094)

***Abstract***—At the meantime, because of the market economy’s growth, more and more investors want to participate in the stock market. Since the stock market trends are complex and unstable, investors can easily lose their investment. In this study, our team hopes to create a sustained, less error-prone and can be freely interacting tool by using ARIMA, ARIMAX, RF, LN, LSTM, CNN, DLM, NNAR, RNN models to compare the stock price and bring up the conclusion.

*Keywords:* ARIMA, ARIMAX, RF, LR, LSTM, CNN, DLM, NNAR, RNN

# **Introduction**

Stocks are units of ownership in a company, also known as shares of stock or equities. When you buy a share of stock, you’re purchasing a partial ownership stake in a company, entitling you to certain benefits. The first modern stock trading market was created in Amsterdam when the Dutch East India Company was the first publicly traded company. For many years, the only trading activity on the exchange was trading shares of the Dutch East India Company.

The purpose of stock valuation is to determine the true value of a stock at a given point in time, find out the potential of the stock, and make relevant investment decisions.

For businesses, stock valuation is one of the important steps of a joint stock company when it wants to offer shares, raise capital and increase its influence in the market.

For investors, stock valuation helps investors know which stocks are worth buying and have the greatest return potential.

In this study, we use ARIAMA, ARIMAX, RF, LR, LSTM, CNN, DLM, NNAR, RNN models to predict the stock prices from the Vietnamese banks so it can be easily access by the team and also the Vietnamese investors.

# **Related reasearch**

Over the years, numerous time-series prediction methods have been proposed. Our research aims to evaluate the accuracy of nine models in predicting stock prices: ARIMA, ARIMAX, RF, LR, LSTM, CNN, DLM, NNAR, RNN. In term of work, Shahzad Zaheer et al. (2023) [1] used LSTM and Deep Learning Model like CNN, RNN to take the input stock data and forecasts two stock parameters close price and high price for the next day, and result shows which model perform the best. Another case in point is Vaishnavi Gururaj, Shriya V R and Dr. Ashwini K (2019) [2] used Linear Regression and SVM to analyze and examine the patterns of previous close stock prices of The Coca-Cola Company, from January 2017 to 2018. Aside from that, Rajat Patil (2021) [3] used ARIMA, ARIMAX and LSTM. The results of the research show that ARIMAX model has outperformed the ARIMA and LSTM models. In addition, Mehar Vijh et al. (2020) [4] used ANN and RF to predict the next day closing price for five companies belonging to different sectors of opera-tion. Others, such as M K Ho, Hazlina Darman, Sarah Musa (2021) [5] using ARIMA, NNAR and LSTM Models for Stock Price Prediction. And last but not least, Marko Laine (2020) [6] employs Dynamic Linear Models (DLM) for Time Series Analysis that can be used for Stock Price Dataset.

# **data and methodology**.

## Dataset source

Three datasets are about the Bank Stock from 19/05/2018 to 19/05/2023 from [www.investing.com](http://www.investing.com), include: Joint Stock Commercial Bank for Foreign Trade of Vietnam (VCB) [7], Sai Gon Thuong Tin Commercial Joint Stock Bank (STB) [8], Military Commercial Joint Stock Bank (MBB) [9]. The datasets include seven attributes columns which as Date, Price, Open, High, Low, Vol, Change %. In this research, we will choose the two best model to predict the “Price” value of 30 days later (20/05/2023 -> 19/06/2023). We will use MBB as the representative dataset to demonstrate in this research.

Ảnh có chứa văn bản, biểu đồ, hàng, Sơ đồ

Mô tả được tạo tự động

Figure 1. The MBB data

## Measurement

To evaluate the predictive performance of our proposed models, we employ three performance measures: the root mean square error (RMSE), the mean absolute error (MAE) and the mean absolute percentage error (MAPE). RMSE is used as the loss function during model training, MAE calculates the average absolute difference between the predicted values and the actual values, while MAPE serves as a statistical measure to assess the accuracy of predictions. The equations for these measures are as follows [10]:

A picture containing font, text, white, line

Description automatically generated

A picture containing font, text, white, diagram

Description automatically generated

A picture containing font, text, white, diagram

Description automatically generated

Where m is the number of data points, *Xi*is the predicted *ith* value and *Yi*element is the actual*ith*value.

## Modeling

### Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).[11]

Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.[12]

The following is an example of a resulting linear regression equation:

Y = β0 + β1 X1 + …

* Y: The dependent variable.
* X: The independent variable.
* β0: The intercept or the value of Y when other parameters equal zero.
* β0: The coefficient effect on X­­1.

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

You’ll find that linear regression is used in everything from biological, behavioral, environmental and social sciences to business. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, hàng

Mô tả được tạo tự động

Figure 2. Result of Linear Regression with MBB divided by 7-2-1 ratio.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train-Test | RMSE | MAE | MAPE |
| LR | 5-3 | 13067.80 | 12255.78 | 53.44 |
| 6-3 | 9063.30 | 8518.06 | 35.32 |
| 7-2 | 4319.41 | 3580.65 | 15.58 |

Table 1. MBB measure result after using LR.

### Random Forest

Random Forest is an ensemble learning algorithm for classification, regression, and other tasks in machine learning.

It is based on the idea of creating multiple decision trees at random and combining their predictions to make a final decision.

Given the requirement of stock price prediction, using Random Forest regression as an appropriate solution demonstrates sound decision-making.

To create a Random Forest Regression model, we must follow these steps:

Step 1: Building Decision Trees:

* Random Forest consists of multiple decision trees, each trained on a random subset of the training data.
* Decision trees are binary trees that recursively split the data based on selected features and their thresholds.
* Each internal node represents a split based on a feature and threshold, while each leaf node represents a predicted value.

Step 2: Bagging and Random Subspace Method:

* Random Forest employs a technique called "bagging" (bootstrap aggregating) to create multiple training datasets by randomly sampling from the original training data with replacement.
* In addition, for each split in a decision tree, only a random subset of features is considered, which is known as the "random subspace" method or feature bagging.

Step 3: Ensemble of Decision Trees:

* Once the decision trees are constructed, predictions are made by aggregating the predictions of all the trees.
* For regression, the final prediction is typically the average or mean of the predicted values from all the decision trees.

Collection of tree predictors h (x;θk)

For regression, the random forest prediction is the unweighted average over the collection:



As k -> ∞ the Law of Large Numbers ensures

A picture containing font, handwriting, text, calligraphy

Description automatically generated

The quantity on the right is the prediction (or generalization) error for therandom forest, designated PE\*f. The convergence above implies that randomforests do not overfit

Now define the average prediction error for an individual tree h (X;θ) as

A black text on a white background

Description automatically generated with low confidence

Assume that for all θ the tree is unbiased, i.e.,EY = EXh(X;θ). Then

A picture containing text, font, white, graphics

Description automatically generated

Where ̄ρ is the weighted correlation between residuals Y - h(X;θ) andY- (hX;θ’) for independent θ, θ’.

We might choose Random Forest by these advantages:

* Random Forest has many applications in various fields, including finance, medicine, and marketing.
* Reduced risk of overfitting.
* Provides flexibility: Since random forest can handle both regression and classification tasks with a high degree of accuracy.
* Easy to determine feature importance: Random Forest makes it easy to evaluate variable importance, or contribution, to the model.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

Figure 3. Result of Random Forest with MBB divided by 7-2-1 ratio

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train-Test | RMSE | MAE | MAPE |
| RF | 5-3 | 13326.13 | 12497.71 | 53.71 |
| 6-3 | 9246.47 | 8708.28 | 35.62 |
| 7-2 | 4374.18 | 3621.68 | 15.37 |

. Table 2. MBB measure result after using RF.

### ARIMA

#### Definition

The Autoregressive Integrated Moving Average Model, or ARIMA for short is a standard statistical model with:

* AR: Auto regressive: A model that uses the dependent relationship between an observation and some number of lagged observations.
* I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
* MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

It is used in statistics and econometrics to measure events that happen over a period of time. The model is used to understand past data or predict future data in a series. It’s used when a metric is recorded in regular intervals, from fractions of a second to daily, weekly or monthly periods.

#### ARIMA Model Parameters

Each component in ARIMA functions as a parameter with a standard notation.

For ARIMA models, a standard notation would be ARIMA with p, d, and q (corresponding to AR, I ,MA), where non-negative integer values substitute for the parameters to indicate the type of ARIMA model used.

The parameters can be defined as:

* p: the number of lag observations in the model, also known as the lag order.
* d: the number of times the raw observations are differenced; also known as the degree of differencing.
* q: the size of the moving average window, also known as the order of the moving average.

#### Stationary Time Series :

The time series has its statistical properties remain constant across time, which also mean that the mean and the variance are constant over time. This will help to perform an ARIMA model easier.[13]

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

Figure 4. Result of ARIMA with MBB divided by 7-2-1 ratio.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train-Test | RMSE | MAE | MAPE |
| ARIMA | 5-3 | 10821.08 | 9897.82 | 42.33 |
| 6-3 | 13470.40 | 10669.26 | 50.1 |
| 7-2 | 4755.67 | 3747.6 | 18.36 |

. Table 3. MBB measure result after using ARIMA.

Before starting to calculate a ARIMA model we will hence to review all the definition of an ARIMA model but this time with some formula:

**AR(Auto-Regressive)**: the time series is linearly regressed on its own past values

p: the number of past values inclued in the AR model



With:

yt : The series of metrics over time under consideration

yt-1 : The time series stops at t -1 (for example t is today t -1 is yesterday)

c : is a constant

t : parameters

t  : errors

**I (Intergrated)**: if a time series is not stationary, the time series can be differenced to become stationary by computing the differences between consecutive observations

d: the number of time the time series is differenced



With:

: First differencing

yt : The series of metrics over time under consideration

yt-1  : The time series stops at t -1 (for example t is today t -1 is yesterday)

**MA (Moving Average )**: the time series is ‘regressed’ on the past forecast errors

q: the number of past forecast errors included in the MA model



With:

yt : The series of metrics over time under consideration

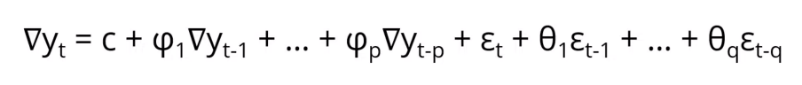
c : is a constant

t  : errors

t : parameters

* Summary the AR and the MA models are usually work on stationary time series , if the time series is not stationary, we can use the I model to make it stationary

**ARIMA** :



With

: difference in time series (can be more than 1)

t  : parameters

t : parameters

c : is a constant

t  : errors

: The difference at time series stops at t -1

Base on the different values of the parameters, ARIMA model can be:

AR model (p,0,0)

MR model(0,0,q)

Or Mix model ARMA(p,0,q)

### ARIMAX

An Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) model can be viewed as a multiple regression model with one or more autoregressive (AR) terms and/or one or more moving average (MA) terms. This method is suitable for forecasting when data is stationary/nonstationary, and multivariate with any type of data pattern, i.e., level/trend /seasonality/cyclicity.

ARIMAX is related to the ARIMA technique but, while ARIMA is suitable for datasets that are univariate.

ARIMAX is suitable for analysis where there are additional explanatory variables (multivariate) in categorical and/or numeric format.

The X in ARIMAX stands for “Exogenous” variable, it is a factor in the model whose value is identified by the factors of variables outside the model study.

The formula of ARIMAX is

A picture containing font, handwriting, calligraphy, typography

Description automatically generated

Pt and Pt-1 represent the values in the current period and 1 period ago respectively.

ϵt and ϵt-1 are the error terms for the same two periods.

c is a baseline constant factor.

ϕ1 and θ1, express what parts of the value Pt-1 and error ϵt-1 last period are relevant in estimating the current one.

X is the exogenous variable

ARIMA model but this time with some formula:

**AR(Auto-Regressive)**: the time series is linearly regressed on its own past values

p: the number of past values inclued in the AR model



With:

yt : The series of metrics over time under consideration

yt-1 : The time series stops at t -1 (for example t is today t -1 is yesterday)

c : is a constant

t : parameters

t  : errors

**I (Intergrated)**: if a time series is not stationary, the time series can be differenced to become stationary by computing the differences between consecutive observations

d: the number of time the time series is differenced



With:

: First differencing

yt : The series of metrics over time under consideration

yt-1  : The time series stops at t -1 (for example t is today t -1 is yesterday)

**MA (Moving Average )**: the time series is ‘regressed’ on the past forecast errors

q: the number of past forecast errors included in the MA model



With:

yt : The series of metrics over time under consideration

c : is a constant

t  : errors

t : parameters

* Summary the AR and the MA models are usually work on stationary time series , if the time series is not stationary, we can use the I model to make it stationary

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

Figure 5. Result of ARIMAX with MBB divided by 7-3 ratio.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train-Test | RMSE | MAE | MAPE |
| ARIMAX | 5-5 | 331.29 | 214.72 | 1.04 |
| 6-4 | 1681.76 | 1303.7 | 6.77 |
| 7-3 | 1436.18 | 1197.2 | 6.3 |

. Table 4. MBB measure result after using ARIMAX.

### Dynamic Linear Model[14]

Statistical analysis of time series data is usually faced with the fact that we have only one realization of a process whose properties we might not fully understand. In analysis of correlation structures, for example, we need to assume that some distributional properties of the process generating the variability stay unchanged in time. In linear trend analysis, we assume that there is an underlying change in the background that stays approximately constant over time.

Dynamic regression with state space approach tries to avoid some of the problems. By explicitly allowing for variability in the regression coefficients we let the system properties change in time. Furthermore, the use of unobservable state variables allows direct modelling of the processes that are driving the observed variability, such as seasonality or external forcing, and we can explicitly allow for some modelling error.

General dynamic linear model can be written with a help of observation equation and model equation as

Yt = Ftxt + vt, vt∼N(0,Vt)

Xt = Gtxt – 1 + wt,wt∼N(0,Wt).

Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, biểu đồ

Mô tả được tạo tự động

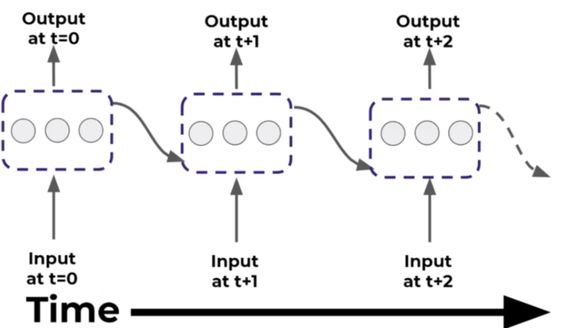
Figure 6. Result of Dynamic Linear Model with MBB divided by 7-2-1 ratio.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train-Test | RMSE | MAE | MAPE |
| ARIMAX | 5-3 | 1186.90 | 1065.69 | 4.60 |
| 6-3 | 1419.32 | 1030.61 | 4.96 |
| 7-2 | 1898.03 | 1466.01 | 7.26 |

*Table 5. MBB measure result after using DLM*

### Recurrent Neural Network[15]

A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data. Like feedforward and convolutional neural networks (CNNs), recurrent neural networks utilize training data to learn. They are distinguished by their “memory” as they take information from prior inputs to influence the current input and output. While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural networks depend on the prior elements within the sequence. While future events would also be helpful in determining the output of a given sequence, unidirectional recurrent neural networks cannot account for these events in their predictions.



The problem of recurrent neural network is the more we unroll, the harder it is to train.

To train the network, we have to modify weights and biases to keep the loss value at lowest at possible.

Let take a look on weight W2.

Ảnh có chứa văn bản, biểu đồ, thiết kế, ảnh chụp màn hình

Mô tả được tạo tự động

If we unrolled the recurrent network 4 times, we must multiply the input value by W2 raise to the number of times we unrolled.

If we unroll 50 times, this can cause a problem called Exploding Gradient

Ảnh có chứa biểu đồ

Mô tả được tạo tự động

One way to prevent the exploding gradient would be limit W2 lower then value 1.

How ever, this results in the vanishing gradient.

Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, biểu đồ

Mô tả được tạo tự động

Figure 7. Result of RNN with MBB divided by 7-2-1 ratio.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train-Test | RMSE | MAE | MAPE |
| RNN | 5-3 | 4823.42 | 3680.6 | 20.5 |
| 6-3 | 3709.37 | 2894 | 12.84 |
| 7-2 | 4217.38 | 3373.69 | 14.73 |

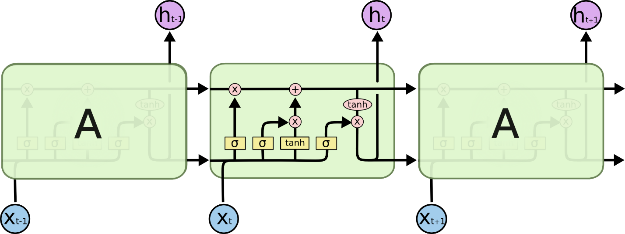
. Table 6. MBB measure result after using RNN.

### Long Short Term Memory Neural Network

LSTM stands for long short-term memory networks, used in the field of Deep Learning. It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems. LSTM has feedback connections, i.e., it is capable of processing the entire sequence of data, apart from single data points such as images. This finds application in speech recognition, machine translation, etc. LSTM is a special kind of RNN, which shows outstanding performance on a large variety of problems.

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

Instead of having a single neural network layer, LSTMs have four of them, interacting in a very special way.



Diagram

Description automatically generated

In the above diagram, each line carries an entire vector, from the output of one node to the inputs of others. The pink circles represent pointwise operations, like vector addition, while the yellow boxes are learned neural network layers. Lines merging denote concatenation, while a line forking denote its content being copied and the copies going to different locations.

**Step-by-Step LSTM Walk Through**

Before presenting the equations that describe the internal mechanisms of an LSTM cell, we will establish a set of conventions for the symbols used as follows:

* is the input vector at each time step *t.*
* , , , are the weight matrices in the LSTM cell.
* , , , are the bias vectors.
* , , are the activation values for the forget gate, input gate, and output gate, respectively.
* , represent the cell internal state and candidate value vectors, respectively.
* is the output value of the LSTM cell.

A picture containing sketch, screenshot

Description automatically generated

A picture containing rectangle, sketch, jack

Description automatically generated

A picture containing screenshot, sketch, line

Description automatically generated

A picture containing jack

Description automatically generated

Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, biểu đồ

Mô tả được tạo tự động

Figure 8. Result of LSTM with MBB divided by 7-2-1 ratio.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train-Test | RMSE | MAE | MAPE |
| LSTM | 5-3 | 7206.96 | 5664.58 | 31.53 |
| 6-3 | 4043.66 | 3178.95 | 14.39 |
| 7-2 | 4081.6 | 3267.48 | 14.27 |

. Table 7. MBB measure result after using LSTM.

### Neural Network Auto Regressive[16]

With time series data, lagged values of the time series can be used as inputs to a neural network. We call this a neural network autoregression or NNAR model.

Consider feed-forward networks with one hidden layer, and we use the notation NNAR(p,k) to indicate there are p lagged inputs and k nodes in the hidden layer.

With seasonal data, it is useful to also add the last observed values from the same season as inputs.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

Figure 9.. Result of NNAR with MBB divided by 7-2-1 ratio.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train-Test | RMSE | MAE | MAPE |
| NNAR | 5-3 | 7172.52 | 5780.46 | 30.07 |
| 6-3 | 3981.72 | 3084.93 | 13.96 |
| 7-2 | 4481.11 | 3580.95 | 16.3 |

. Table 8. MBB measure result after using NNAR.

### Convolutional Neural Network

A convolutional neural network (CNN) is essentially a neural network that employs the convolution operation (instead of a fully connected layer) as one of its layers. CNNs are an incredibly successful technology that has been applied to problems where in the input data on which predictions are to be made has a known grid-like topology, like a time series (a 1-D grid) or an image (a 2-D grid). CNNs ushered deep learning into modern times, solving one of the most crucial computational problems in the digital era of computer vision. With the popularity of CNNs, a surge in the research for deep learning was witnessed that continues today. [17]With the popularity of CNNs, a surge in the research for deep learning was witnessed that continues today.[17]

CNN is renowned for its capability in processing image data, leveraging its strength in detecting spatial patterns and extracting relevant features from visual input., if we are able to convert the 1D time-series sequence to an input image matrix shape, we could apply a CNN model for the forecasting problem.

A CNN is made up of three primary layers a convolution layer, a pooling layer, and

a fully connected layer, as shown in Figure 10. The convolution layer makes an effort to retrieve the best features from the 1-D matrix and perform calculation to provide a convoluted output, as shown in equation below. [1]

CLt = tanh(xt ∗ wt + bt) (1)

Where CLt is the convolution output, activation function is tanh, xt is input value, wt is the weight, and bt is the bias.

A diagram of convolution structure

Description automatically generated with medium confidence

Figure 10. CNN one-dimensional structure.

A The pooling layer takes the output of the convolutions as an input. The max pooling function is used to choose the heavily weighted features in the pooling layer. The pooling layer’s output is passed to the flatten layer. The flatten layer’s primary function is to convert the data into a single array form. The fully connected layer receives the flatten layer’s output and processes it to obtain the results. [1]

Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, biểu đồ

Mô tả được tạo tự động

Figure 11. Result of CNN with MBB divided by 7-2-1 ratio.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train-Test | RMSE | MAE | MAPE |
| CNN | 5-3 | 5902.99 | 4599.95 | 23.99 |
| 6-3 | 3795.53 | 2933.31 | 13.13 |
| 7-2 | 4439.05 | 3551.37 | 16.16 |

. Table 9. MBB measure result after using CNN.

## Result

After building the models on our datasets, the following tables are the evaluation of error predicted on the test set.

### Linear Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Train-Test | RMSE | MAE | MAPE |
| MBB | 5-3 | 13067.80 | 12255.78 | 53.44 |
| 6-3 | 9063.30 | 8518.06 | 35.32 |
| 7-2 | 4319.41 | 3580.65 | 15.58 |
| VCB | 5-3 | 4954.80 | 3911.49 | 4.90 |
| 6-3 | 11052.26 | 8484.90 | 11.19 |
| 7-2 | 9732.89 | 9403.91 | 12.50 |
| STB | 5-3 | 12928.16 | 15442.26 | 56.6 |
| 6-3 | 4983.06 | 9498.37 | 33.35 |
| 7-2 | 5149.15 | 5961.45 | 24.58 |

### Random Forest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Train-Test | RMSE | MAE | MAPE |
| MBB | 5-3 | 13326.13 | 12497.71 | 53.71 |
| 6-3 | 9246.47 | 8708.28 | 35.62 |
| 7-2 | 4374.18 | 3621.68 | 15.37 |
| VCB | 5-3 | 5024.71 | 3939.6 | 4.88 |
| 6-3 | 5677.04 | 4936.91 | 6.15 |
| 7-2 | 11286.29 | 9269.67 | 12.14 |
| STB | 5-3 | 16799.05 | 15749.06 | 56.7 |
| 6-3 | 10845.38 | 9659.88 | 33.24 |
| 7-2 | 6843.76 | 6038.2 | 24.14 |

### ARIMA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Train-Test | RMSE | MAE | MAPE |
| MBB | 5-3 | 10821.08 | 9897.82 | 42.33 |
| 6-3 | 13470.40 | 10669.26 | 50.1 |
| 7-2 | 4755.67 | 3747.6 | 18.36 |
| VCB | 5-3 | 12073.24 | 1141.45 | 13.92 |
| 6-3 | 6036.9 | 4460.24 | 5.49 |
| 7-2 | 6567.22 | 5162.3 | 6.38 |
| STB | 5-3 | 13175.94 | 11916.19 | 42.49 |
| 6-3 | 29287.9 | 24407.82 | 106.28 |
| 7-2 | 6024.07 | 5132.46 | 23.15 |

### ARIMAX

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Train-Test | RMSE | MAE | MAPE |
| MBB | 5-3 | 331.29 | 214.72 | 1.04 |
| 6-3 | 1681.76 | 1303.69 | 6.77 |
| 7-2 | 1436.18 | 1197.20 | 6.28 |
| VCB | 5-3 | 822.63 | 710.78 | 0.90 |
| 6-3 | 925.42 | 701.52 | 0.87 |
| 7-2 | 5328.58 | 4236.14 | 4.94 |
| STB | 5-3 | 367.73 | 412.47 | 1.66 |
| 6-3 | 2865.86 | 2338.81 | 10.43 |
| 7-2 | 1616.16 | 1304.14 | 5.89 |

### Dynamic Linear Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Train-Test | RMSE | MAE | MAPE |
| MBB | 5-3 | 1186.90 | 1065.69 | 4.60 |
| 6-3 | 1419.32 | 1030.61 | 4.96 |
| 7-2 | 1898.03 | 1466.01 | 7.26 |
| VCB | 5-3 | 2821.38 | 2302.79 | 2.93 |
| 6-3 | 4383.01 | 3523.51 | 4.63 |
| 7-2 | 3122.34 | 2462.00 | 3.26 |
| STB | 5-3 | 1706.42 | 1490.17 | 5.33 |
| 6-3 | 2984.13 | 2309.78 | 10.45 |
| 7-2 | 2676.71 | 2159.65 | 10.20 |

### Recurrent Neural Network

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Train-Test | RMSE | MAE | MAPE |
| MBB | 5-3 | 4823.41 | 3680.58 | 20.49 |
| 6-3 | 3709.36 | 2894.00 | 12.83 |
| 7-2 | 4217.37 | 3373.69 | 14.73 |
| VCB | 5-3 | 6030.63 | 4618.59 | 5.7 |
| 6-3 | 7275.39 | 5716.60 | 7.32 |
| 7-2 | 7455.52 | 5848.45 | 7.37 |
| STB | 5-3 | 6051.03 | 5022.78 | 21.78 |
| 6-3 | 7318.63 | 5929.90 | 24.01 |
| 7-2 | 7483.63 | 6076.02 | 22.83 |

### Long Short Term Meomory Neural Network

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Train-Test | RMSE | MAE | MAPE |
| MBB | 5-3 | 7206.96 | 5664.58 | 31.53 |
| 6-3 | 4043.66 | 3178.95 | 14.39 |
| 7-2 | 4081.6 | 3267.48 | 14.27 |
| VCB | 5-3 | 6534.36 | 5024.84 | 6.35 |
| 6-3 | 7175.87 | 5617.94 | 7.13 |
| 7-2 | 7293.76 | 5797.28 | 7.31 |
| STB | 5-3 | 11066.25 | 9010.22 | 41.38 |
| 6-3 | 8504.94 | 6912.37 | 28.38 |
| 7-2 | 7256.78 | 5905.52 | 22.45 |

### Neural Network Auto Regressive

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Train-Test | RMSE | MAE | MAPE |
| MBB | 5-3 | 7172.52 | 5780.46 | 30.07 |
| 6-3 | 3981.72 | 3084.93 | 13.96 |
| 7-2 | 4481.11 | 3580.95 | 16.3 |
| VCB | 5-3 | 7860.2 | 5940.98 | 7.59 |
| 6-3 | 7515.31 | 5834.64 | 7.46 |
| 7-2 | 7981.22 | 6289.61 | 8.05 |
| STB | 5-3 | 9286.3 | 7539.32 | 32.79 |
| 6-3 | 7141.51 | 5726.62 | 23.46 |
| 7-2 | 8019.53 | 6515.72 | 27.07 |

### Convolutional Neural Network

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Train-Test | RMSE | MAE | MAPE |
| MBB | 5902.99 | 5902.99 | 4599.95 | 23.99 |
| 3795.53 | 3795.53 | 2933.31 | 13.13 |
| 4439.05 | 4439.05 | 3551.37 | 16.16 |
| VCB | 5-3 | 6371.98 | 4868.13 | 6.13 |
| 6-3 | 7234.11 | 5650.2 | 7.2 |
| 7-2 | 7875.02 | 6239.02 | 7.98 |
| STB | 5-3 | 7862.8 | 6288.43 | 26.99 |
| 6-3 | 6984.17 | 5609.67 | 22.76 |
| 7-2 | 7943.88 | 6452.86 | 26.73 |

# **Conclusion**

Based on the above evaluation results, we can see that for each dataset there will be its own model and appropriate grave division ratio. For the best predictive performance, we predict the "close" price of each data set over the next 30 days based on the selected models and ratios. The selected models will have the MAPE value closest to 10%.

### MBB

For the MBB dataset, the model selected was Recurrent Neural Network (6-3) with a MAPE score of 12,83%.

Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, biểu đồ

Mô tả được tạo tự động

### VCB

For the VCB dataset, the chosen model is Linear Regression (6-3) with a MAPE score of 11.19%.

Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, biểu đồ

Mô tả được tạo tự động

### STB

For the STB dataset, the chosen model is Linear Regression (7-2) with a MAPE score of 10,2%

Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, biểu đồ

Mô tả được tạo tự động

##### Acknowledgment

First and foremost, we would like to express our deepest gratitude to Assoc.Prof .Dr.Nguyễn Đình Thuân. I appreciate your lecturing and your enthusiasm, which provide the team the acknowledge to accomplish the project. We sincerely give our thank Mr. Nguyễn Minh Nhựt, who supports our team throughout the course.

I would like to take this opportunity to express our appreciate to Assoc.Prof .Dr.Nguyễn Đình Thuân for permitting us to carry out our project.

Last but not least,we deeply express our gratefulness to Assoc.Prof .Dr.Nguyễn Đình Thuân and Mr. Nguyễn Minh Nhựt for your tireless efforts in guiding the team throughout the project and encouraging our team to keep moving forward.

##### References

[1] S. Zaheer *et al.*, “A Multi Parameter Forecasting for Stock Time Series Data Using LSTM and Deep Learning Model,” *Mathematics*, vol. 11, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/math11030590.

[2] V. Gururaj, “Stock Market Prediction using Linear Regression and Support Vector Machines,” vol. 14, no. 8, 2019.

[3] R. Patil, “Time Series Analysis and Stock Price Forecasting using Machine Learning Techniques,” p. 19, Feb. 2021, doi: 10.1994/Rajat/AI.

[4] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, “Stock Closing Price Prediction using Machine Learning Techniques,” *Procedia Computer Science*, vol. 167, pp. 599–606, Jan. 2020, doi: 10.1016/j.procs.2020.03.326.

[5] M. K. Ho, H. Darman, and S. Musa, “Stock Price Prediction Using ARIMA, Neural Network and LSTM Models,” *J. Phys.: Conf. Ser.*, vol. 1988, no. 1, p. 012041, Jul. 2021, doi: 10.1088/1742-6596/1988/1/012041.

[6] M. Laine, “Introduction to Dynamic Linear Models for Time Series Analysis,” 2020, pp. 139–156. doi: 10.1007/978-3-030-21718-1\_4.

[7] “Joint Stock Commercial Bank for Foreign Trade of Vietnam Stock Price Today | HM VCB Live Ticker - Investing.com.” https://www.investing.com/equities/joint-stock-commercial-bank (accessed Jun. 17, 2023).

[8] “Sai Gon Thuong Tin Commercial Joint Stock Bank Stock Price Today | HM STB Live Ticker - Investing.com.” https://www.investing.com/equities/sai-gon-thuong-tin-commercial (accessed Jun. 17, 2023).

[9] “Military Commercial Joint Stock Bank Stock Price Today | HM MBB Live Ticker - Investing.com.” https://www.investing.com/equities/military-commercial-bank (accessed Jun. 17, 2023).

[10] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” *PeerJ Comput. Sci.*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[11] “About Linear Regression | IBM.” https://www.ibm.com/topics/linear-regression (accessed Jun. 16, 2023).

[12] “How Linear regression algorithm works—ArcGIS Pro | Documentation.” https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-linear-regression-works.htm (accessed Jun. 16, 2023).

[13] “Autoregressive Integrated Moving Average (ARIMA),” *Corporate Finance Institute*. https://corporatefinanceinstitute.com/resources/data-science/autoregressive-integrated-moving-average-arima/ (accessed Jun. 16, 2023).

[14] M. Laine, N. Latva-Pukkila, and E. Kyrölä, “Analysing time-varying trends in stratospheric ozone time series using the state space approach,” *Atmos. Chem. Phys.*, vol. 14, no. 18, pp. 9707–9725, Sep. 2014, doi: 10.5194/acp-14-9707-2014.

[15] “Recurrent neural network,” *Wikipedia*. Apr. 07, 2023. Accessed: Jun. 17, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Recurrent\_neural\_network&oldid=1148729712

[16] *12.4 Neural network models | Forecasting: Principles and Practice (3rd ed)*. Accessed: Jun. 17, 2023. [Online]. Available: https://otexts.com/fpp3/nnetar.html

[17] N. Ketkar and J. Moolayil, “Convolutional Neural Networks,” in *Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch*, N. Ketkar and J. Moolayil, Eds., Berkeley, CA: Apress, 2021, pp. 197–242. doi: 10.1007/978-1-4842-5364-9\_6.