VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

ASSIGNMENT REPORT

CO1007 - GROUP 5

DISCRETE STRUCTURES FOR COMPUTING

Instructor: PROF. Tran Tuan Anh

HO CHI MINH CITY, DECEMBER 2022

Member list & Contribution

No.	Full name	Student ID	Percentage of work
1	Nguyen Van Thanh Dat	2152055	20%
2	Nguyen Khanh Nam	2153599	20%
3	Le Van Phuc	2152241	20%
4	Nguyen Quang Thien	2152994	20%
5	Tran Minh Tuan	2152336	20%

Contents

1 Introduction

In these days and ages, it is immediately evident that the world has been through the darkest period of time due to being affected by COVID-19, leading to the large-scale destruction in innumerable aspects of life. Every country has been implementing several policies and attempts to recover those, especially the economic growth. Nevertheless, this seems to be long and takes a lot of efforts whereas countless countries are currently facing the financial crisis, and state banks keep rising interest rates. These would be definitely the most fundamental and principal factors for the adverse effects on the stock market, especially in Vietnam.

Stock market is a term referring to numerous exchanges in which shares of publicly held companies are bought and sold. These financial activities are carried out through formal exchanges and via over-the-counter (OTC) marketplaces that operate under a certain set of regulations. The stock market paves the way for buyers and sellers of securities to meet, interact, and transact. The markets permit for price discovery for shares of corporations and serve as a barometer for the overall economy. Buyers and sellers are assured of a fair price, high degree of liquidity, and transparency as market participants compete in the open market. In addition to this, when the stock market is being up or down, it means one of the major market indexes. A market index tracks a group of stocks' performance, which both represents the market as a whole or a specific sector of the market such as technology, retail companies. Investors use indexes to benchmark the performance of their own portfolios and, in several cases, to inform their stock trading decisions. Besides, *stock exchange* is also a term that could be used interchangeably with stock market. Traders in the stock market buy or sell shares on one or more of the stock exchanges that are part of the overall stock market.

Figure 1: Stock market

The year 1773 saw the appearance of the first market, the London Stock Exchange, and it began in a coffee house, where traders met to exchange shares. The first stock exchange in the United States appeared in Philadelphia in 1790. The Buttonwood Agreement, so named since it was signed under a buttonwood tree, marked the milestone of the beginning of New York's Wall Street in 1792. The agreement was signed by 24 traders and was the first American organization of its kind to trade in securities. The traders renamed their venture the New York Stock and Exchange Board in 1817. A stock market is a regulated and controlled environment. In the United States, the main regulators include the Securities and

Exchange Commission (SEC) and the Financial Industry Regulatory Authority (FINRA). The earliest stock markets issued and dealt in paper-based physical share certificates. At the present, stock markets operate electronically.

Figure 2: The London Stock Exchange (LSE), one of the oldest stock exchanges in the world

About its operation, it is instinctively known that stock markets provide a secure and regulated environment where market participants can transact in shares and other eligible financial instruments with confidence, with zero to low operational risk. Operating under the defined rules as stated by the regulator, the stock markets act as primary markets and secondary markets. As a primary market, the stock market allows companies to issue and sell their shares to the public for the first time through the process of an initial public offering (IPO). This activity helps companies raise necessary capital from investors. A company divides itself into several shares and sells some of those shares to the public at a price per share. To facilitate this process, a company needs a marketplace where these shares can be sold and this is achieved by the stock market. A listed company may also offer new, additional shares through other offerings at a later stage, such as through rights issues or follow-on offerings. They may even buy back or delist their shares. Investors will own company shares in the expectation that share value will rise or that they will receive dividend payments or both. The stock exchange acts as a facilitator for this capital-raising process and receives a fee for its services from the company and its financial partners. Using the stock exchanges, investors can also buy and sell securities they already own in what is called the secondary market.

With the thorough understanding and knowledge associated with stock markets as well as technology, our team fully recognize how urgent the phenomenon mentioned initially is; therefore, we have conducted this research so named *Vietnamese Stock Market Analysis* so as to analyze scientifically the movement of the stock markets nationally by collecting all the data of Vietnam's stock VNINDEX during the past one year and that of various sectors. What is more, we will build models of prediction for the next three months, carry out analysis to opt for the best option that is closet to the actuality, and then forecast the future movements. This assignment report will perform in depth how we processed from collecting data to building the models.

2 Tools and environment

With the accumulated data about Vietnam's stock indexes, we will utilize Python, Matlab and Google Colaboratory. Several important programs, such as Matplotlib, Seaborn, NumPy, SkLearn, and Pandas that handle certain tasks to assist us in the study are built on top of Python.

2.1 Python

Python is a helpful and straightforward programming language that allows developers to write programs with fewer lines than some other programming languages with its simple syntax and can be treated in a procedural way, an object-oriented way or a functional way. Python runs on an interpreter system, meaning that code can be executed as soon as it is written. On top of that, this kind of language also comes with a mixed diversity of tools for developing models and analyzing data. The tools are referred known as "packages" which pave the way for us to analyze, visualize and manipulate sets of data in this project.

Figure 3: Python programming language

2.1.1 Pandas

Pandas is built on top of the Python programming language that is widely renowned as a fast, powerful, flexible and easy to use open source data analysis and manipulation tool. Due to the fact that we are likely to go through a complex and lengthy process of handling the data accumulated, Pandas would be definitely an ideal and beneficial solution.

2.1.2 Matplotlib

Matplotlib is a comprehensive library for serving as the creation of static, animated, and interactive visualizations in Python. This means that we can utilize it depicting graphs based on the collected data in order to manipulate and analyze it in a more effective way and effortlessly undertake towards the success of the predictive models.

2.1.3 Seaborn

Seaborn is a library for making statistical graphics in Python. This tool is built on top of Matplotlib and integrated closely with Pandas data structures. Its plotting functions operate on dataframes and arrays holding the whole sets of data and internally perform the fundamental semantic mapping and statistical aggregation to make informative plots. Its dataset-oriented, declarative API lets us concentrate on what the different elements of our plots mean, rather than on the details of how to draw them.

2.1.4 Numpy

NumPy, standing for *Numerical Python*, is a Python library used for working with mathematical problems, such as linear algebra, fourier transform, matrices. This tool would pave the way for us to facilitate an abundance of issues related to maths when we build the models of prediction.

2.1.5 SkLearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It is built on NumPy, SciPy, and Matplotlib with simple and efficient tools for predictive data analysis. What is more, this library provides a wide range of valuable tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

2.2 Matlab

Matlab has an extensive set of built-in functions as well as additional toolboxes that consist of functions related to more specialized topics like fuzzy logic, neural networks, signal processing, and so on. It can be used in two different ways: as a traditional programming environment and as an interactive calculator. In calculator mode, the built-in and toolbox functions provide a convenient means of performing one-off calculations and graphical plotting; in programming mode, it provides a programming environment (editor, debugger, and profiler) that enables the users to write their own functions and scripts. In Matlab, everything that can be done using the *GUI* interface (e.g., plotting) can also be accomplished using a command-line equivalent.

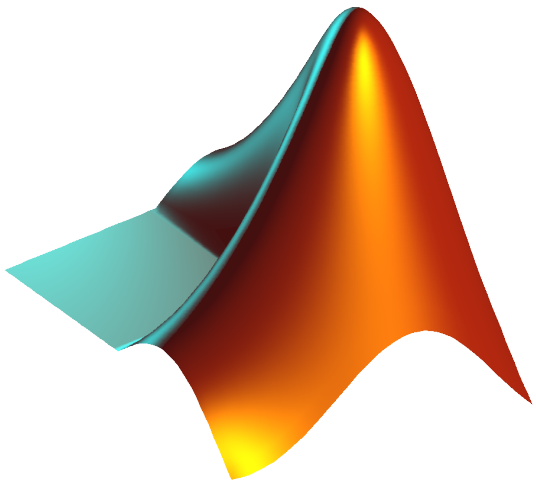

Figure 4: Matlab logo

Based on the renowned functions mentioned of Matlab, it is of fundamental importance to assisting us in finding the best distribution that fits our gathered data via *GUI*. The picture below shows that we can select whether we desire to fit continuous of discrete distributions, and whether we desire to display the

PDFs of the CDFs (with the pop-down menus). The lists of optional distributions are displayed on the left side of the *GUI*. It also provides us with the 4 best fits on the graph, and detailed parameters of the best 4 distributions under the graph, namely Distribution Name, NLogL - *Negative of the log likelihood*, BIC - Bayesian information criterion, AIC - *Akaike information criterion,* AICc - *AIC with a correction for finite sample sizes*, Parameters names, Parameters values.

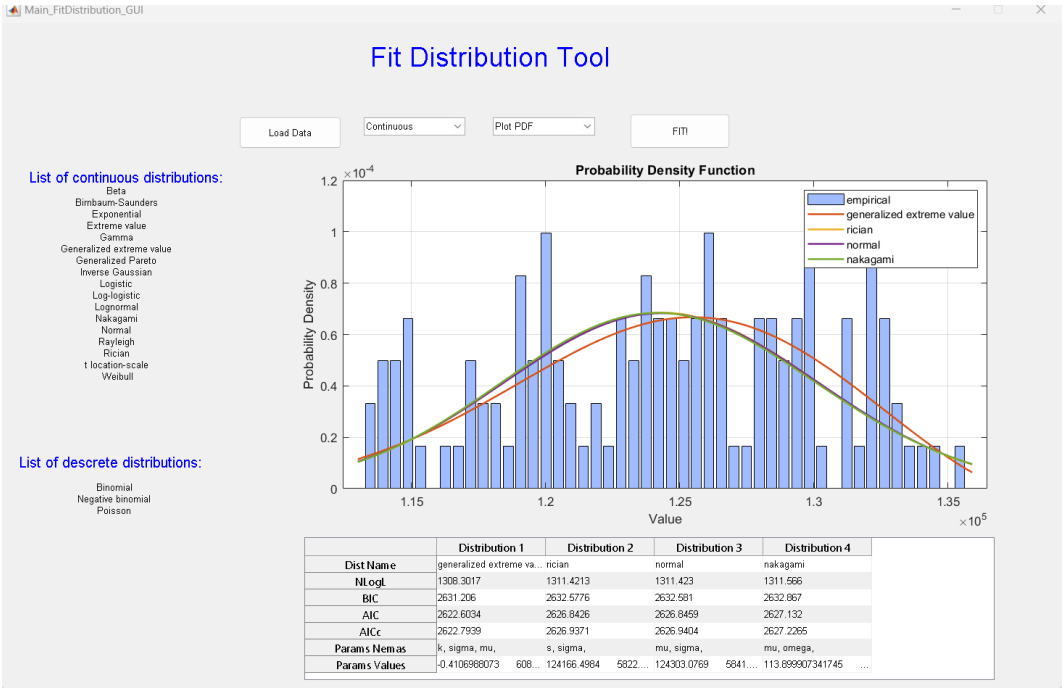

Figure 5: Probability Density Function

2.3 Google Colaboratory

Colaboratory, or "Colab" for short, is a product from Google Research. Colab assists users in writing and executing arbitrary python code through the browser, and is eminently suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, whereas offering access free of charge to computing resources including GPUs; thus, we have decided to use this supportive environment so as to associate, perform live code and run it altogether.

Figure 6: Google Colaboratory

3 Basic concepts of Machine Learning and Deep Learning

3.1 Introduction to Machine Learning

Machine Learning is a subfield of Artificial Intelligence *(AI)* that allows computers to train on data inputs and use statistical analysis so as to output values that fall within a specific range, leading to facilitating computers in building models from sample data in order to automate decision-making processes based on training data.

AI is the broader concept – machines making decisions, learning new skills, and solving problems in a similar way to humans – whereas Machine Learning is a branch of AI that enables intelligent systems to autonomously learn new things from data. An analogous pattern could be seen in Deep Learning, which is also a subset of Machine Learning that is going to be introduce later in this assignment report.

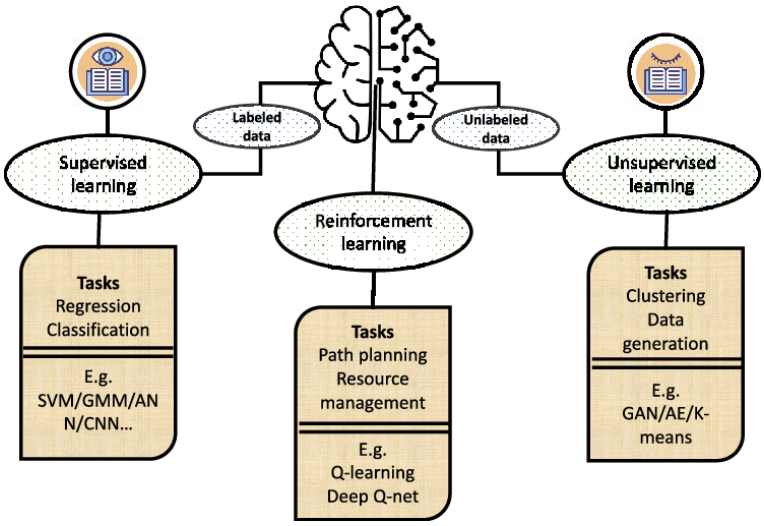

Figure 7: Machine Learning

Two of the most ubiquitously adopted Machine Learning methods are *supervised learning* which trains algorithms based on example input and output data that is labeled by humans, and *unsupervised learning* which provides the algorithm with no labeled data in order to help it to find structure within its input data.

3.2 Supervised Learning

With supervised learning, computers are able to make predictions based on labeled training data. Each training sample includes an input and a desired output. A supervised learning algorithm analyzes this sample data and makes an inference – basically, an educated guess while determining the labels for unlabeled data. These models need to be fed manually tagged sample data to learn from. Data is labeled to tell the machine what patterns, *(e.g. similar words and images)*, it should be searching for and establish connections with.

An extremely typical case of supervised learning is to use historical data to predict statistically likely future events. That should be particularly helpful as our team may use historical stock market information to foresee upcoming oscillations.

There are two types of supervised learning tasks that are *classification* and *regression*.

- **Classification**: Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters. Additionally, a computer program is trained on the

training dataset and based on that training, it categorizes the data into different classes. The output variable must be a discrete value.

- **Regression**: Regression is a process of finding the correlations between dependent and independent variables. It paves the way to forecast the continuous variables such as prediction of Market Trends, prediction of House prices, and so on. The output variable must be of continuous nature or real value.

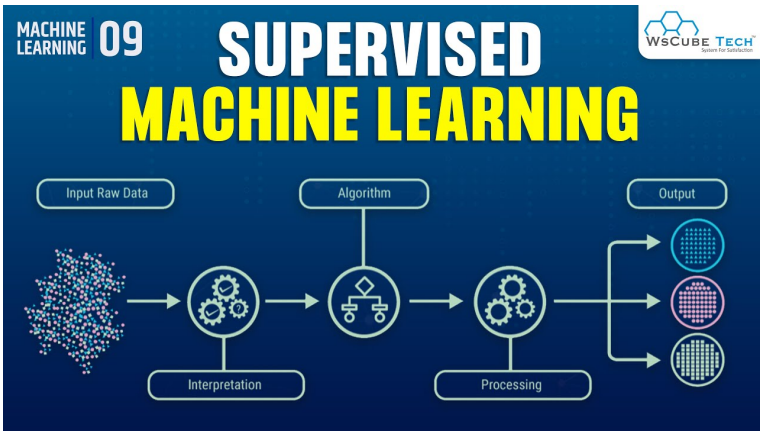

Figure 8: The process of supervised learning

3.3 Unsupervised Learning

Unsupervised learning algorithms uncover insights and relationships in unlabeled data. In this case, models are fed input data but the desired outcomes are unknown, so they have to make inferences based on circumstantial evidence, without any guidance or training. The models are not trained with the "right answer," so they must find patterns on their own.

One of the most common sorts of unsupervised learning is *clustering*, which includes grouping similar data. This method is mostly utilized for exploratory analysis and allows users to detect hidden patterns or trends.

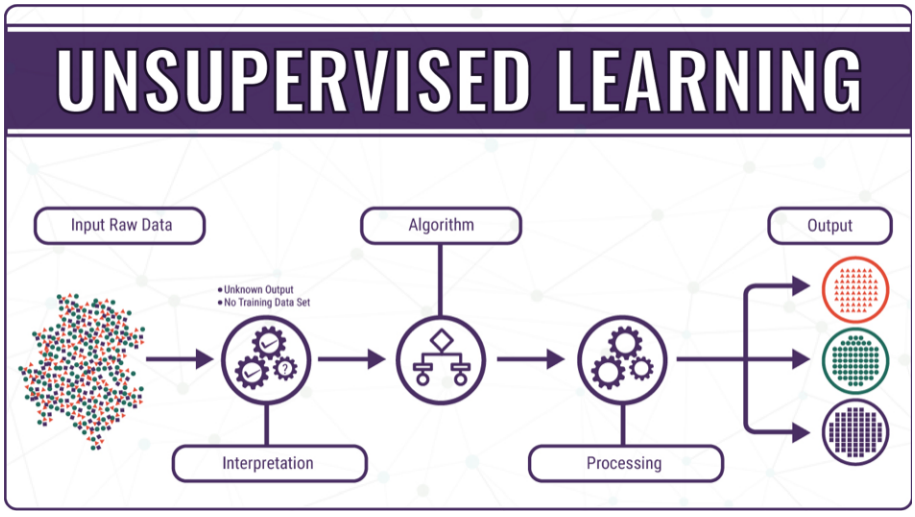

Figure 9: The process of unsupervised learning

3.4 Semisupervised Learning

In semisupervised learning, it takes a middle-ground approach that developers can be able to enter a relatively small set of labeled training data, as well as a larger corpus of unlabeled data. The algorithm is then instructed to extrapolate what it learns from the labeled data to the unlabeled data and draw conclusions from the set as a whole.

3.5 Deep Learning

This technique that teaches computers to do what comes naturally to humans: learn by example. Driverless cars use deep learning as a vital technology to detect stop signs and tell a pedestrian from a lamppost apart. It is of fundamental importance for voice control on consumer electronics including hands-free speakers, tablets, TVs, and smartphones. Recently, deep learning has attracted a lot of interest, and for good reason. It is producing outcomes that were previously unattainable. A computer model learns to carry out categorization tasks directly from pictures, text, or voice using deep learning. Cut-edge precision may be attained by deep learning models, sometimes even outperforming human ability. A sizable collection of labeled data and multi-layered neural network architectures are used to train models.

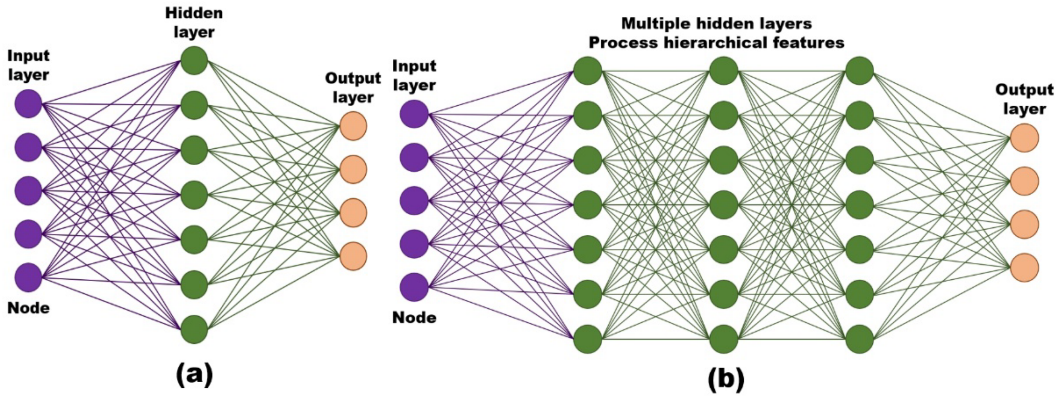

Figure 10: Deep Learning

Since neural network architectures are known to be used in the techniques' workings, deep learning models are frequently referred to as deep neural networks. The number of hidden layers in the neural network is typically indicated by the term "deep." While deep networks can have as many as 150 hidden layers, traditional neural networks typically have two or three.

Large collections of labeled data and neural network topologies that automatically extract characteristics from the data are used to train deep learning models.

4 Distribution

4.1 Theory of distribution

4.1.1 What is data distribution?

A probability distribution is a formula or a table used to assign probabilities to each possible value of a random variable X. A probability distribution may be either discrete or continuous.

A discrete distribution is one in which the data can only take on certain values, for example, integers. For a discrete distribution, probabilities can be assigned to the values in the distribution - for example, "the probability that the web page will have 12 clicks in an hour is 0.15."

A continuous distribution is one in which data can take on any value within a specified range (which may be infinite).

In contrast, a continuous distribution has an infinite number of possible values, and the probability associated with any particular value of a continuous distribution is null.

Therefore, continuous distributions are normally described in terms of probability density, which can be converted into the probability that a value will fall within a certain range.

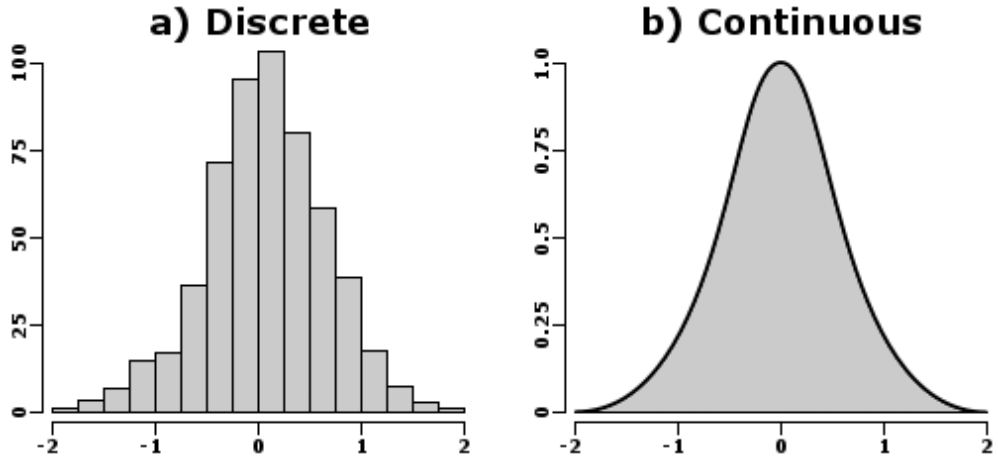

Figure 11: Discrete distribution and continuous distribution

All in all, data distributions are used often in statistics (stock analysis). They are graphical methods of organizing and displaying useful information of any dataset. Like the histogram, the distribution tells us about the shape and spread of the data.

Figure 12: Movement of VNINDEX

As we can see from figure 12, an instance of the stock movement of VNINDEX is a continuous distribution (the number of values is infinite) after examining the movement of the stock indexes during the period from 7/2021 to 12/2021. The opening and closing of stock prices occur often, constantly, and nearly without interruption throughout the year. The stock price fluctuation is also always less than 5 % and always within a particular range, often only 3 % less than the previous trading day. As a result, the continuous distribution can clearly display how the stock indexes are moving.

4.1.2 What data distribution can do?

The basic advantage of data distribution is to estimate the probability of any specific observation in a sample space. Probability distribution is a mathematical model that calculates the probability of occurrence of different possible outcomes in a test or experiment.

Data distribution has a huge application in the stock market.

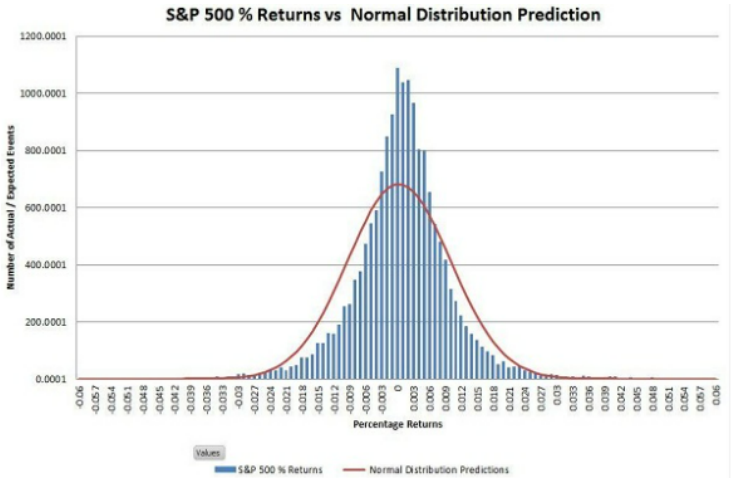

Figure 13: Source: StudiousGuy

Most of us have heard about the rise and fall in the prices of the shares in the stock market. Our parents or in the news about falling and hiking in the price of the shares. These changes in the log values of Forex rates, price indices, and stock prices return often form a bell-shaped curve. For stock returns, the standard deviation is often called volatility. If returns are normally distributed, more than 99 percent of the returns are expected to fall within the deviations of the mean value. Such characteristics of the bell-shaped normal distribution allow analysts and investors to make statistical inferences about the expected return and risk of stocks.

4.1.3 Normal Distribution

It is also known as the bell curve, symmetric distribution, or Gaussian distribution (named after Carl Friedrich Gauss). This is a type of continuous probability distribution for a real-valued random variable.

Let's start with these 3 ideas: mean, median, and mode to gain a greater knowledge of the Gaussian distribution. Describe them.

Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

Type	Description	Example	Result
Arithmetic mean	*Sum of values of a data set divided by number of values*	(1+2+2+3+4+7+9) / 7	4
Median	*Middle value separating the greater and lesser halves of a data set*	1, 2, 2, **3**, 4, 7, 9	3
Mode	*Most frequent value in a data set*	1, **2**, **2**, 3, 4, 7, 9	2

Figure 14: Definition and example of Mean, Median, and Mode (From Wikipedia)

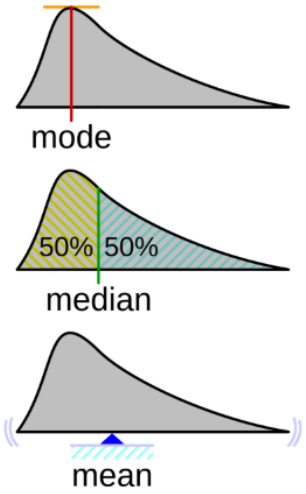

Figure 15: Geometric visualization of the Mode, Median, and Mean (From Wikipedia)

The distribution is symmetrical and unimodal (there is only one peak) since the normal distribution's mean, median, and mode are all the same.

And data is not always symmetrical. Frequently, it has some extremely high or extremely low numbers on one side, rendering it slightly skewed. The skewness of a probability distribution of a real-valued random variable is a measure of its asymmetry relative to its mean in probability theory and statistics. The value of the skewness might be positive, negative, zero, or undefined.

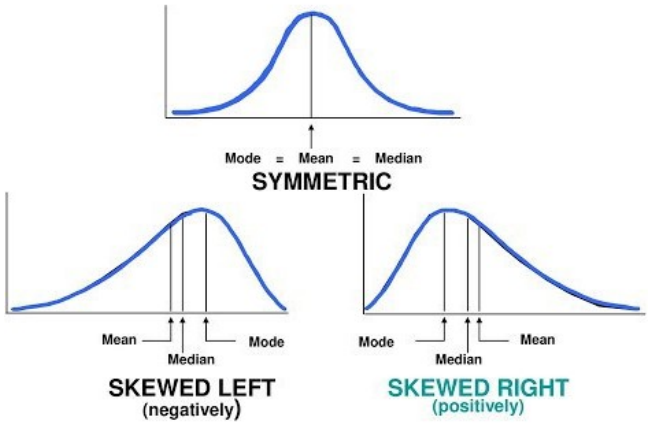

Figure 16: Skewness of Normal distribution (From Medium)

Boxplot of Normal distribution:

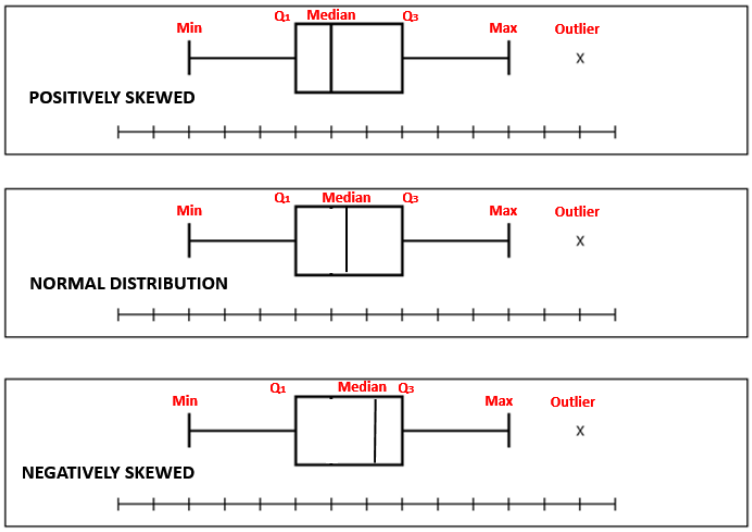

Figure 17: Boxplot of Gaussian distribution (From Studywalk)

The mean and standard deviation of the sample determines the form of the Normal Distribution curve; the curve will be symmetrical and centered on the mean and extended by the standard deviation. The Probability density function curve is non-zero over the whole real line because it never crosses the x-axis. The chance of any event occurring may be determined using the Gaussian distribution, but as an event deviates from the mean, its likelihood of occurring will approach progressively closer to zero.

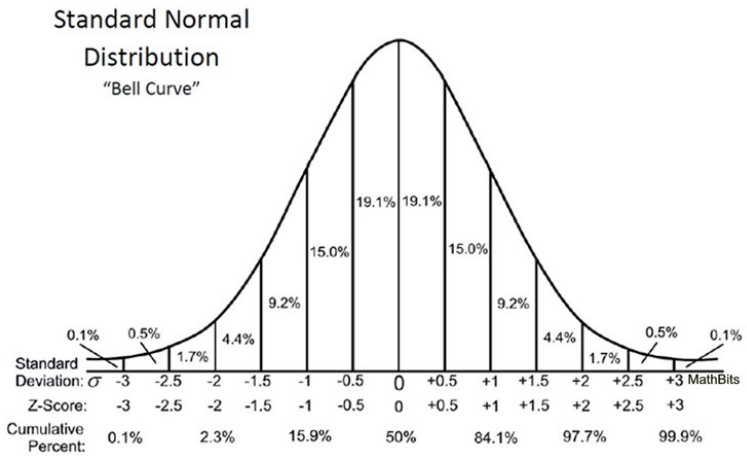

Figure 18: Standard Normal Distribution (From MathBitsNotebook)

Around 68 % of values drawn from a normal distribution are within one standard deviation σ away from the mean; about 95 % of the values lie within two standard deviations; about 99.7 % are within three standard deviations. This fact is known as the 68-95-99.7 (empirical) rule or the 3-sigma rule. This is highly helpful when attempting to spot outliers in your data or even when determining the normality of the distribution.

The formula of Gaussian distribution:

$$f(x|\mu, \sigma^2) = \frac{1}{2\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where:

μ : Mean (In normal distribution: Mean = Mode = Median).

σ : Standard Deviation.

x: Normal random variable.

When we examine categories of things such as IQ or on exams such as the SAT, the normal distribution frequently appears. As a result, it is frequently seen in measurement errors, test points, salary, human height, blood pressure, and IQ scores and is widely utilized in statistics, business, and government organizations. Skew is another helpful metric for contrasting data. For instance, to determine how challenging a test was, teachers frequently look at the distribution of test scores. When students perform poorly on really difficult examinations, which often provide negatively skewed scores, they can analyze a variety of information from the distribution.

4.1.4 Gamma Distribution

The two-parameter family of continuous probability distributions is known as the gamma distribution. This is a particular case of the exponential distribution, Erlang distribution, and chi-square distribution. In practice, there are two equivalent parameterizations:

With a shape parameter k and a scale parameter θ.

With a shape parameter α = k and an inverse scale parameter β = 1 / θ, called a rate parameter.

It deals with variables that are continuous and have a tremendous range of values. Using a gamma distribution function as a foundation, we may describe probabilities throughout any range of conceivable values.

The formula of gamma distribution:

In the case when **X** is a continuous random variable, the probability distribution function is:

$$f(n) = \begin{cases} \dfrac{1}{\beta\Gamma(\alpha)} x^{\alpha-1} e^{\dfrac{-x}{\beta}} & \text{if x} \geqslant 0 \\ 0 & \text{if otherwise} \end{cases}$$

Where:

$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$: Gamma function.

α : the shape parameter.

β : the rate parameter (the reciprocal of the scale parameter). (sometimes σ is used instead)

α and β are both greater than 1.

When $\alpha = 1 \Rightarrow$ Exponential distribution.

When $\beta = 1 \Rightarrow$ Standard gamma distribution.

Additional special examples of the Gamma distribution include the Erlang distribution and the chi-square distribution.

Below is a plot of its Probability density function and Cumulative distribution function, respectively:

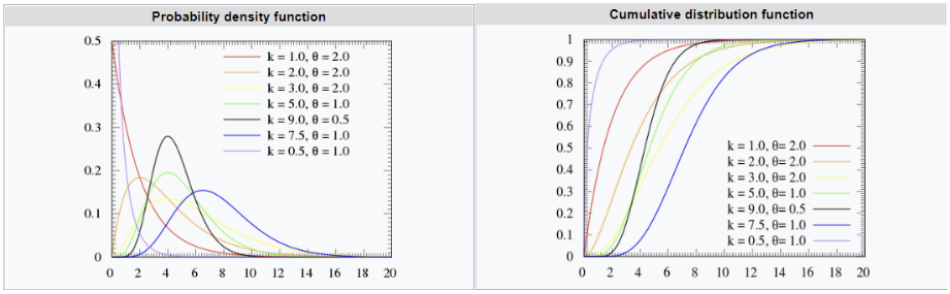

Figure 19: Gamma distribution (From Wikipedia)

If we consider α to be the total number of events we are anticipating (although α need not be an integer, it may be any positive number), and β to be the average amount of time before the first occurrence. It makes it reasonable that the graph shifts to the right as waiting times grow if α (number of occurrences) stays constant but β (mean time between events) rises. The graph will also move to the right if the mean waiting time (β) remains constant but more events occur (α). The gamma closely resembles the normal distribution as α nears infinity. As a result, the demand on web servers, the quantity of rainwater that has been collected in a reservoir, etc. are applications of the gamma distribution.

4.1.5 Exponential Distribution

As was already noted, the exponential distribution is a specific instance of the gamma distribution. It is the geometric distribution's continuous equivalent and possesses the crucial quality of being memoryless.

The formula of exponential distribution:

For every b $>$ 0, the probability density function and cumulative distribution function for the exponential random variable are as follows:

$$P(x) = \frac{1}{b} exp \frac{-x}{b} u(x)$$

$$P(X \leqslant x) = [1 - \frac{1}{b} exp \frac{-x}{b}] u(x)$$

Below is a plot of its Probability density function and Cumulative distribution function, respectively:

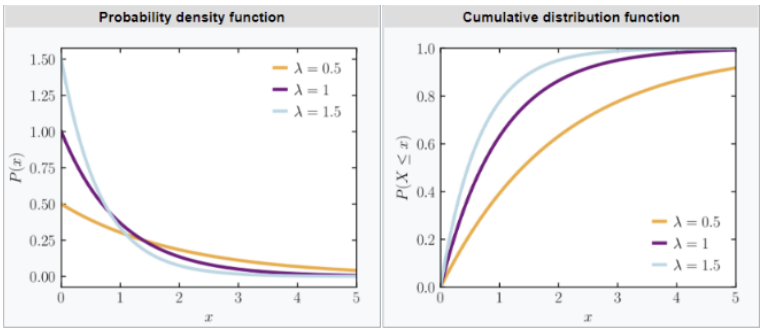

Figure 20: Exponential distribution (From Wikipedia)

One of the widely used continuous distributions is the exponential distribution. It helps to determine the time elapsed between the events. It is used in a range of applications such as reliability theory, queuing theory, physics, and so on. Some of the fields that are modeled by the exponential distribution are as follows:

- Exponential distribution helps to find the distance between mutations on a DNA strand.
- Calculating the time until the radioactive particle decays.
- Helps in finding the height of different molecules in a gas at the stable temperature and pressure in a uniform gravitational field.
- Helps to compute the monthly and annual highest values of regular rainfall and river outflow volumes.

4.2 Code to plot histogram

```
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import csv
\%matplotlib inline
from google.colab import drive
drive.mount('/content/drive')

raw_data=pd.read_excel('/content/drive/MyDrive/BTL discrete/distribution/Ban sao cua PSH.xlsx')
raw_data

sns.distplot(raw_data.Open)
```

4.3 Distribution Analysis(According to MATLAB tool)

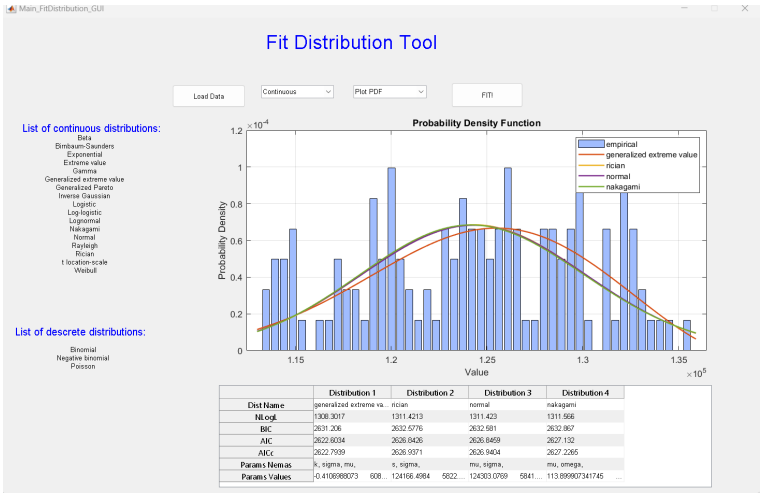

Figure 21: An instance of evaluating the VNINDEX distribution

And similarly applies to the other 9 charts of 9 stocks.

4.3.1 Distribution of VNINDEX

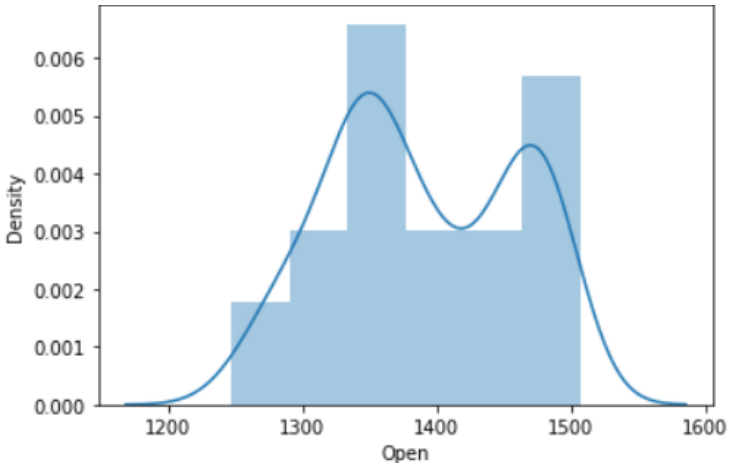

Figure 22: Distribution of VNINDEX

From the histogram distribution of VNINDEX, it does not follow any typical distribution(normal, exponential,...). However, it looks like a gamma distribution, the difference here is that the density of the value 1420 to 1490 suddenly increases.

In general, the most popular value stock price of VNINDEX are in the range from 1300 to 1350. The higher the stock price is, the lower density is.

4.3.2 Distribution of 3 stocks in three different sectors

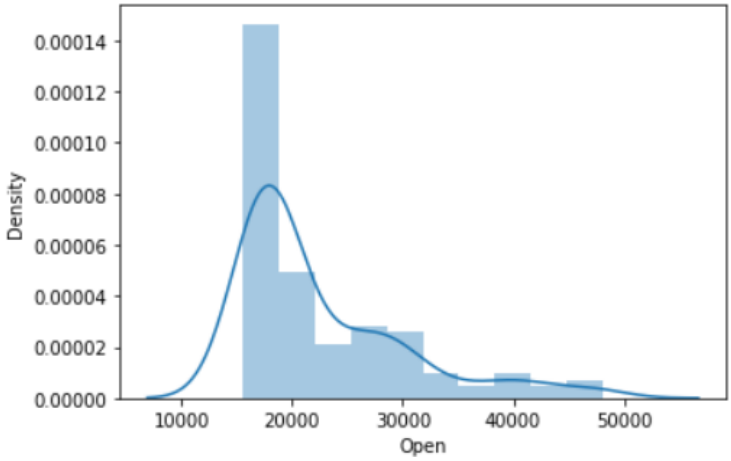

Figure 23: Distribution of CII

The distribution of CII is nearly lognormal distribution. the pdf rises very sharply in the beginning and essentially follows the ordinate axis, peaks out early, and then decreases sharply like an exponential pdf or a Weibull pdf.

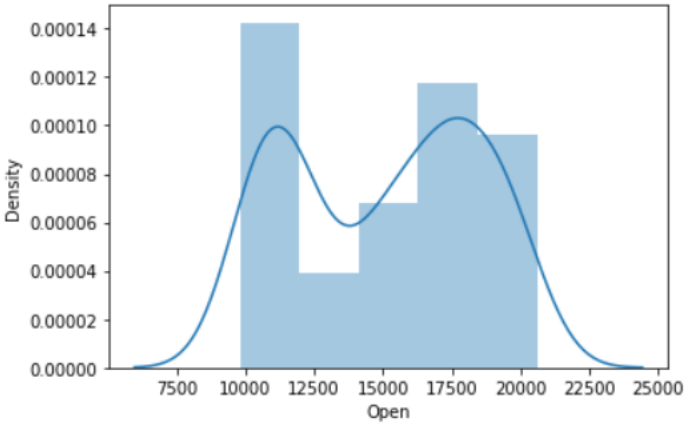

Figure 24: Distribution of SKG

The figure distribution of SKG seems like the distribution of VNINDEX. Therefore, it also does not follow any well-known distribution. However,in general, it looks like the Weibull distribution(another type of exponential distribution). Nevertheless, it has a decrease in density at the middle.

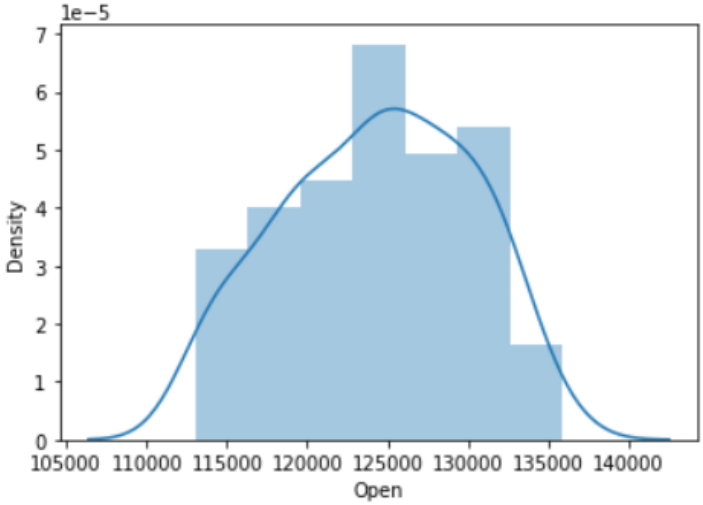

Figure 25: Distribution of VJC

From the histogram distribution of VJC, it can be easily seen that this is a normal(Gaussian) distribution. It is symmetrical, which means that if the distribution is cut in half, each side would be the mirror of the other.

4.3.3 Distribution of 3 stocks with the strongest drop

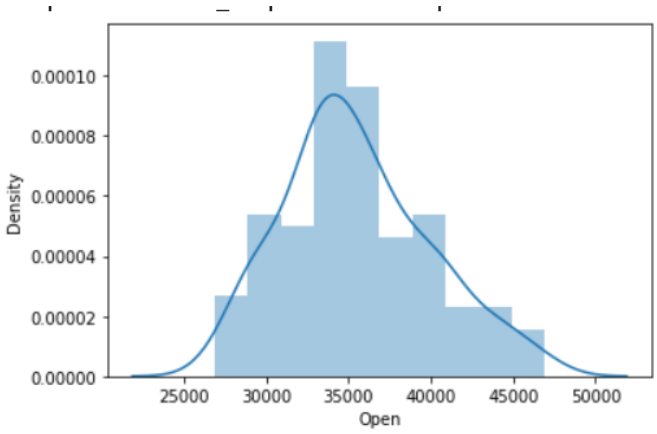

Figure 26: Distribution of APH

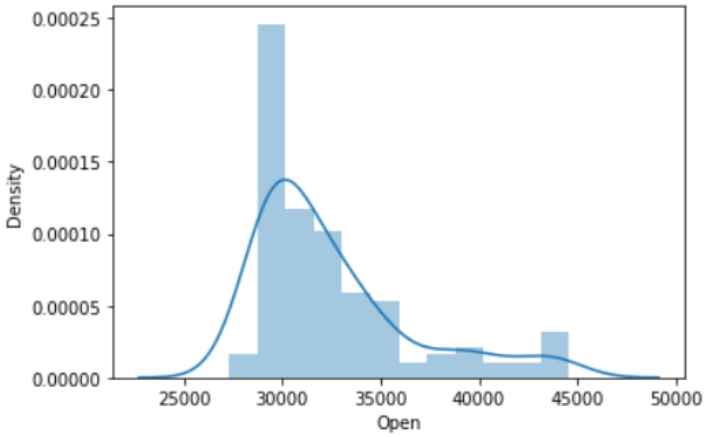

Figure 27: Distribution of NHH

The distribution of APH and NHH are quite the same and are lognormal distribution. The density increase sharply at the beginning of the range of the values, then it decreases sharply and approaches the Ox axis very fast.

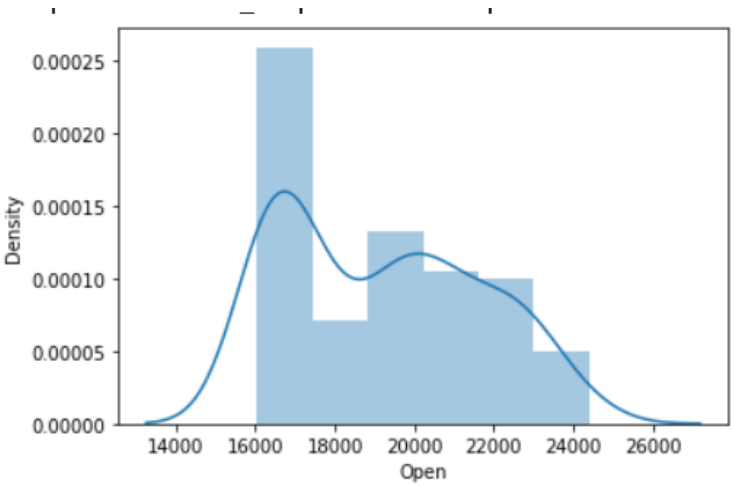

Figure 28: Distribution of PSH

This figure looks like distribution of VNINDEX(Figure 17), it seems to be a gamma distribution. And there is also a increase in the density after reaching the peak (similar to figure 17).

4.3.4 Distribution of 3 stocks with the highest growth

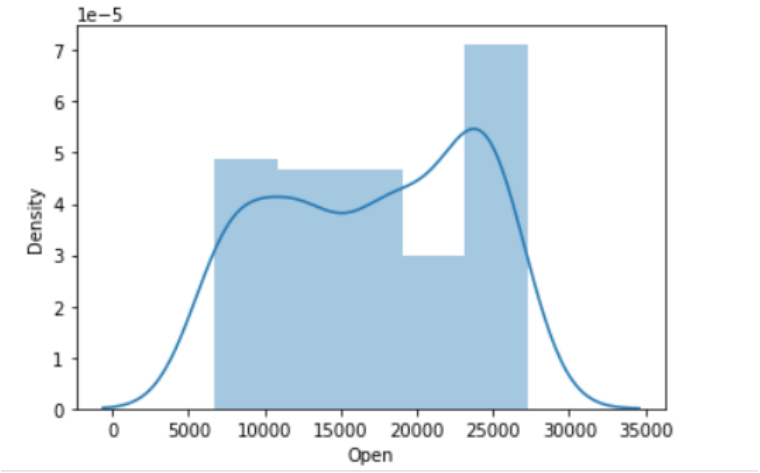

Figure 29: Distribution of TCD

The figure distribution of TCD is nearly normal distribution, it seems to be symmetric and reach the peak at the middle.

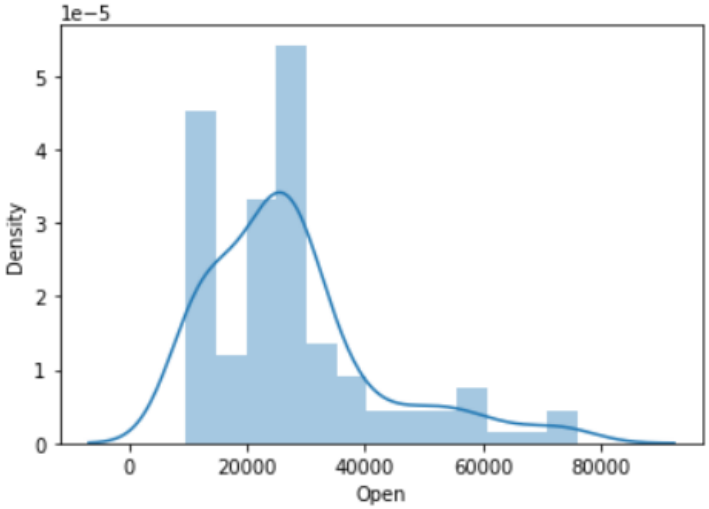

Figure 30: Distribution of TGG

The distribution of TGG is quite similar to the figures distribution of HNG and NHH. Absolutely, it is lognormal distribution.

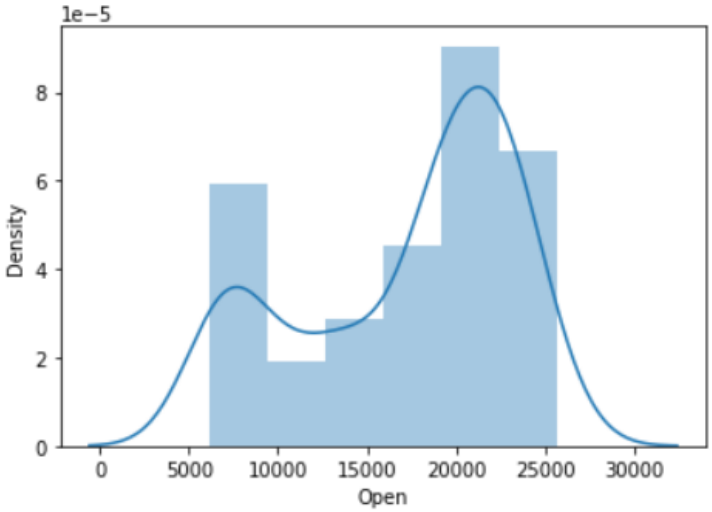

Figure 31: Distribution of VOS

The distribution of VOS stock index looks like Weibull distribution, which can approximate many other distributions: normal, exponential and so on. The Weibull curve is called a "bathtub curve".

5 Prediction Models

5.1 Time series data

Time series data is a collection of observations obtained through repeated measurements over time. Plot the points on a graph, and one of your axes would always be time.

There are several terms about this type of data:

1. Trend:trend of the data (up or down). This is detectable through the slope of the data on the graph.
2. Seasonality:The data is affected by the time measure like: hour, week, month, year, ... In this case, the data has a cycle that repeats according to the above time measure with a fixed frequency.
3. Cyclicity: a cycle occurs when data increases and decreases at a fixed frequency. The most obvious here is in the field of economics.
4. Residuals: the difference between actual value and predicted value at a particular time.

A data

5.2 Linear Regression model

5.2.1 What is Regression ?

Regression is defined as a statistical method that helps us to analyze and understand the relationship between two or more variables of interest. The process that is adapted to perform regression analysis helps to understand which factors are important, which factors can be ignored, and how they are influencing each other.

In regression, we normally have one dependent variable and one or more independent variables. Here we try to “regress” the value of the dependent variable “Y” with the help of the independent variables. In other words, we are trying to understand, how the value of ‘Y’ changes w.r.t change in ‘X’.

5.2.2 Regression Analysis

Regression analysis is used for prediction and forecasting. This has substantial overlap with the field of machine learning. This sta tistical method is used across difthe ferent industries such as,

Financial Industry - Understand the trend in the stock prices, forecast thinprices, and evaluate risks in the insurance domain.

Marketing - Understand the effectiveness of market campaigns, and forecast pricing and sales of the product.

Manufacturing - Evaluate the relationship of variables that determine to define a bewhich engine to provide better performance.

Medicine - Forecast the different combinations of medicines to prepare generic medicines for diseases.

Regression equation

$$Y = f(X, \theta)$$

- Y: The predict variable (dependent variable).
- X: The variable that is used to predict (independent variable).
- θ: Regression coefficient (describes the relative influence of X on Y).

5.2.3 Linear regression and its algorithm

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

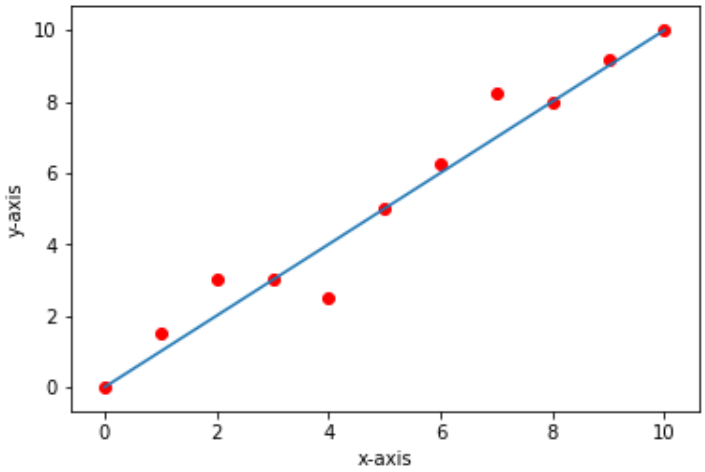

Figure 32: Sample of linear regression model.

5.2.4 Hypothesis function for linear regression

This is a mathematical function that helps us to convert the value from x to y so that the predicted value of y is correct or approximate to the actual value of the data samples we have collected.

The equation of the hypothesis function:

$$h_\theta(x) = \theta_0 + \theta_1 \times x$$

. $h_\theta(x)$: Hypothesis function.
. x: Value used to predict.
. θ_0, θ_1: Coefficients.

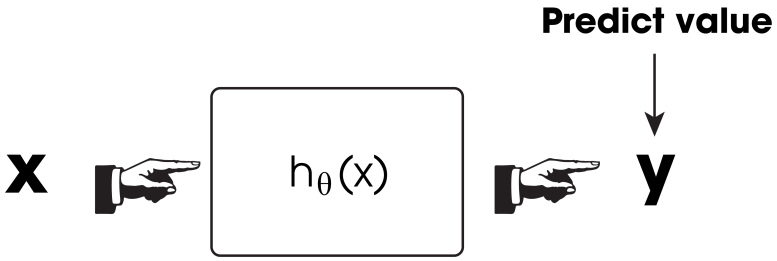

Figure 33: Hypothesis function

The Hypothesis function can be extended with even more weights:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

5.2.5 The cost function of linear regression

The next task is to determine the weights θ_0 and θ_1 so that we can draw a line that is as close to every point on our graph as appropriate (the distance between the line and the points is minimal) in order to minimize the error.

In Machine Learning, Cost function is the calculation of the error between predicted values and actual values, represented as a single real number. The function is defined as follow:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

The term $h_\theta(x^{(i)}) - y^{(i)}$ is called error term or residual analysis (in statistics). This term represents the distance between the line and the points in the graph.

. m: Number of data samples collected.

. x: Value used to predict.

. y: The real value in data.

. $x^{(i)}$: The i^{th} value of x.

. $y^{(i)}$: The i^{th} value of y.

. (x, y): The complete record of the data we collect.

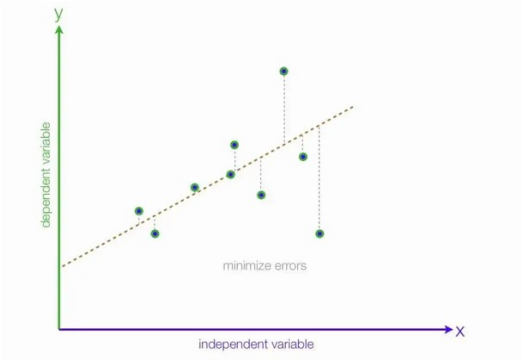

Figure 34: Minimize Errors

5.2.6 Gradient Descent technique

Gradient Descent is an iterative optimization algorithm, used to find the minimum value for a function. The general idea is to initialize the parameters to random values, and then take small steps in the direction

of the "slope" at each iteration. Gradient descent is highly used in supervised learning to minimize the error function and find the optimal values for the parameters.

5.2.7 Prediction of VNINDEX

```
df = pd.read_excel("/content/drive/MyDrive/Colab Notebooks/Data/VNI year.xlsx",
parse_dates=['Date'], index_col='Date')
df = df[['Close']]
```

```
future_days = 59
df['Prediction'] = df[['Close']].shift(-future_days)
```

```
X = np.array(df.drop(['Prediction'], 1))[:-future_days]
Y = np.array(df['Prediction'])[:-future_days]
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.3122,
random_state = 0)
```

```
model = LinearRegression()
model.fit(X_train,Y_train)
```

```
X_future = df.drop(['Prediction'], 1)[:-future_days]
X_future = X_future.tail(future_days)
X_future = np.array(X_future)
# This variable is the previous 'future_days' days (59 days previous in 2021)
# We will use this to predict the next 59 days (till 1st April 2022)
```

```
linear_prediction = model.predict(X_future)
```

```
predictions = linear_prediction

valid = df[X.shape[0]:]
valid['Predictions'] = predictions
plt.figure(figsize = (16,8))
plt.title('Predicting VNINDEX stock indexes using Linear Regression')
plt.xlabel('Date')
plt.ylabel('Close Price')
plt.plot(df['Close'])
plt.plot(valid[['Close','Predictions']])
plt.legend(['Orig', 'Val', 'Pred'])
plt.show()
```

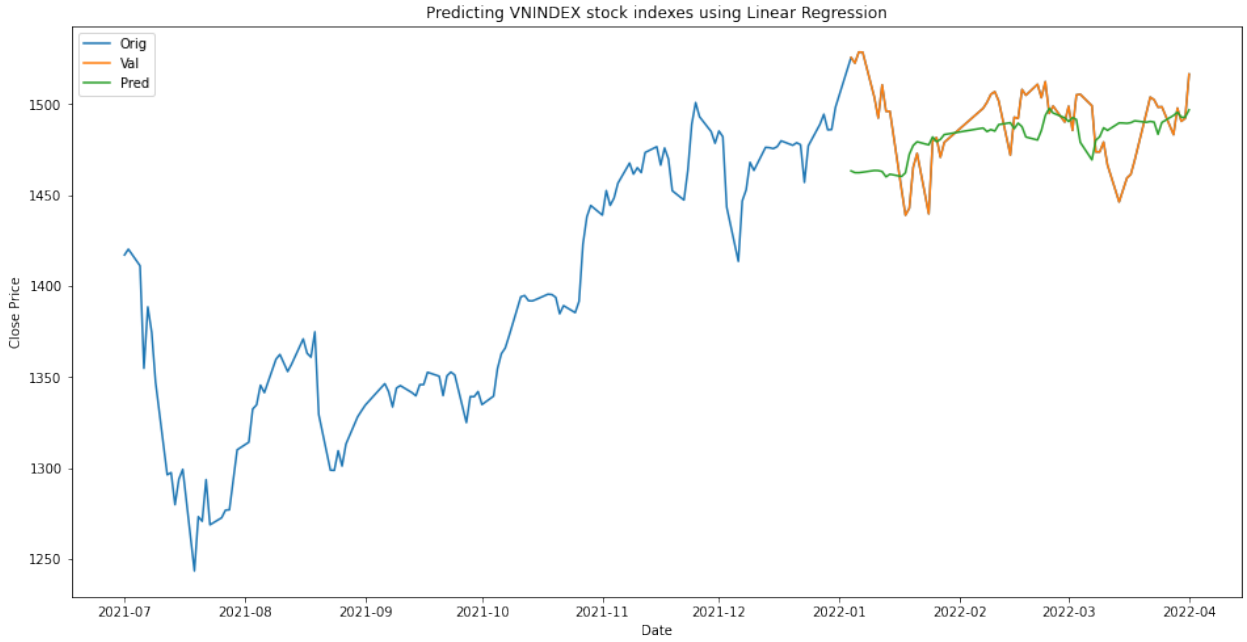

Figure 35: Predicting VNINDEX stock using Linear Regressive

5.2.8 Predictions of 3 stocks in three different sectors

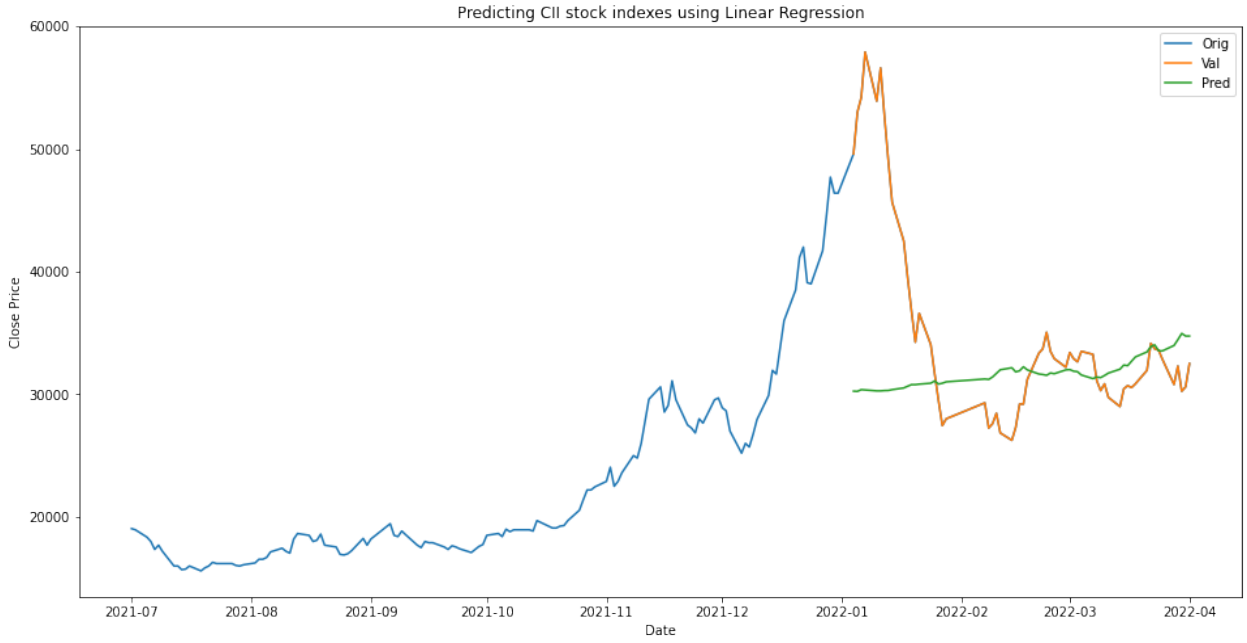

Figure 36: Predicting CII stock using Linear Regressive

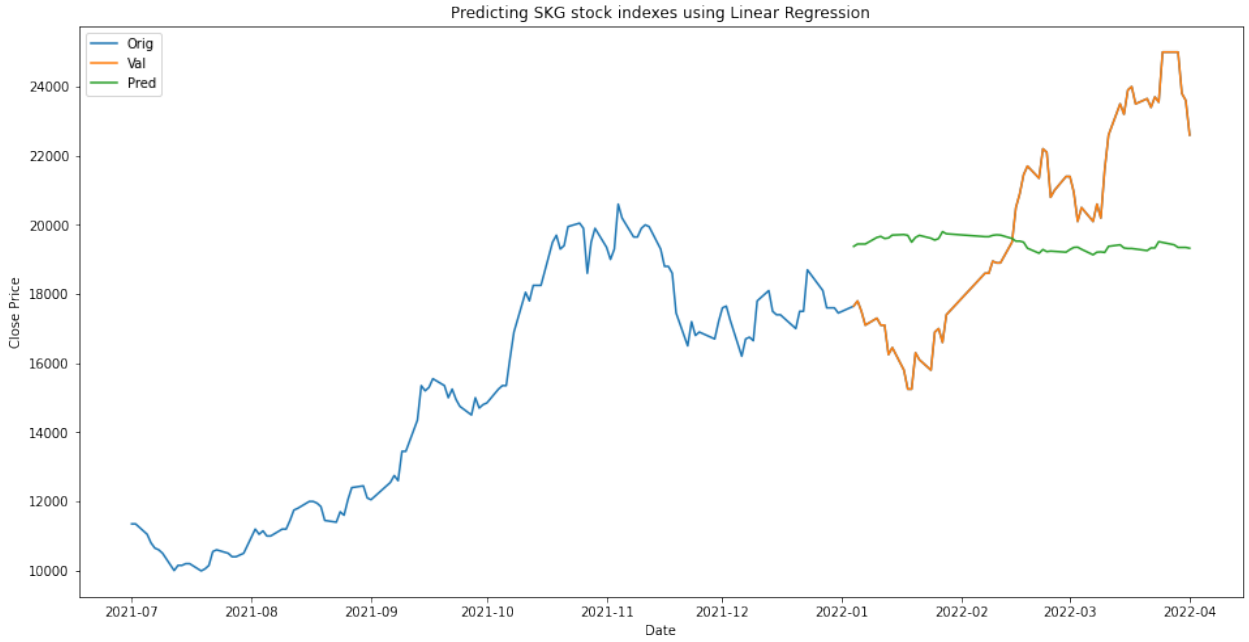

Figure 37: Predicting SKG stock using Linear Regressive

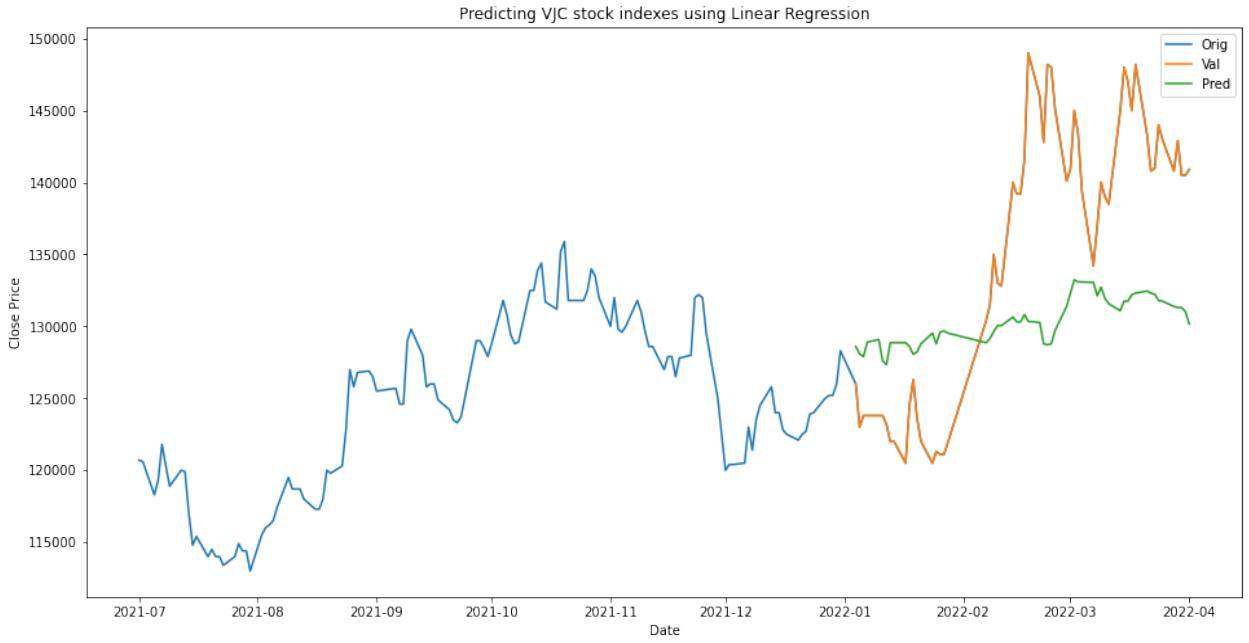

Figure 38: Predicting VJC stock using Linear Regressive

5.2.9 Predictions of 3 stocks with the strongest drop

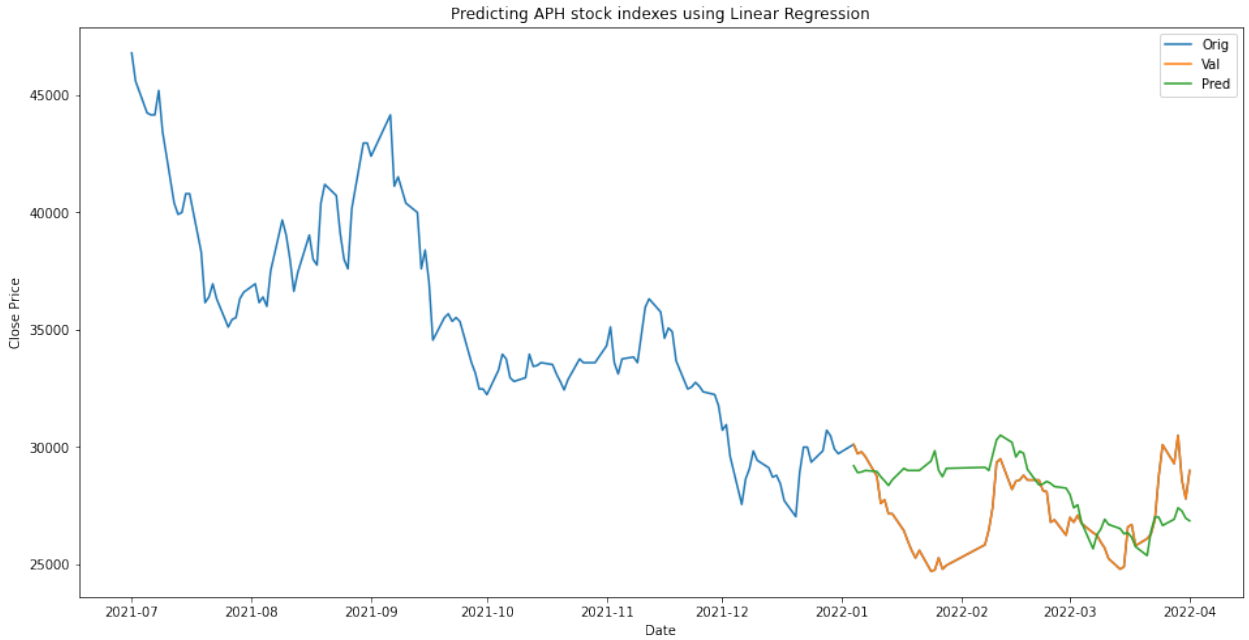

Figure 39: Predicting APH stock using Linear Regressive

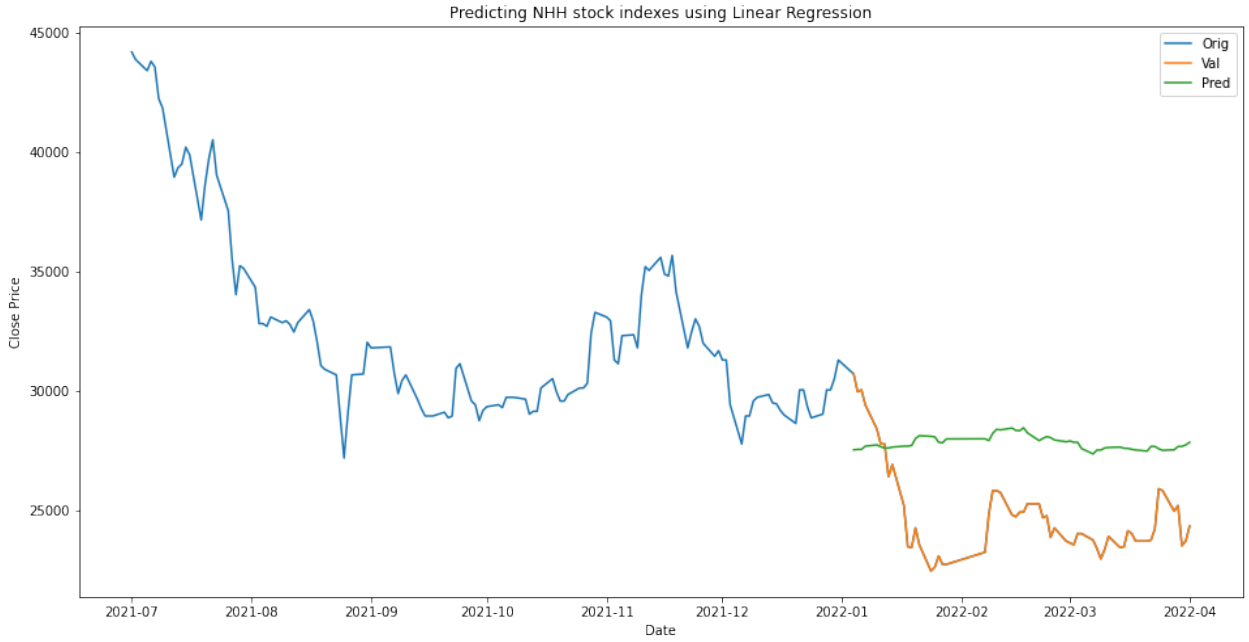

Figure 40: Predicting NHH stock using Linear Regressive

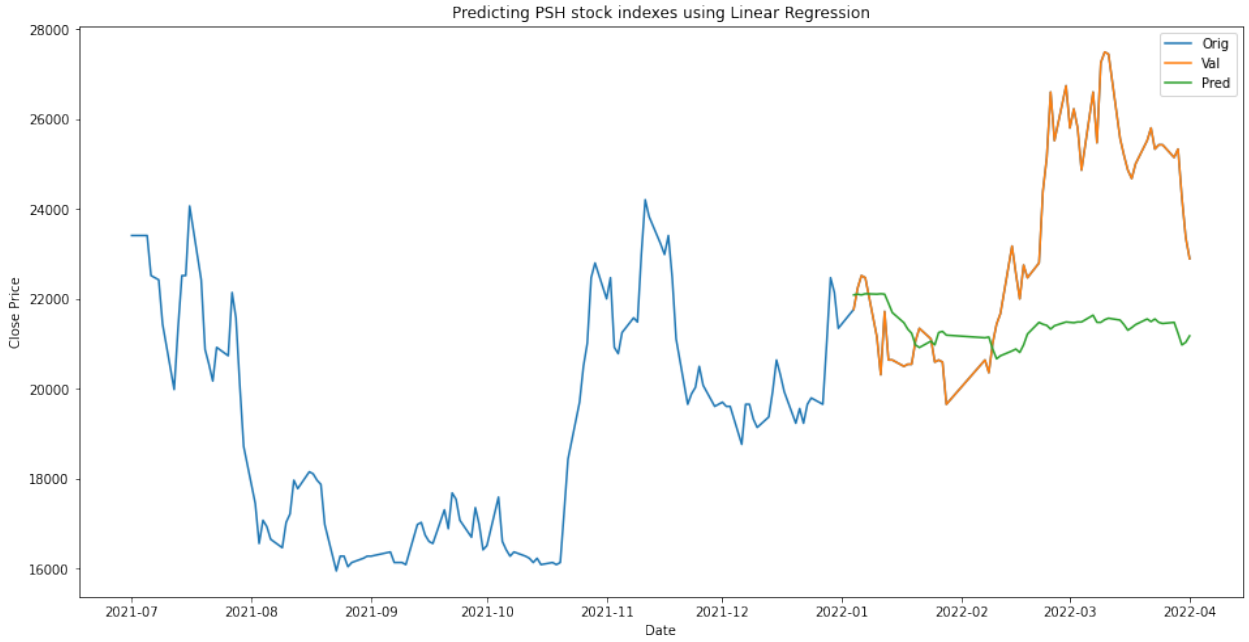

Figure 41: Predicting PSH stock using Linear Regressive

5.2.10 Predictions of 3 stocks with the highest growth

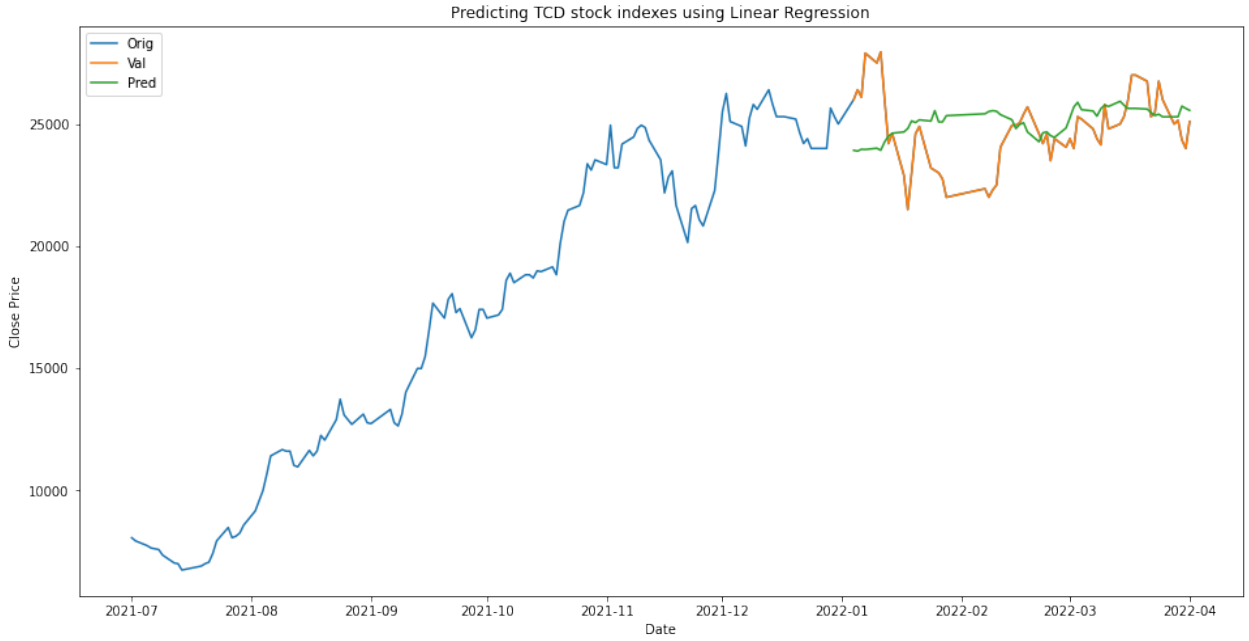

Figure 42: Predicting TCD stock using Linear Regressive

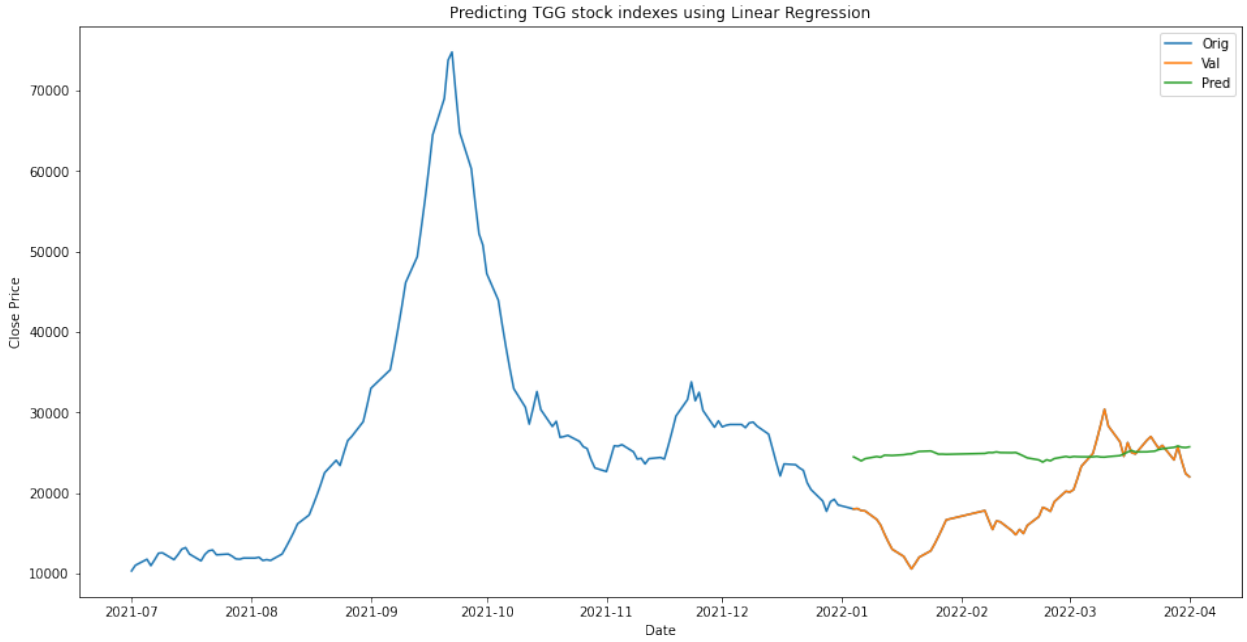

Figure 43: Predicting TGG stock using Linear Regressive

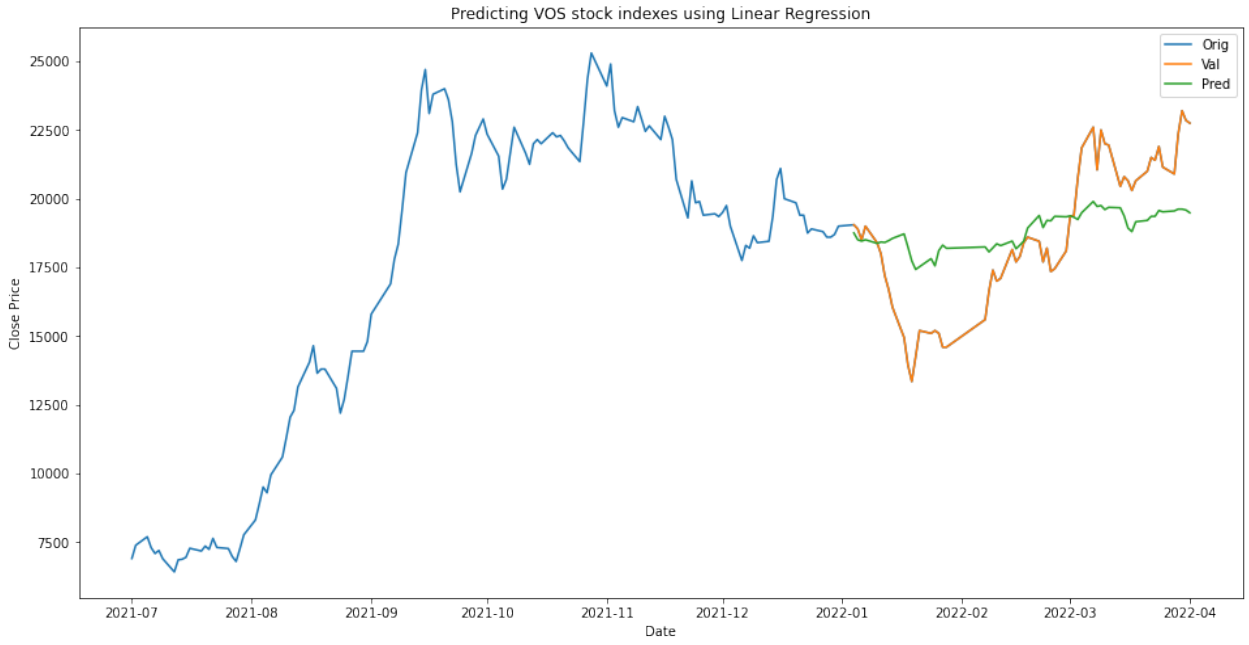

Figure 44: Predicting VOS stock using Linear Regressive

From those figures above, it can be seen that the Linear Regression model is not recommended to predict stock indexes due to its inaccuracy.

5.3 Decision Tree model

5.3.1 What is a Decision Tree?

A Decision Tree is a support tool with a tree-like structure that models probable outcomes, cost of resources, utilities, and possible consequences. Decision trees provide a way to present algorithms with conditional control statements. They include branches that represent decision-making steps that can lead to a favorable result.

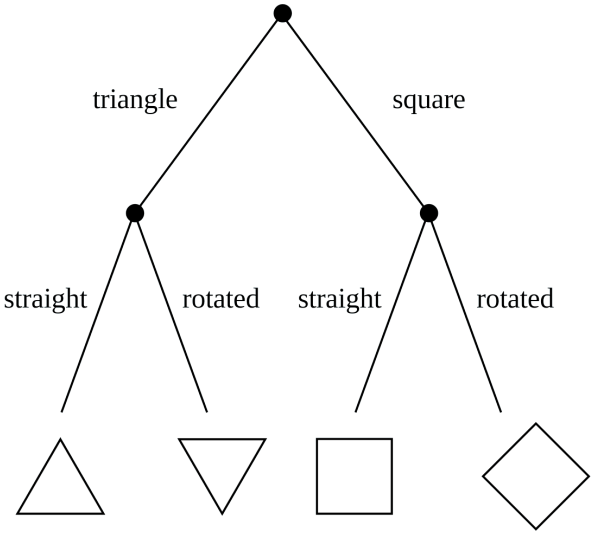

Figure 45: Decision Tree example.

5.3.2 Decision Tree summary

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

By learning straightforward decision rules derived from previous data, a Decision Tree is used to build a training model that may be used to predict the class or value of the target variable (training data).

In Decision Trees, we begin at the tree's root when anticipating a record's class label. We contrast the root attribute's values with that of the attribute on the record. We follow the branch that corresponds to that value and go on to the next node based on the comparison.

5.3.3 Types of Decision Trees

There are 2 types: *Categorical Variable Decision Tree* and *Continuous Variable Decision Tree.*

Categorical variable Decision Tree: A categorical variable decision tree includes categorical target variables that are divided into categories. For example, the categories can be yes or no. The categories mean that every stage of the decision process falls into one category, and there are no in-betweens.

Continuous Variable Decision Tree: A continuous variable decision tree is a decision tree with a continuous target variable. For example, the income of an individual whose income is unknown can be predicted based on available information such as their occupation, age, and other continuous variables.

5.3.4 Important Terminologies related to Decision Trees

Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets.

Splitting: It is a process of dividing a node into two or more sub-nodes.

Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node.

Leaf / Terminal Node: Nodes do not split is called Leaf or Terminal node.

Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.

Branch / Sub-Tree: A subsection of the entire tree is called branch or sub-tree.

Parent and Child Node: A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

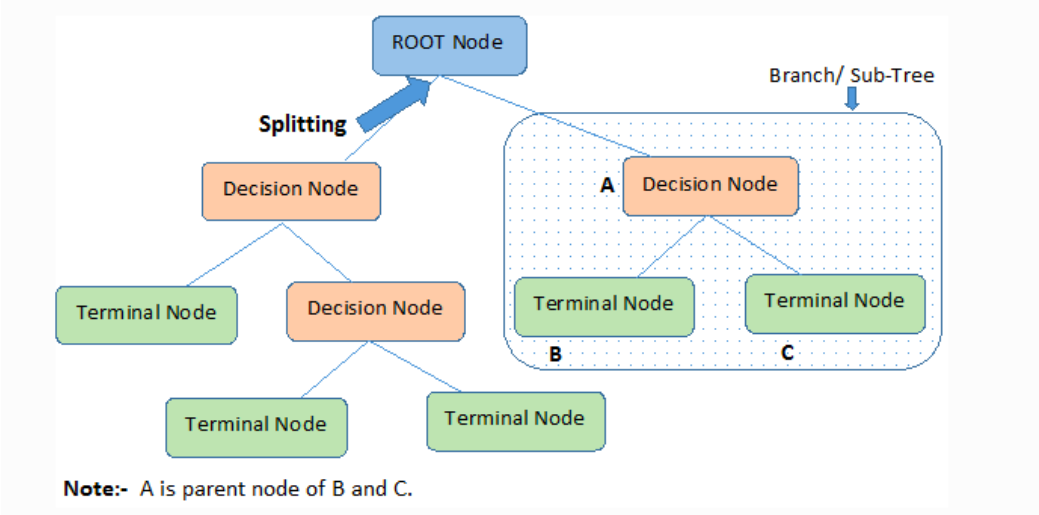

Figure 46: Basic Decision Tree.

5.3.5 Attribute Selection Measures

If the dataset consists of N attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can't solve the issue. If we follow a random approach, it may give us bad results with low accuracy.

For solving this attribute selection problem, researchers worked and devised some solutions. They suggested using some criteria like:

- Entropy
- Information gain
- Gini index
- Gain Ratio
- Reduction in Variance
- Chi-Square

These criteria will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by following the order i.e, the attribute with a high value(in case of information gain) is placed at the root.

While using Information Gain as a criterion, we assume attributes to be categorical, and for the Gini index, attributes are assumed to be continuous.

Entropy

Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information. Flipping a coin is an example of an action that provides information that is random.

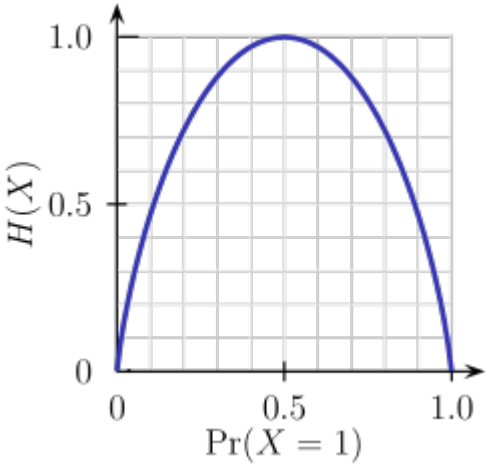

Figure 47: Graph representing flipping a coin.

From the above graph, it is quite evident that the entropy H(X) is zero when the probability is either 0 or 1. The Entropy is maximum when the probability is 0.5 because it projects perfect randomness in the data and there is no chance if perfectly determining the outcome.

Information Gain

Information gain or IG is a statistical property that measures how well a given attribute separates the training examples according to their target classification. Constructing a decision tree is all about finding an attribute that returns the highest information gain and the smallest entropy.

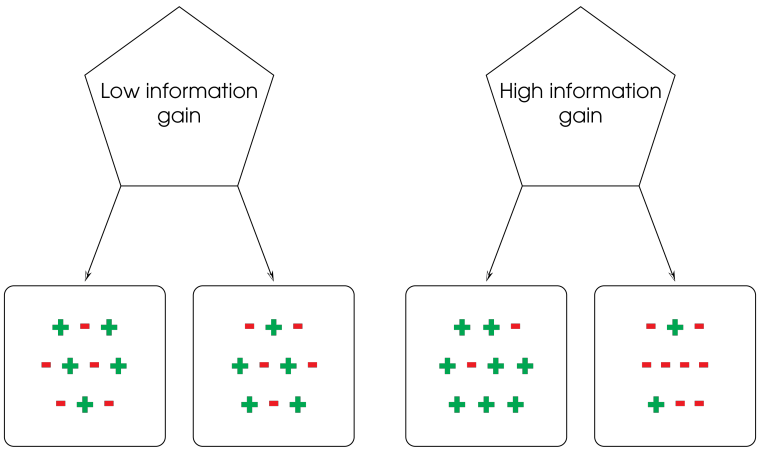

Figure 48: Information gain

Gini Index

You can understand the Gini index as a cost function used to evaluate splits in the dataset. It is calculated by subtracting the sum of the squared probabilities of each class from one. It favors larger partitions and easy to implement whereas information gain favors smaller partitions with distinct values.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Figure 49: Gini Index

Gain ratio

Information gain is biased towards choosing attributes with a large number of values as root nodes. It means it prefers the attribute with a large number of distinct values.

$$Gain\ Ratio = \frac{Information\ Gain}{SplitInfo} = \frac{Entropy\ (before) - \sum_{j=1}^{K} Entropy(j, after)}{\sum_{j=1}^{K} w_j\ log_2\ w_j}$$

Figure 50: Gain Ratio

Reduction in Variance

Reduction in variance is an algorithm used for continuous target variables (regression problems). This algorithm uses the standard formula of variance to choose the best split. The split with lower variance is selected as the criteria to split the population:

$$\text{Variance} = \frac{\Sigma(X - \overline{X})^2}{n}$$

Figure 51: Reduction in Variance

Chi-Square

The acronym CHAID stands for Chi-squared Automatic Interaction Detector. It is one of the oldest tree classification methods. It finds out the statistical significance between the differences between sub-nodes and parent node. We measure it by the sum of squares of standardized differences between observed and expected frequencies of the target variable.

It works with the categorical target variable "Success" or "Failure". It can perform two or more splits. Higher the value of Chi-Square higher the statistical significance of differences between sub-node and Parent node.

It generates a tree called CHAID (Chi-square Automatic Interaction Detector).

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared
O_i = observed value
E_i = expected value

Figure 52: Chi-Square

5.3.6 Pruning Decision Trees

The splitting process results in fully grown trees until the stopping criteria are reached. But, the fully grown tree is likely to overfit the data, leading to poor accuracy on unseen data.

In pruning, you trim off the branches of the tree, i.e., remove the decision nodes starting from the leaf node such that the overall accuracy is not disturbed. This is done by segregating the actual training

set into two sets: training data set, D and validation data set, V. Prepare the decision tree using the segregated training data set, D. Then continue trimming the tree accordingly to optimize the accuracy of the validation data set, V.

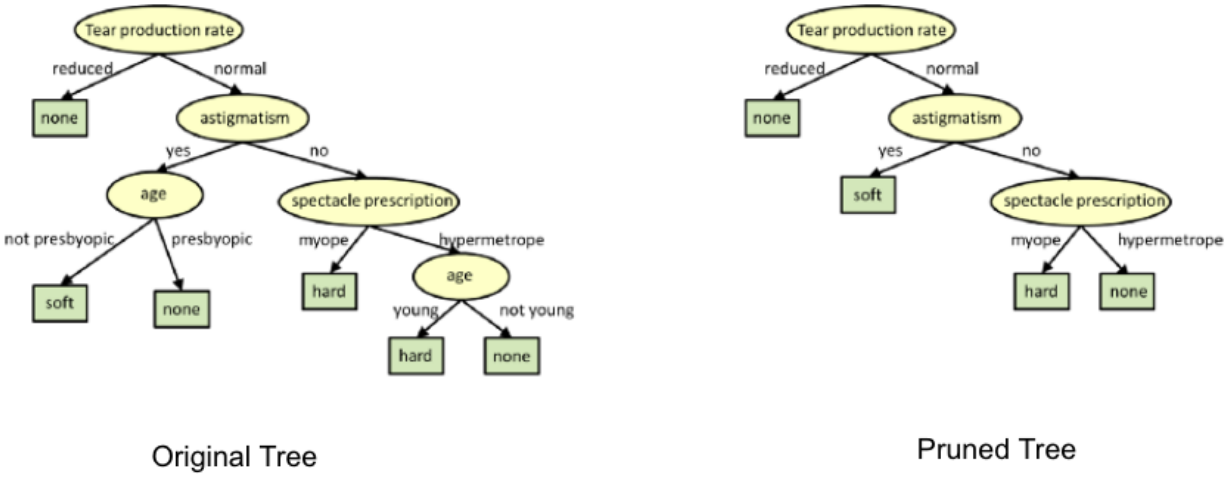

Figure 53: Pruning

In the above diagram, the 'Age' attribute in the left-hand side of the tree has been pruned as it has more importance on the right-hand side of the tree, hence removing overfitting.

Random Forest

Random Forest is an example of ensemble learning, in which we combine multiple machine learning algorithms to obtain better predictive performance.

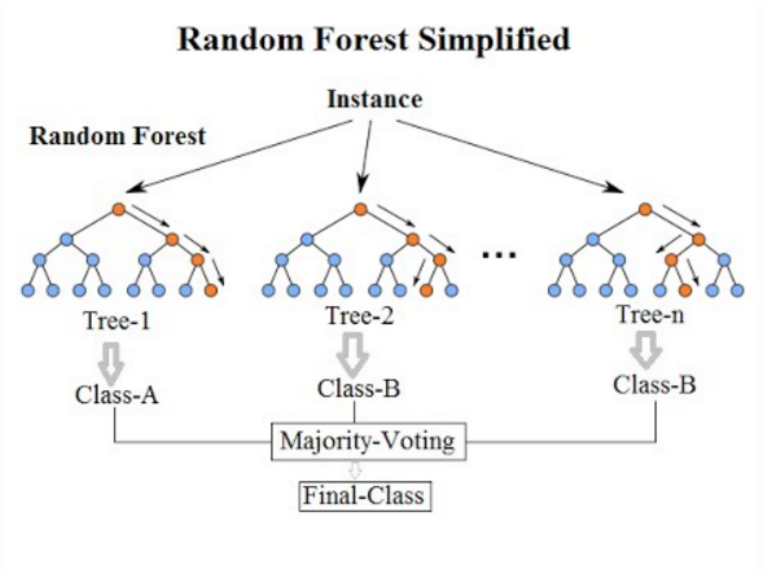

Figure 54: Random Forest

5.3.7 Prediction of VNINDEX

```
df = pd.read_excel("/content/drive/MyDrive/Colab Notebooks/Data/VNI year.xlsx",
parse_dates=['Date'], index_col='Date')
df = df[['Close']]
```

```
future_days = 59
df['Prediction'] = df[['Close']].shift(-future_days)
X = np.array(df.drop(['Prediction'], 1))[:-future_days]
Y = np.array(df['Prediction'])[:-future_days]
```

```
future_days = 59
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.3122,
random_state = 0)
model = DecisionTreeRegressor()
model.fit(X_train,Y_train)
```

```
X_future = df.drop(['Prediction'], 1)[:-future_days]
X_future = X_future.tail(future_days)
X_future = np.array(X_future)
```

```
tree_prediction = model.predict(X_future)
```

```
predictions = tree_prediction

valid = df[X.shape[0]:]
valid['Predictions'] = predictions
plt.figure(figsize = (16,8))
plt.title('Predict VNINDEX stock indexes using Tree Decision')
plt.xlabel('Date')
plt.ylabel('Close Price')
plt.plot(df['Close'])
plt.plot(valid[['Close','Predictions']])
plt.legend(['Orig', 'Val', 'Pred'])
plt.show()
```

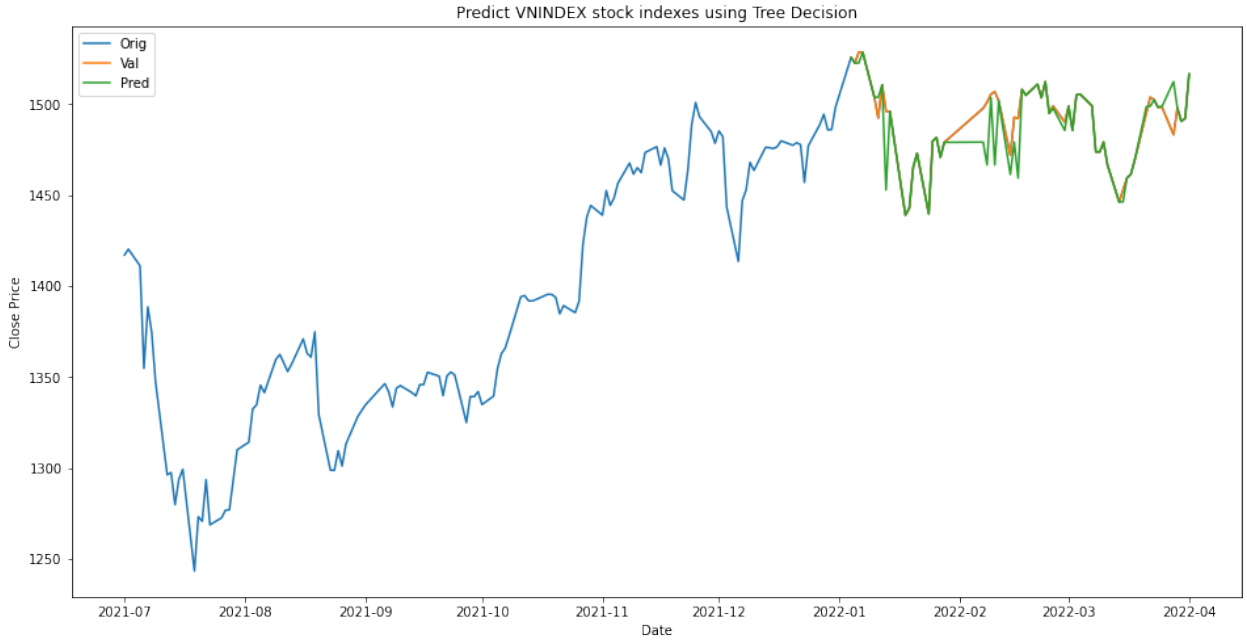

Figure 55: Predicting VNINDEX stock using Decision Tree

5.3.8 Predictions of 3 stocks in three different sectors

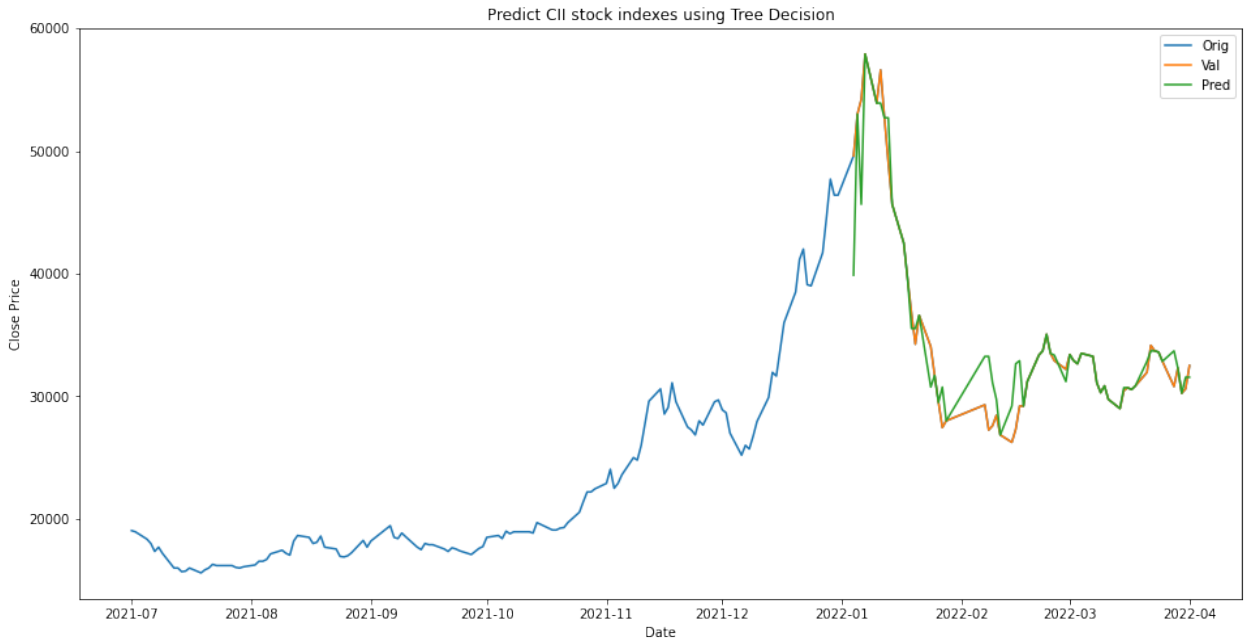

Figure 56: Predicting CII stock using Decision Tree

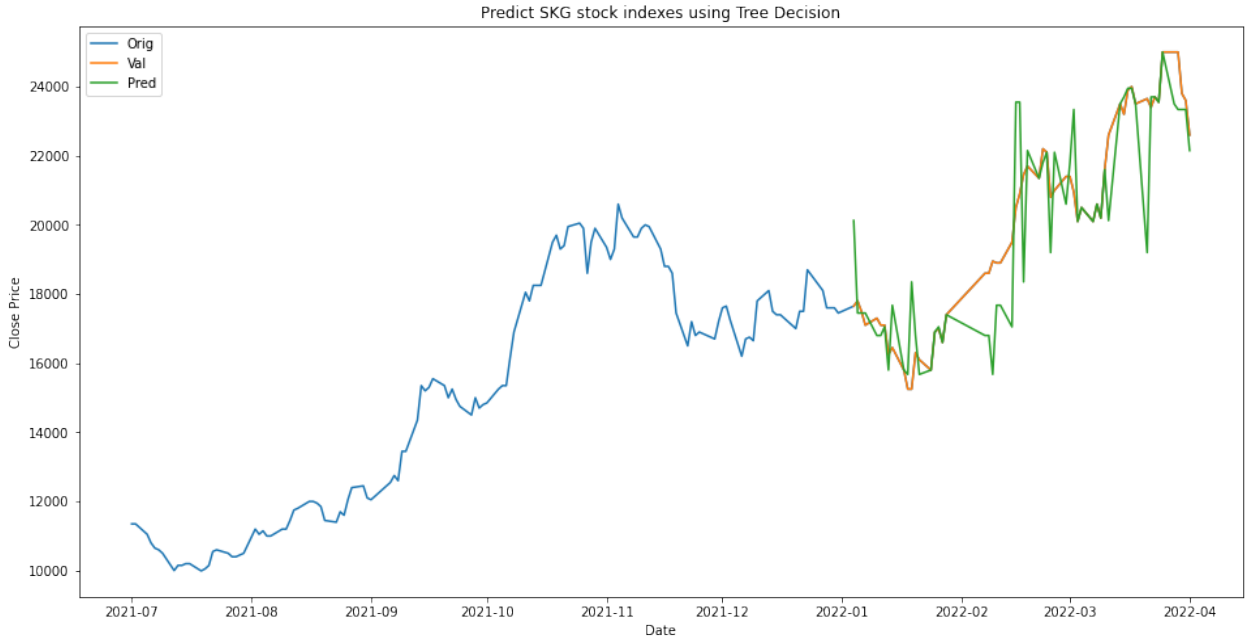

Figure 57: Predicting SKG stock using Decision Tree

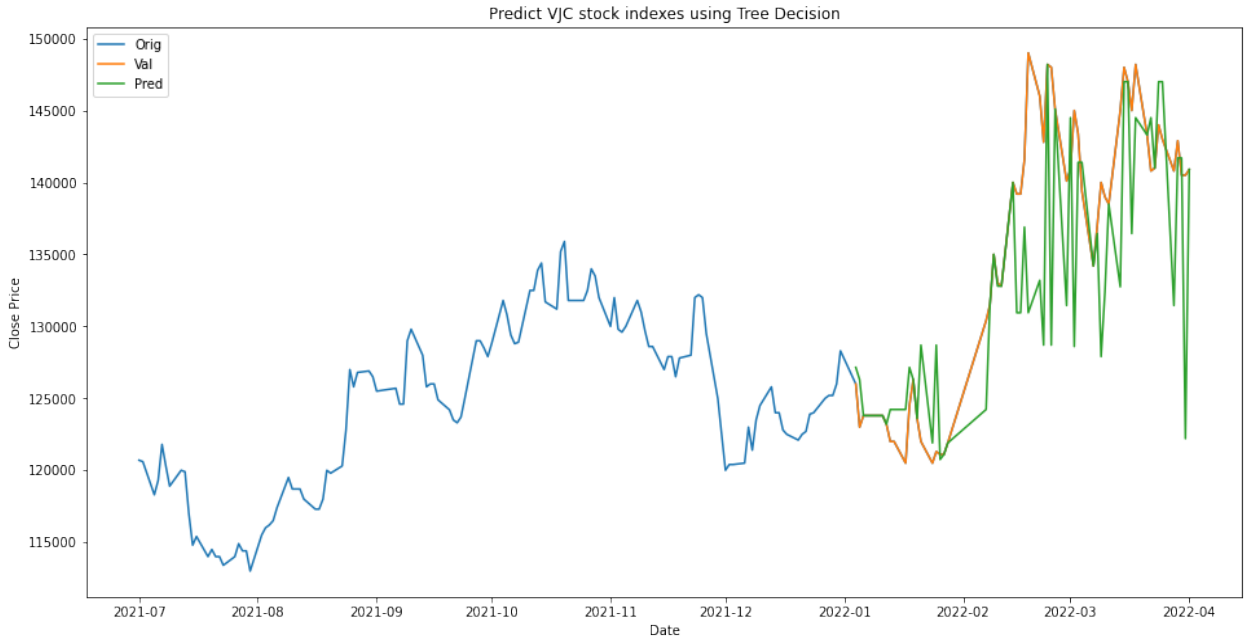

Figure 58: Predicting VJC stock using Decision Tree

5.3.9 Predictions of 3 stocks with the strongest drop

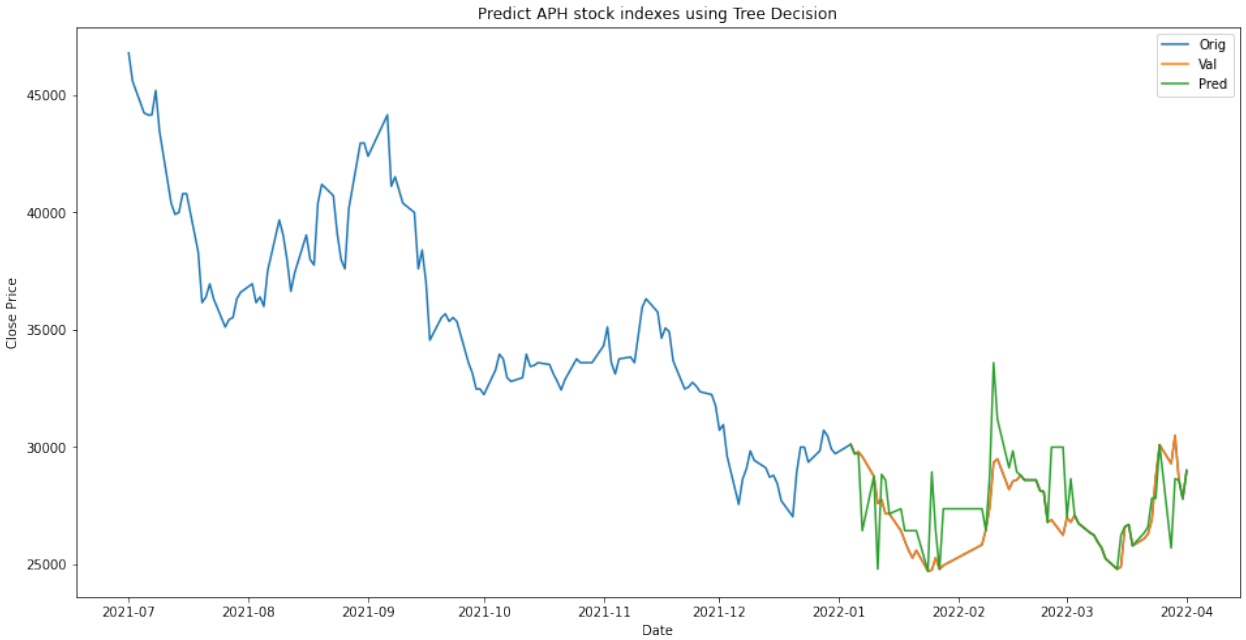

Figure 59: Predicting APH stock using Decision Tree

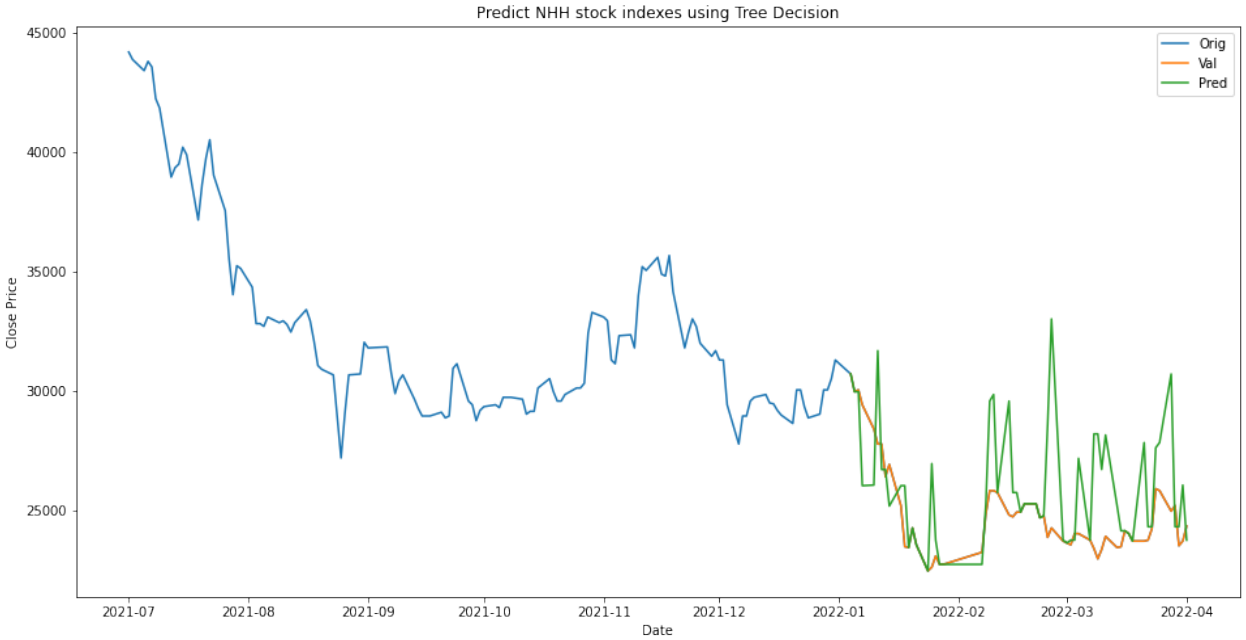

Figure 60: Predicting NHH stock using Decision Tree

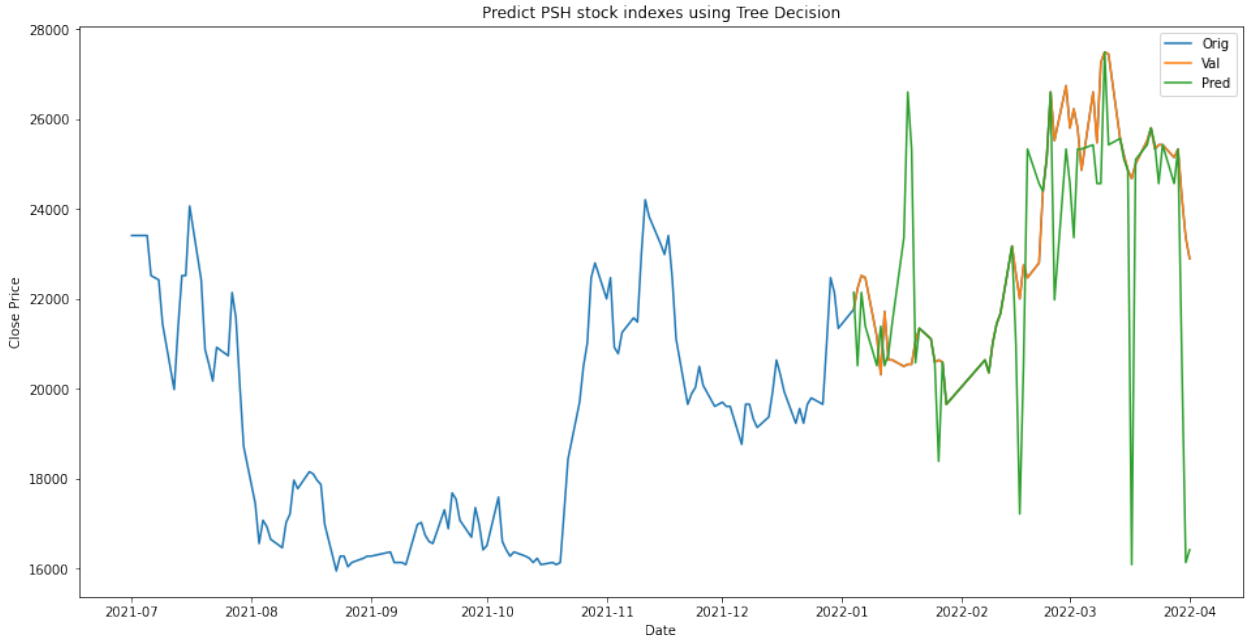

Figure 61: Predicting PSH stock using Decision Tree

5.3.10 Predictions of 3 stocks with the highest growth

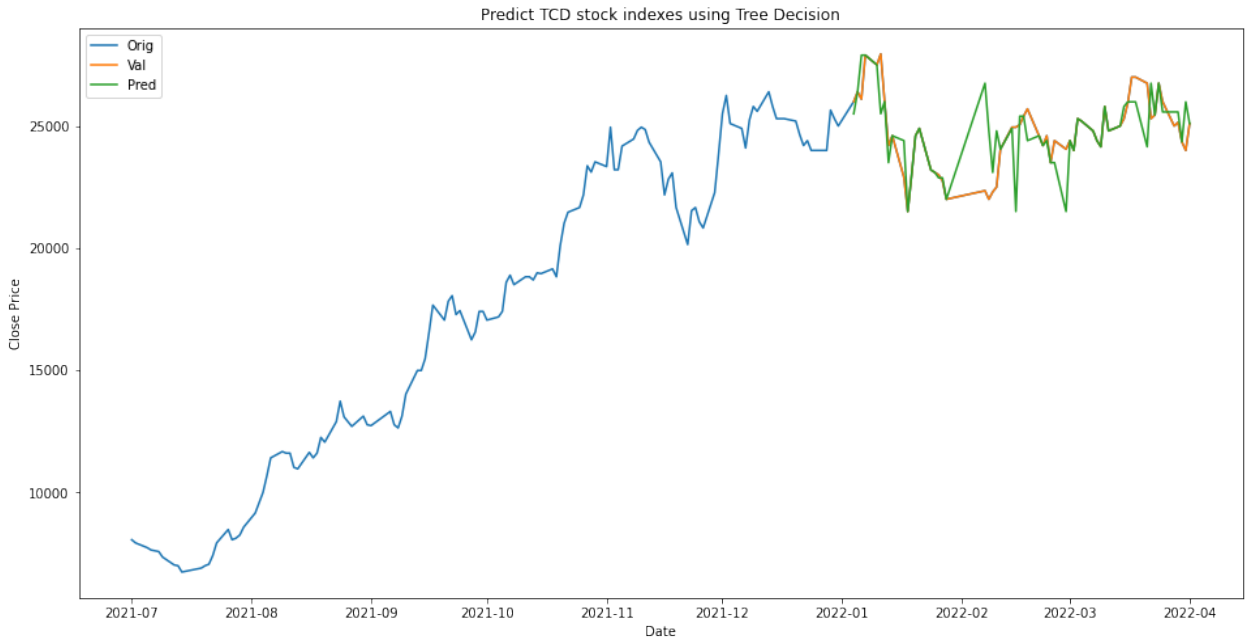

Figure 62: Predicting TCD stock using Decision Tree

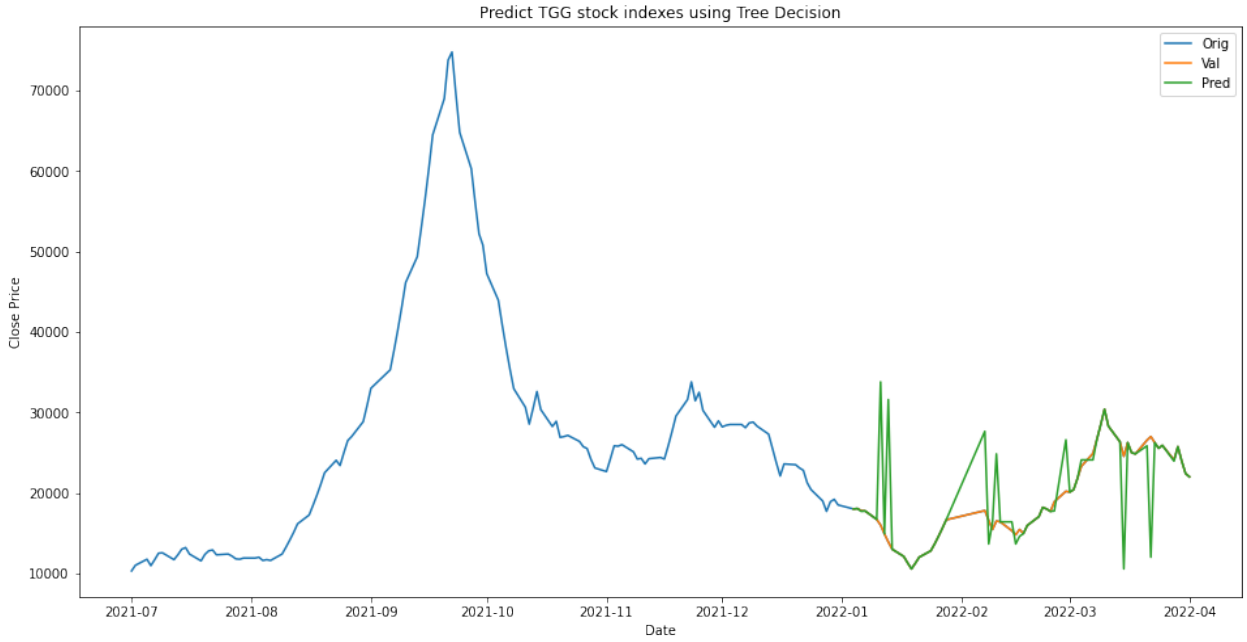

Figure 63: Predicting TGG stock using Decision Tree

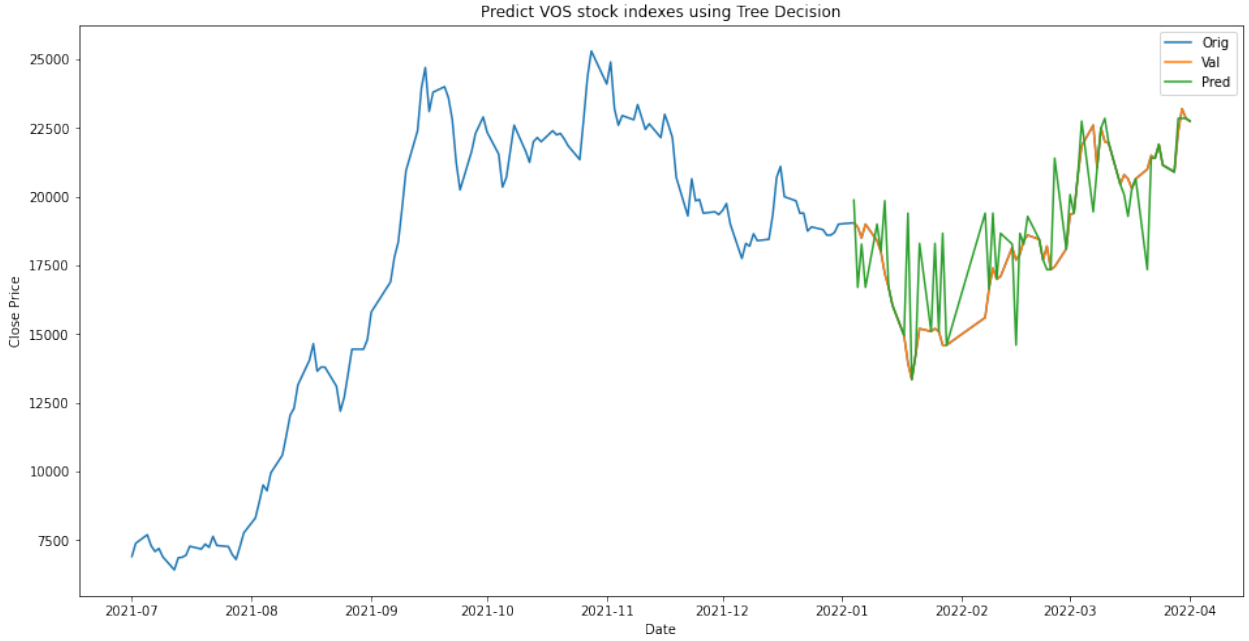

Figure 64: Predicting VOS stock using Decision Tree

5.4 Autoregressive integrated moving average (ARIMA) model

5.4.1 ARIMA summary

What is ARIMA model ?

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to predict future values in the series (forecasting) in order to generate a better fitting for the value of the series. Due to being fitted to time series data, ARIMA model is used in statistics and econometrics, and in particular, in time series analysis.

ARIMA Analysis ?

(A-R-I-M-A) This abbreviation is descriptive and captures the salient features of the model. In a nutshell, they are:

- **Autoregression (AR)**: refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
- **Integrated (I)**: represents the differencing of raw observations to allow for the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).
- **Moving average (MA)**: incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

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, where 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 included in the model, also called the lag order.
- d: The number of times that the raw observations are differenced, also called the degree of differencing.
- q: The size of the moving average window, also called the order of moving average.

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

In an autoregressive integrated moving average model, the data are differenced in order to make it stationary. A model that shows stationarity is one that shows there is constancy to the data over time. Most economic and market data show trends, so the purpose of differencing is to remove any trends or seasonal structures.

Seasonality, or when data show regular and predictable patterns that repeat over a calendar year, could negatively affect the regression model. If a trend appears and stationarity is not evident, many of the computations throughout the process cannot be made with great efficacy.

Special Considerations: ARIMA models are based on the assumption that past values have some residual effect on current or future values. For example, an investor using an autoregressive model to predict the performance of U.S. financial stocks would have had good reason to predict an ongoing trend of stable or rising stock prices. But once it became public knowledge that many financial institutions were at risk of imminent collapse, investors suddenly became less concerned with these stocks' recent prices and far more concerned with their underlying risk exposure. Therefore, the market rapidly revalued financial stocks to a much lower level, a move that would have utterly confounded an autoregressive model.

5.4.2 ARIMA and its algorithm

The model will represent multiple linear regression equations of the input variables (also known as dependent variables in statistics) as 3 main components: AR model, MA model and the other is Non-seasonal ARIMA model.

- **Autoregression (AR) model**:

  In a multiple regression model, we forecast the variable of interest using a linear combination of predictors. In an autoregression model, we forecast the variable of interest using a linear combination

of past values of the variable. The term autoregression indicates that it is a regression of the variable against itself. Thus, an autoregressive model of order p can be written as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t,$$

where:

c is the intercept.

$\phi_1, \phi_2, ..., \phi_p$ are the coefficients.

ε_t is white noise (error).

This is like a multiple regression but with lagged values of y_t as predictors and known as an $AR(p)$ model, which stands for an autoregressive model of order p.

Autoregressive models are remarkably flexible at handling a wide range of different time series patterns. The two series in Figure 65 show series from an AR(1) model and an AR(2) model. Changing the coefficients $\phi_1, \phi_2, ..., \phi_p$ results in different time series patterns. The variance of the error term ε_t will only change the scale of the series, not the patterns.

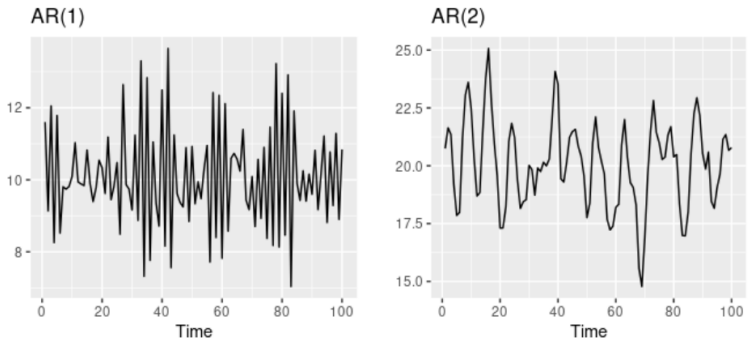

Figure 65: Two examples of data from auto regressive models with different parameters. Left: AR(1) with $y_t = 18 - 0.8y_{t-1} + \varepsilon_t$. Right: AR(2) with $y_t = 8 + 1.3y_{t-1} - 0.7y_{t-2} + \varepsilon_t$. In both cases, ε_t is normally distributed white noise with mean zero and variance one.

For AR(1) model:

- When $\phi_1 = 0, y_t$ is equivalent to white noise.
- When $\phi_1 = 1$ and $c = 0$, y_t is equivalent to a random walk.
- When $\phi_1 = 1$ and $c \neq 0$, y_t is equivalent to a random walk with drift.
- When $\phi_1 < 0, y_t$ tends to oscillate around the mean.

We normally restrict auto regressive models to stationary data, in which case some constraints on the values of the parameters are required.

- For AR(1) model: $-1 < \phi_1 < 1$.
- For AR(2) model: $-1 < \phi_2 < 1, \phi_2 + \phi_1 < 1, \phi_2 - \phi_1 < 1$.

When $p \geqslant 3$, the restrictions are much more complicated.

- **Moving average (MA)**:

  Rather than using past values of the forecast variable in a regression, a moving average model uses past forecast errors in a regression-like model.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q},$$

where:

c is the intercept.

$\theta_1, \theta_2, ..., \theta_p$ are the coefficients.

ε_t is white noise (error); ε_{t-1}, ε_{t-2}, ..., ε_{t-q} are lagged errors.

This is referred to as an $MA(q)$ model, or a moving average model of order q. Of course, we do not observe the values of ε_t, so it is not really a regression in the usual sense.

Notice that each value of y_t can be thought of as a weighted moving average of the past few forecast errors. However, moving average models should not be confused with moving average smoothing. A moving average model is used for forecasting future values while moving average smoothing is used for estimating the trend cycle of past values.

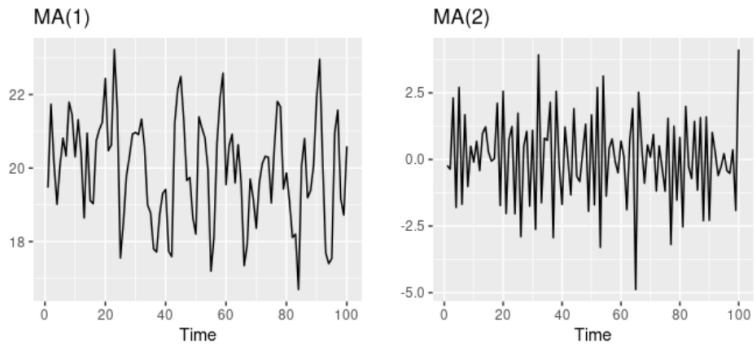

Figure 66: Two examples of data from moving average models with different parameters. Left: MA(1) with $y_t = 20 + \varepsilon_t + 0.8\varepsilon_{t-1}$. Right: MA(2) with $\varepsilon_t - \varepsilon_{t-1} + 0.8\varepsilon_{t-2}$. In both cases, ε_t is normally distributed white noise with mean zero and variance one.

Figure 66 shows some data from an MA(1) model and an MA(2) model. Changing the coefficients $\theta_1, \theta_2, ..., \theta_p$ results in different time series patterns. As with auto regressive models, the variance of the error term ε_t will only change the scale of the series, not the patterns.

It is possible to write any stationary $AR(p)$ model as an $MA(\infty)$ model. For instance, using repeated substitution, we can demonstrate this for an AR(1) model:

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \varepsilon_t \\ &= \phi_1(\phi_1 y_{t-2} + \varepsilon_{t-2}) + \varepsilon_t \\ &= \phi_1^2 y_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= \phi_1^3 y_{t-2} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &etc. \end{aligned}$$

Provided $-1 < \phi_1 < 1$, the value of ϕ_1^k will get smaller as k gets larger. So eventually we obtain

$$y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \phi_1^3 \varepsilon_{t-3} + ...,$$

an MA(∞) process.

The reverse result holds if we impose some constraints on the MA parameters. Then the MA model is called invertible. That is, we can write any invertible $MA(q)$ process as an $AR(\infty)$ process.

Invertible models are not simply introduced to enable us to convert from MA models to AR models. They also have some desirable mathematical properties.

For instance, consider the MA(1) process, $y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1}$. In its $AR(\infty)$ representation, the most recent error can be written as a linear function of current and past observations:

$$\varepsilon_t = \sum_{j=0}^{\infty} (-\theta)^j y_{t-j}.$$

When $|\theta| > 1$, the weights increase as lags increase, so the more distant the observations the greater their influence on the current error. When $|\theta| = 1$, the weights are constant in size, and the distant observations have the same influence as the recent observations. As neither of these situations makes much sense, we require $|\theta| < 1$, so the most recent observations have a higher weight than observations from the more distant past. Thus, the process is invertible when $|\theta| < 1$.

The invertibility constraints for other models are similar to the stationarity constraints.

- For MA(1) model: $-1 < \phi_1 < 1$.
- For MA(2) model: $-1 < \phi_2 < 1, \phi_1 + \phi_2 > -1, \phi_1 - \phi_2 < 1$.

More complicated conditions hold for $q \geqslant 3$.

- **Non-seasonal ARIMA models**:

In this context, "Integration" is the reverse of differencing. If we combine differencing with autoregression and a moving average model, we obtain a non-seasonal ARIMA model. The full model can be written as

$$y'_t = c + \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + ... + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_2 + ... + \theta_1 \varepsilon_{t-1} + \varepsilon_t, (1)$$

where:

y'_t is the differenced series (it may have been differenced more than once).

The "predictors" on the right-hand side include both lagged values of y_t and lagged errors. As we have said, this an **ARIMA** (p, d, f) **model** where:

- p: order of the autoregressive part.
- d: degree of first differencing involved.
- q: order of the moving average part.

The same stationarity and invertibility conditions that are used for autoregressive and moving average models also apply to an ARIMA model.

Many of the models we have already discussed are special cases of the ARIMA model, as shown in the following table.

Table of some special cases of ARIMA models.

White noise	ARIMA$(0, 0, 0)$
Random walk	ARIMA$(0, 1, 0)$ with no constant
Random walk with drift	ARIMA$(0, 1, 0)$ with a constant
Autoregression	ARIMA$(p, 0, 0)$
Moving average	ARIMA$(0, 0, q)$

Once we start combining components in this way to form more complicated models, it is much easier to work with the backshift notation. For instance, equation above can be written in backshift notation as

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p)\ (1 - B)^d y_t = c + (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_p B^p)\varepsilon_t$$

With:

$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) \rightarrow AR(p)$

$(1 - B)^d y_t \rightarrow$ d differences

$c + (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_p B^p)\varepsilon_t \rightarrow MA(q)$

5.4.3 Prediction of VNINDEX

```
!pip install pmdarima
```

```
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import datetime
```

```
path = r"/content/drive/MyDrive/Colab Notebooks/Data/VNI
(1_7_2021-31_12_2021).xlsx"
data = pd.read_excel(path)
stock_train = data.loc[:129]
stock_predict = data.loc[129:]
```

```
data_train_length = len(stock_train)
data_train = stock_train.loc[0:data_train_length-1];
data_train_frame = pd.DataFrame(data_train['Close'].to_numpy(), index =
stock_train.loc[0:data_train_length-1]['Date'], columns=['Price'])
data_predict_frame = pd.DataFrame(stock_predict['Close'].to_numpy(),index =
stock_predict['Date'], columns=['Price'])
```

```
plt.figure(figsize=(16,8))
plt.title('Train data')
plt.xlabel('Date')
plt.ylabel('Close Price (VND)')
plt.plot(data_train_frame)
```

```
stepwise_model = auto_arima(data_train_frame['Price'], test='adf',
                        max_p=3, max_q=3,
                        m=1,
                        d=None,
                        seasonal=False,
                        start_P=0,
                        D=0,
                        trace=True,
                        error_action='ignore',
                        suppress_warnings=True,
                        stepwise=True)
```

```
stepwise_model.fit(data_train_frame)
```

```
future_data = stepwise_model.predict(n_periods=len(data_predict_frame))
future_data_frame = pd.DataFrame(future_data.to_numpy(),
index=data_predict_frame.index, columns = ['Price'])
```

```
plt.figure(figsize=(16,8))
plt.title('NHH')
plt.xlabel('Date')
plt.ylabel('Close Price (VND)')
plt.plot(data_train_frame, color='blue', label='Actual')
plt.plot(data_predict_frame, color='blue')
plt.plot(future_data_frame, color='red', label='Predict future')
plt.legend(loc='lower right', fontsize='x-large')
```

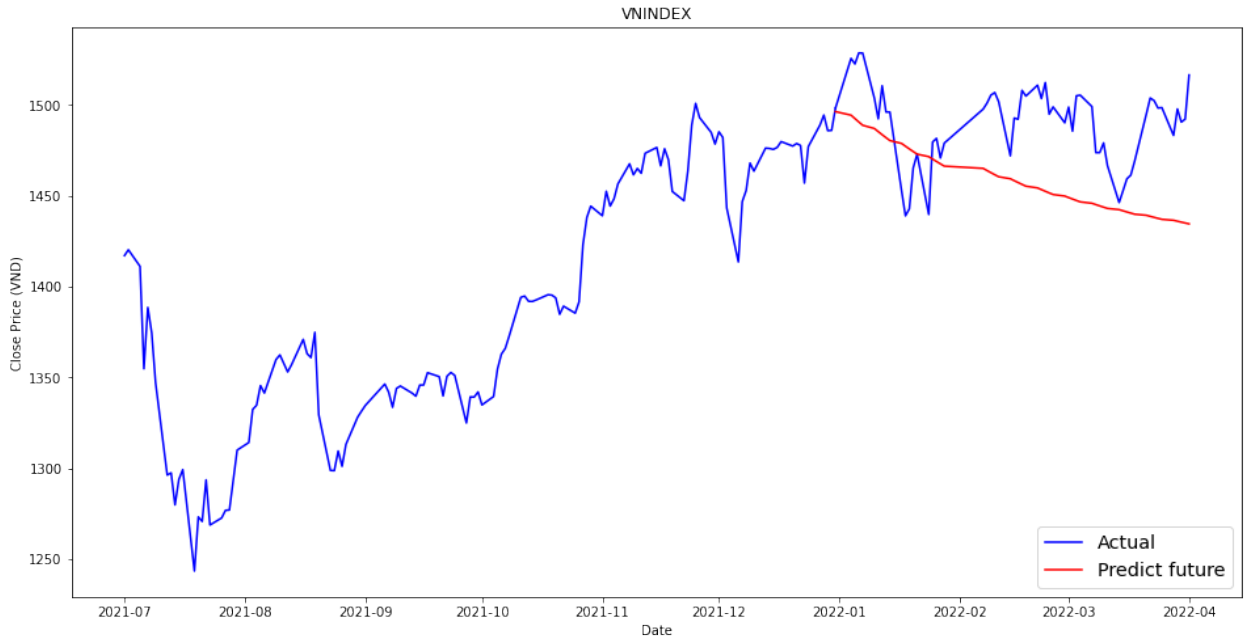

Figure 67: Predicting VNINDEX stock using ARIMA

5.4.4 Predictions of 3 stocks in three different sectors

Figure 68: Predicting CII stock using ARIMA

Figure 69: Predicting SKG stock using ARIMA

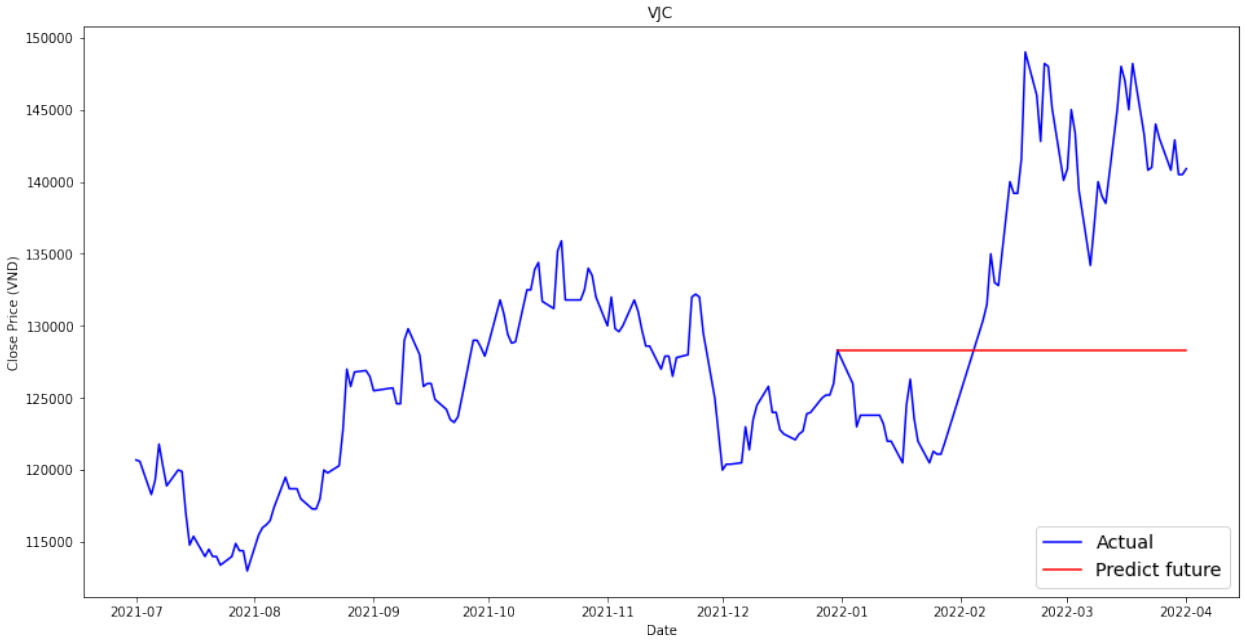

Figure 70: Predicting VJC stock using ARIMA

5.4.5 Predictions of 3 stocks with the strongest drop

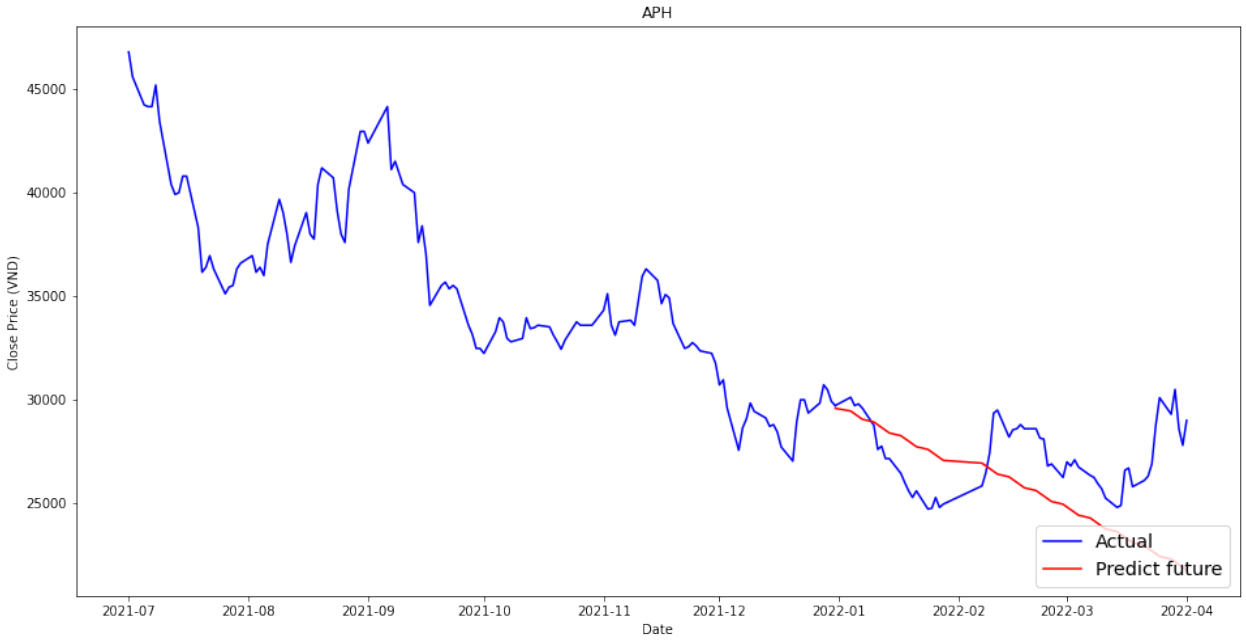

Figure 71: Predicting APH stock using ARIMA

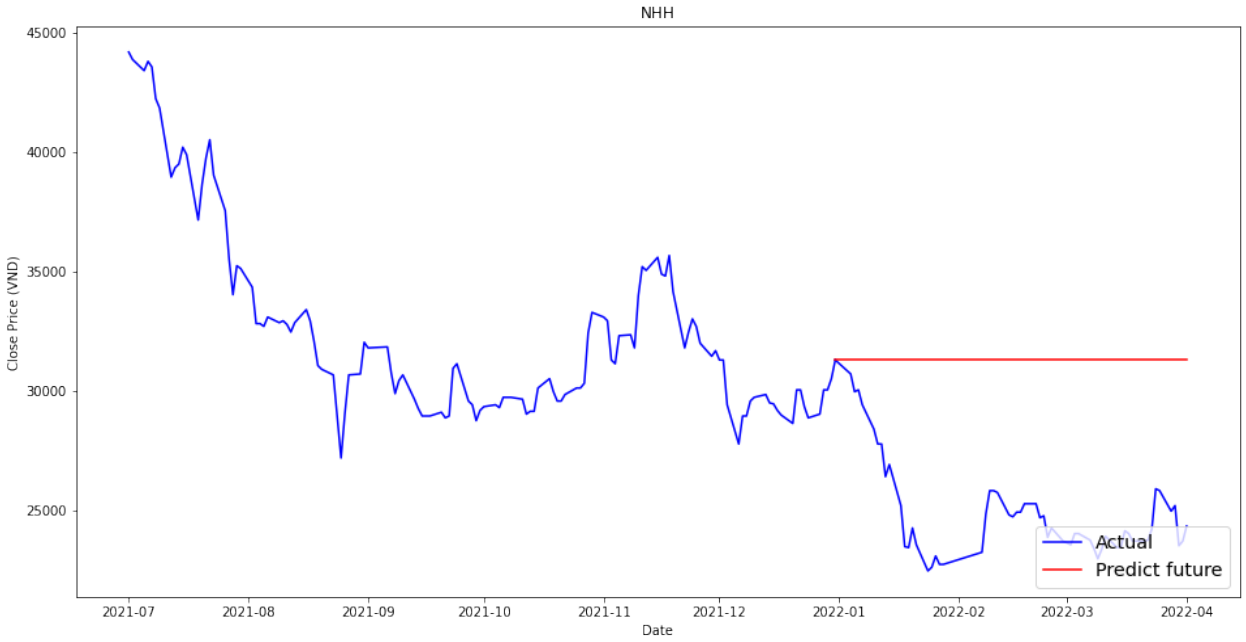

Figure 72: Predicting NHH stock using ARIMA

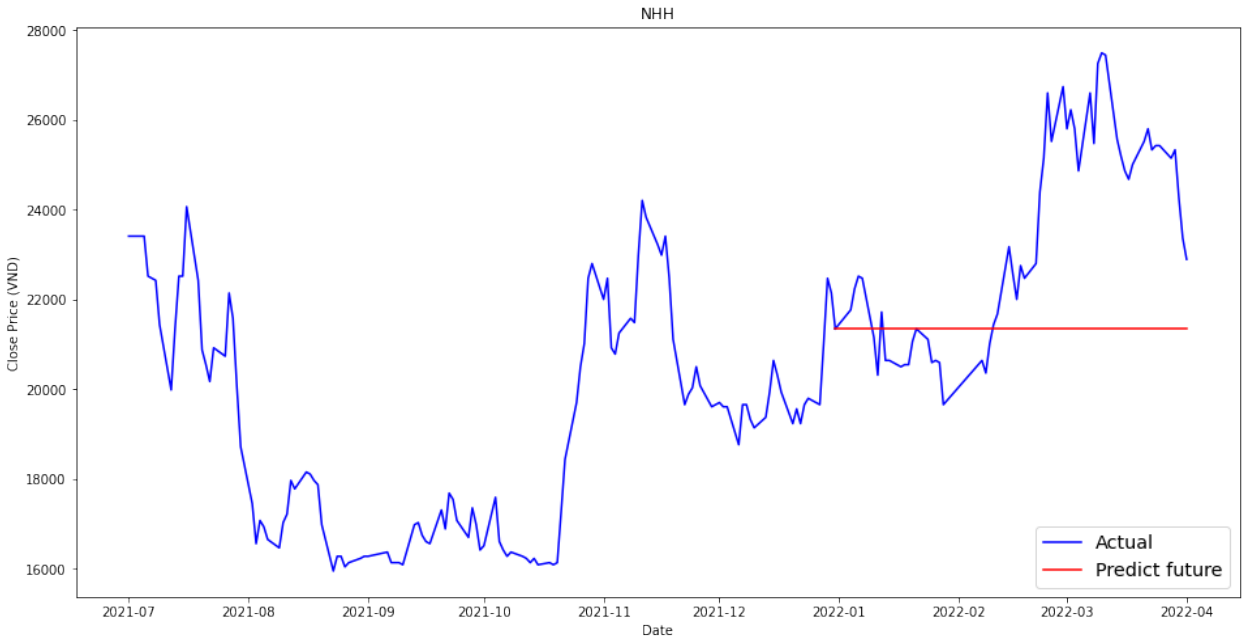

Figure 73: Predicting PSH stock using ARIMA

5.4.6 Predictions of 3 stocks with the highest growth

Figure 74: Predicting TCD stock using ARIMA

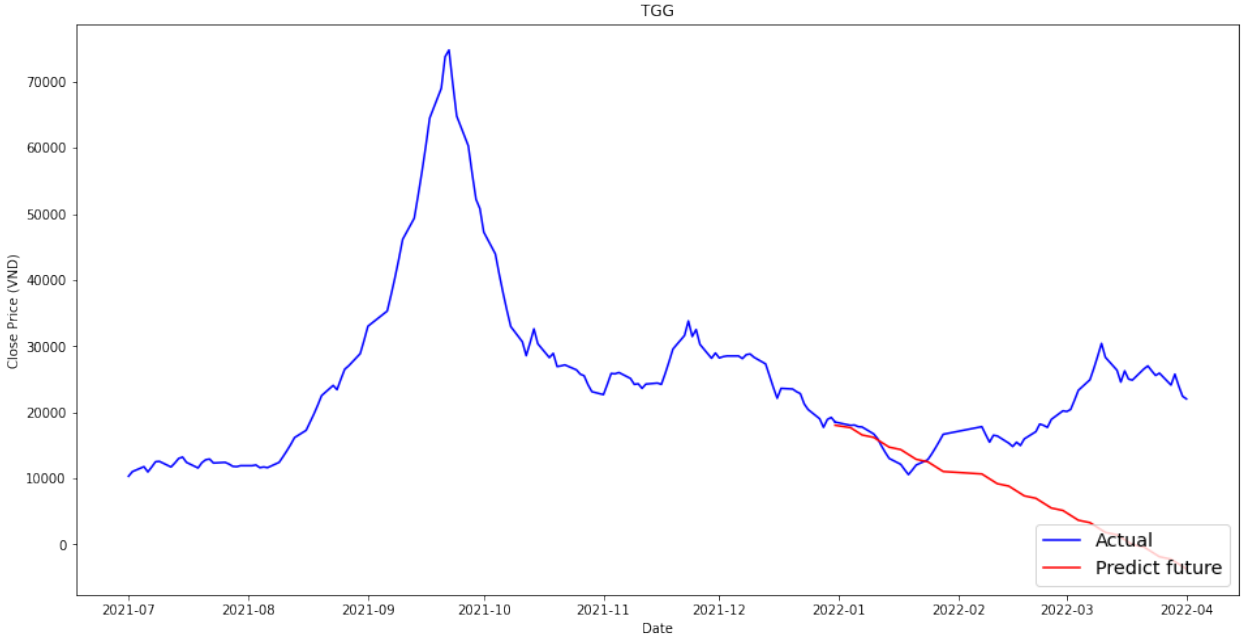

Figure 75: Predicting TGG stock using ARIMA

Figure 76: Predicting VOS stock using ARIMA

5.5 LSTM model

5.5.1 Neural network architecture

A neural network can be thought of as a network of "neurons" which are organised in layers. The predictors (or inputs) form the bottom layer, and the forecasts (or outputs) form the top layer. There may also be intermediate layers containing "hidden neurons".

The simplest networks contain no hidden layers and are equivalent to linear regressions. Figure 11.11 shows the neural network version of a linear regression with four predictors. The coefficients attached to these predictors are called "weights". The forecasts are obtained by a linear combination of the inputs. The weights are selected in the neural network framework using a "learning algorithm" that minimizes a "cost function" such as the MSE. Of course, in this simple example, we can use linear regression which is a much more efficient method of training the model.

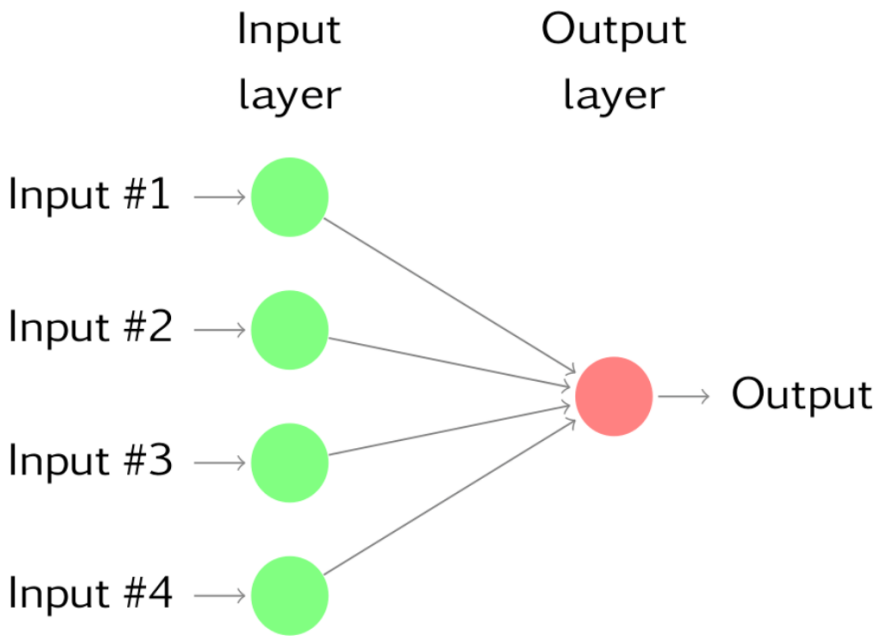

Figure 77: A simple neural network

Once we add an intermediate layer with hidden neurons, the neural network becomes non-linear.

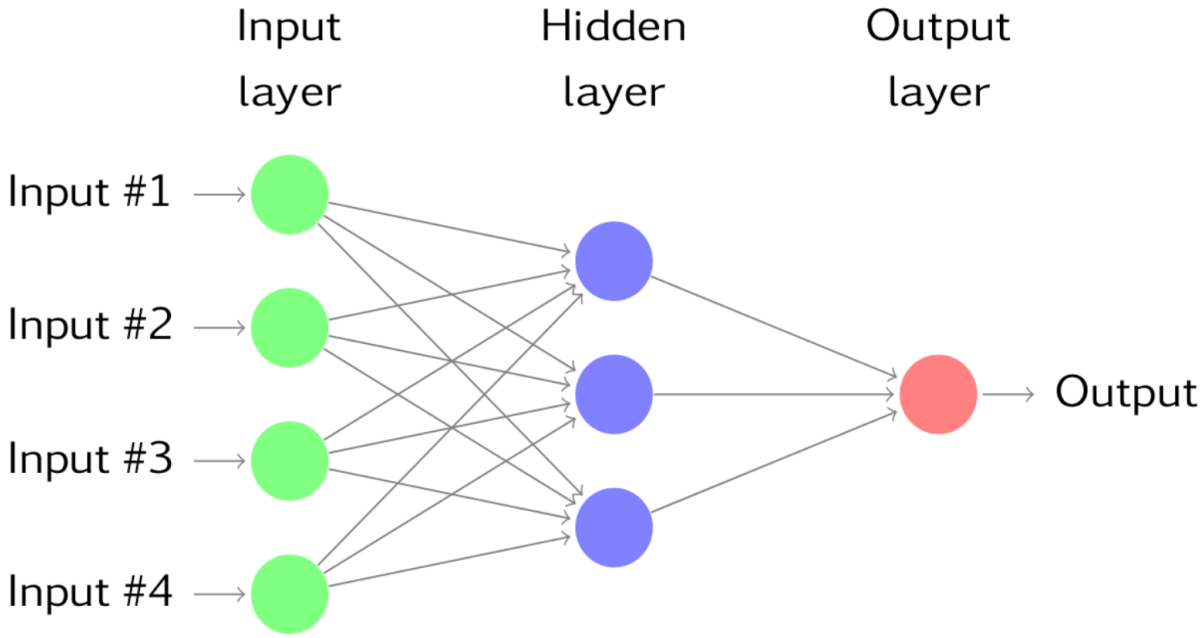

Figure 78: A simple neural network

This is known as a multilayer feed-forward network, where each layer of nodes receives inputs from the previous layers. The outputs of the nodes in one layer are inputs to the next layer. The inputs to each node are combined using a weighted linear combination. The result is then modified by a nonlinear function before being output.

5.5.2 Recurrent Neural Network model (RNN)

Recurrent Neural Network is a generalization of feedforward neural network that has an internal memory. RNN is recurrent in nature as it performs the same function for every input of data while the

output of the current input depends on the past one computation. After producing the output, it is copied and sent back into the recurrent network. For making a decision, it considers the current input and the output that it has learned from the previous input.

Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. In other neural networks, all the inputs are independent of each other. But in RNN, all the inputs are related to each other.

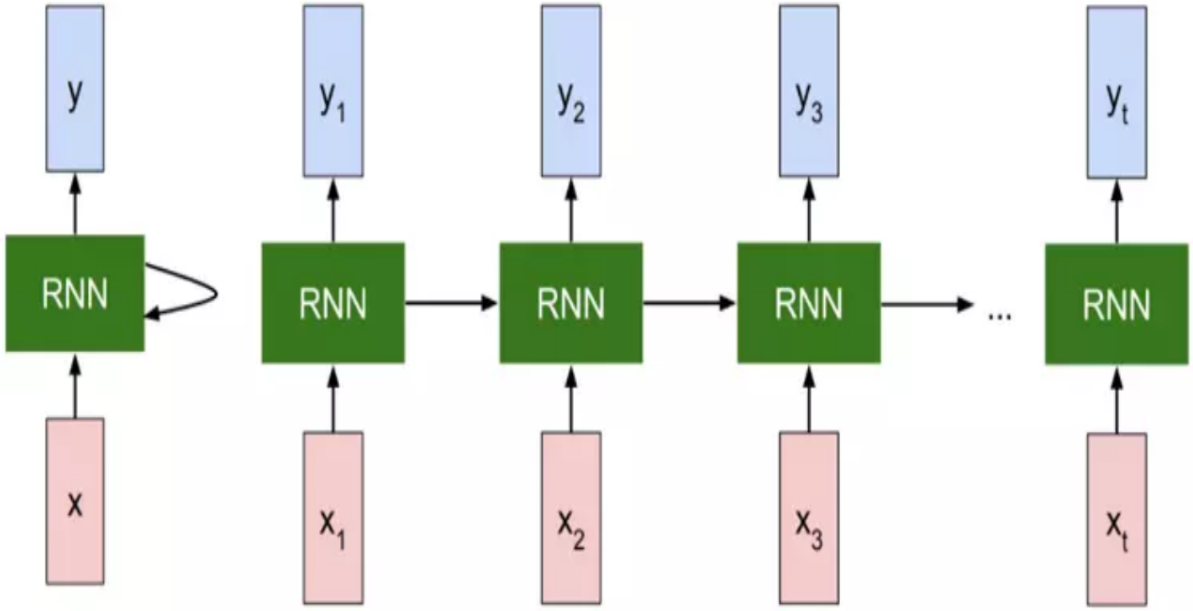

Figure 79: Recurrent Neural Network

First, it takes the X(0) from the sequence of input and then it outputs h(0) which together with X(1) is the input for the next step. So, the h(0) and X(1) is the input for the next step. Similarly, h(1) from the next is the input with X(2) for the next step and so on. This way, it keeps remembering the context while training.

5.5.3 Long Short Term Memory Model(LSTM)

LSTM networks were designed specifically to overcome the long-term dependency problem faced by recurrent neural networks RNNs (due to the vanishing gradient problem). LSTMs have feedback connections which make them different to more traditional feedforward neural networks. This property enables LSTMs to process entire sequences of data (e.g. time series) without treating each point in the sequence independently, but rather, retaining useful information about previous data in the sequence to help with the processing of new data points. As a result, LSTMs are particularly good at processing sequences of data such as text, speech and general time-series.

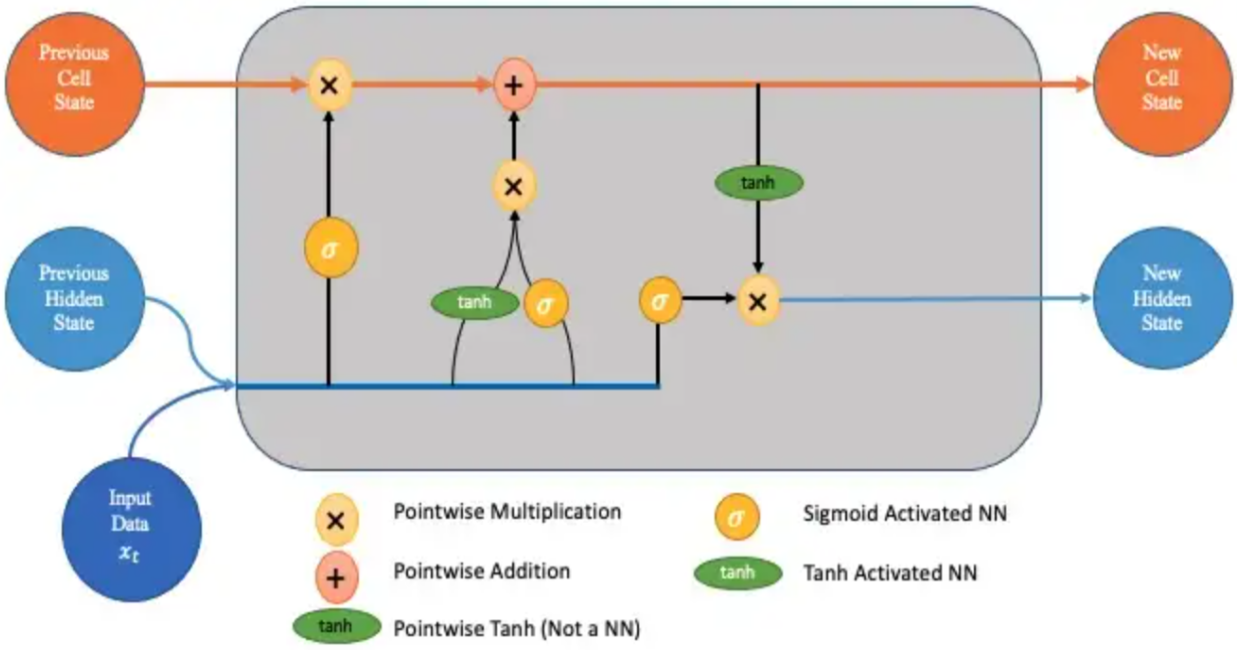

Figure 80: LSTM model

LSTMs use a series of 'gates' which control how the information in a sequence of data comes into, is stored in and leaves the network. There are three gates in a typical LSTM; forget gate, input gate and output gate. These gates can be thought of as filters and are each their own neural network. We will explore them all in detail during the course of this article.

Step 1:

The first step in the process is the forget gate. Here we will decide which bits of the cell state (long term memory of the network) are useful given both the previous hidden state and new input data. The data that the gate deems irrelevant will be multiplied with small weights (close to zero) so that it doesn't affect the next steps.

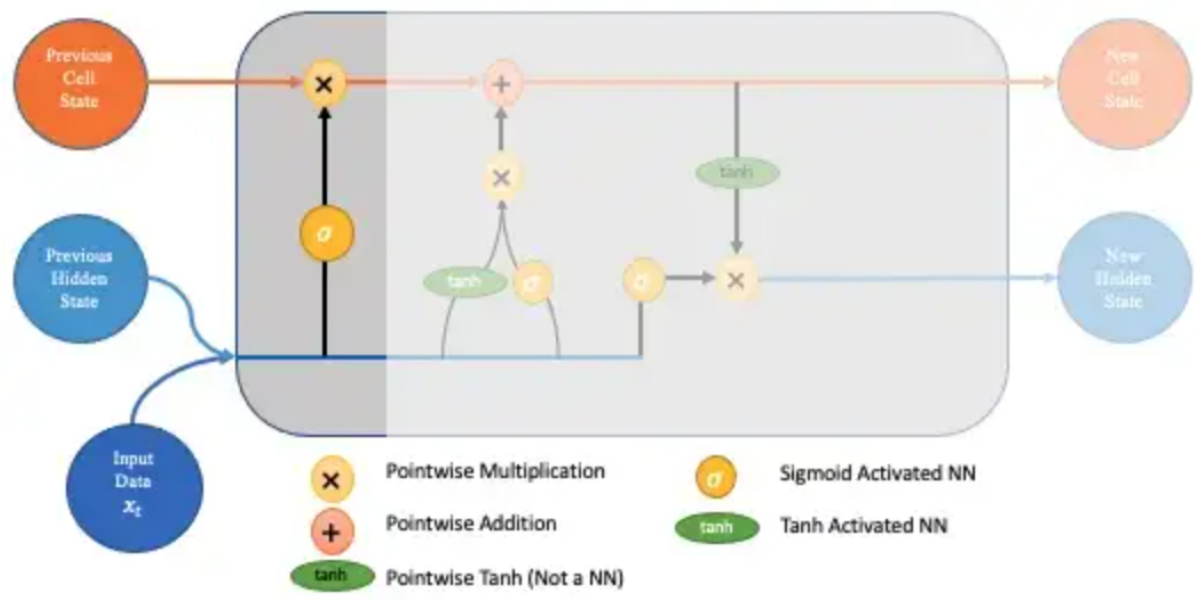

Figure 81: Forget gate

Step 2:

The next step involves the new memory network and the input gate. The goal of this step is to determine what new information should be added to the networks long-term memory (cell state), given the previous hidden state and new input data.

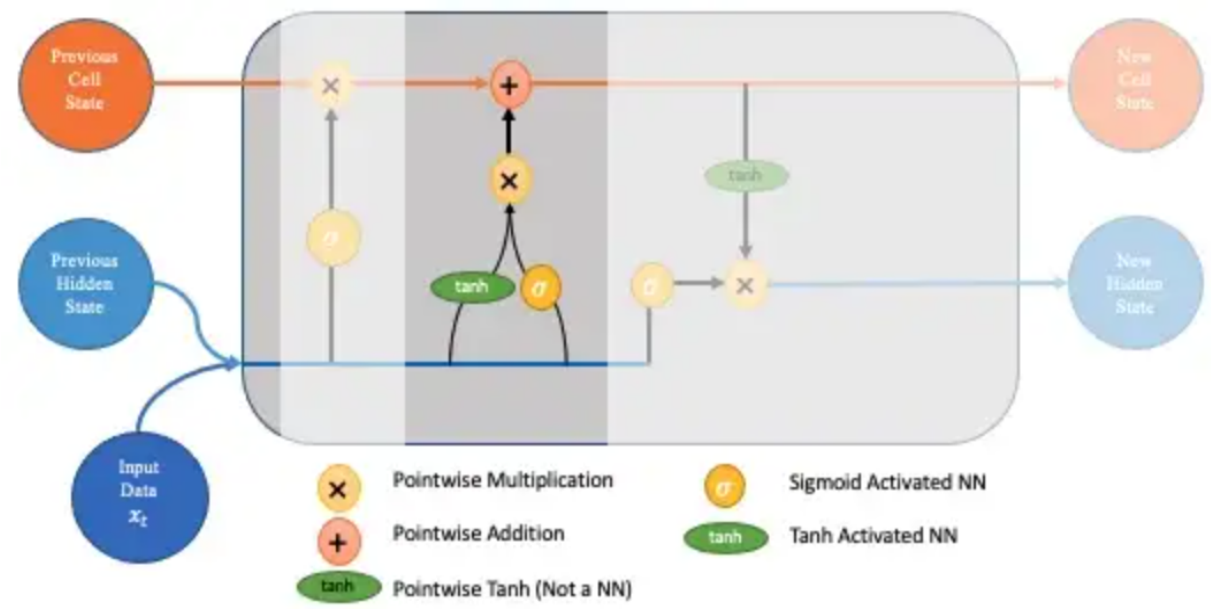

Figure 82: Input gate

Step 3:

Now that our updates to the long-term memory of the network are complete, we can move to the final step, the output gate, deciding the new hidden state. To decide this, we will use three things; the newly updated cell state, the previous hidden state and the new input data. In this step, it exports the result of three steps, the new cell state.

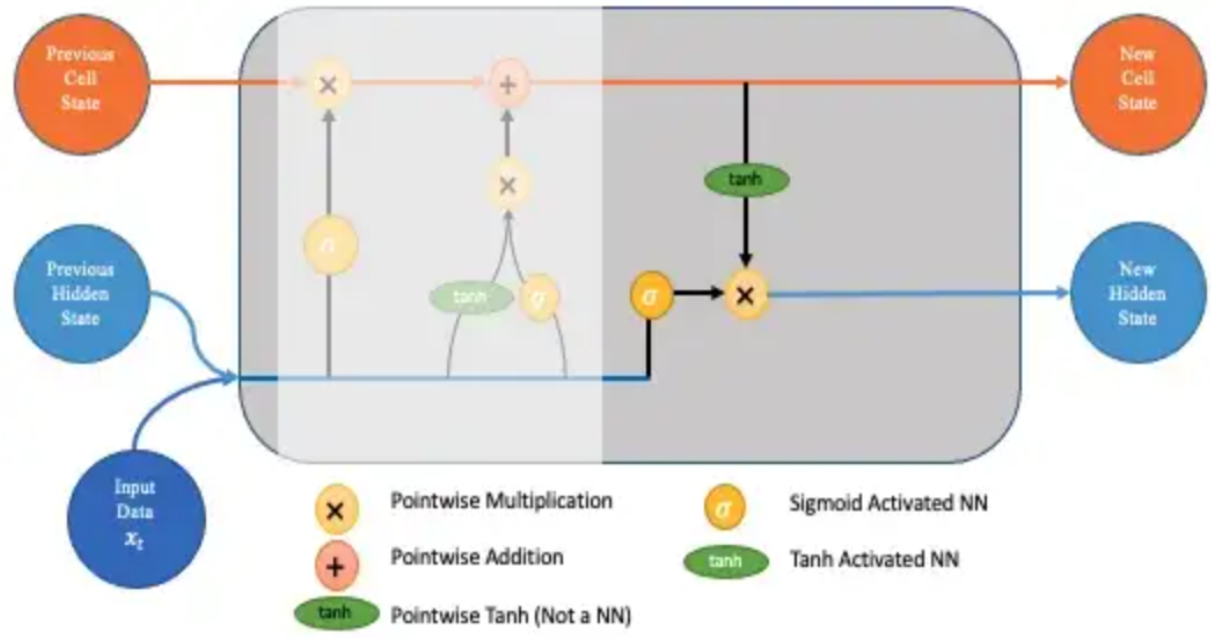

Figure 83: Output gate

In the real model, these steps above are repeated many times depend on the length of model's output.

Another thing we should consider is that: the output in the final step is still hidden state.And so, to convert the hidden state to the output, we actually need to apply a linear layer which only happens once at the very last step in the LSTM.

5.5.4 LSTM mode

```
import datetime
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
```

```
path = r"/content/drive/MyDrive/Colab Notebooks/Data/strongest drop/PSH.xlsx"
data = pd.read_excel(path)
stock_data = data.loc[:129]
stock_predict = data.loc[129:]
feed_data = data.loc[129-40:128]
```

```
stock_data_raw = stock_data['Close'].interpolate(method='linear').copy()
data_predict_raw = stock_predict['Close'].interpolate(method='linear').copy()
data_values = stock_data_raw.values
scaler = MinMaxScaler(feature_range=(0,1))
scaled_data = scaler.fit_transform(data_values.reshape(-1,1))
```

```
data_values = data_values.reshape(-1,1)
```

```
length_train = len(scaled_data)-20
window = 40
future_look = 20
```

```
feed_data = scaled_data[len(scaled_data)-window:]
x_feed = []
for i in range(0,int(len(feed_data))):
  x_feed.append(feed_data[i,0])
x_feed = np.array(x_feed)
x_feed = np.reshape(x_feed, (1,x_feed.shape[0],1))
```

```
x_train = []
y_train = []
for i in range(window+future_look, len(scaled_data)):
    x_train.append(scaled_data[i-window-future_look:i-future_look, 0])
    y_train.append(scaled_data[i-future_look:i, 0])
x_train, y_train = np.array(x_train), np.array(y_train)
x_train = np.reshape(x_train, (x_train.shape[0], x_train.shape[1], 1))
y_train = np.reshape(y_train, (y_train.shape[0], y_train.shape[1],1))
```

```
model = keras.Sequential()
model.add(layers.LSTM(100, return_sequences=True, input_shape=(x_train.shape[1],
1)))
model.add(layers.LSTM(100, return_sequences=False))
model.add(layers.Dense(25))
model.add(layers.Dense(20))
model.summary()
```

```
model.compile(optimizer='adam', loss='mean_squared_error')
model.fit(x_train, y_train, batch_size= 1, epochs=3)
```

```
def future_predict(length, predict = model.predict, feed = x_feed):
  y_future = []
  for i in range(0, length, 20):
    y = predict(feed)
    y_future = np.append(y_future,y[0])
    new_feed = np.append(x_feed[0],y[0])
    new_feed = new_feed[20:]
    feed = np.reshape(new_feed, (1, new_feed.shape[0],1))
  return y_future
```

```
future = future_predict(57)
```

```
future = np.reshape(future, (future.shape[0],1))
rescaled_future = scaler.inverse_transform(future)
```

```
stock_predict_frame =
pd.DataFrame(stock_predict['Close'].to_numpy(),index=stock_predict['Date'],
columns = ['Close'])
future_data_frame = pd.DataFrame(rescaled_future, index=stock_predict['Date'],
columns = ['Close'])
stock_data_frame = pd.DataFrame(stock_data['Close'].to_numpy(),
index=stock_data['Date'], columns = ['Close'])
```

```
plt.figure(figsize=(16,8))
plt.title('VOS')
plt.xlabel('Date')
plt.ylabel('Close Price')
plt.plot(future_data_frame, color='green', label='Predict')
plt.plot(stock_predict_frame, color='red', label='Actual')
plt.plot(stock_data_frame, color='blue', label = 'Train')
plt.legend(loc='upper left', shadow=False, fontsize='x-large')
```

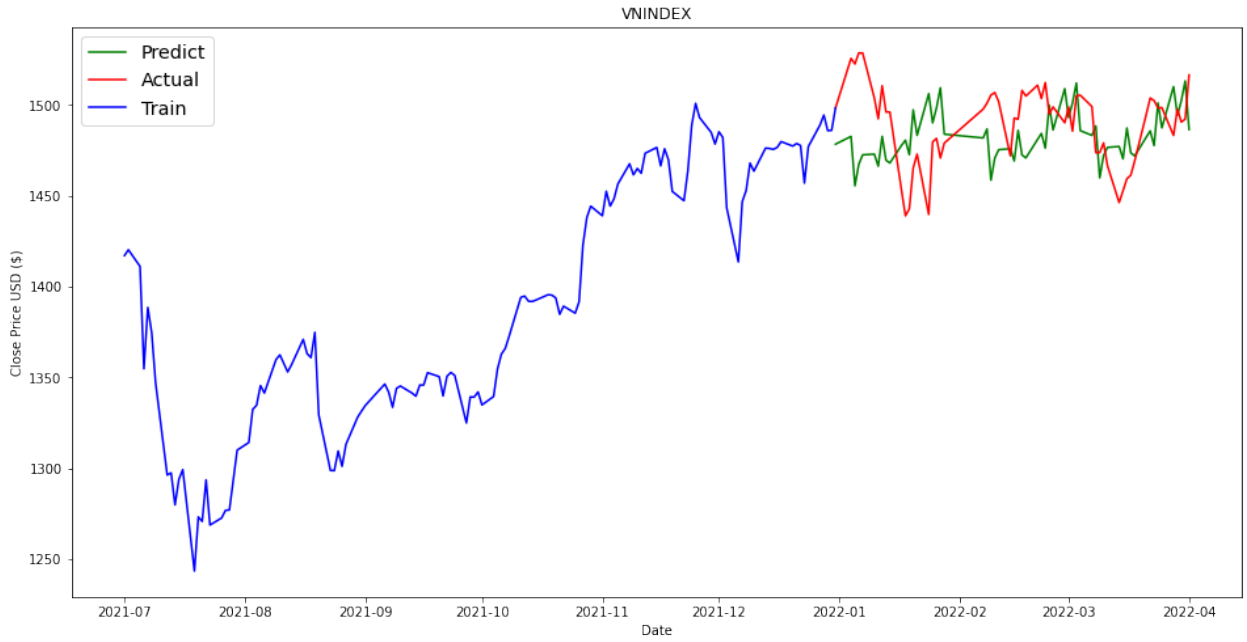

Figure 84: Predicting VNINDEX stock using LSTM

5.5.5 Predictions of 3 stocks in three different sectors

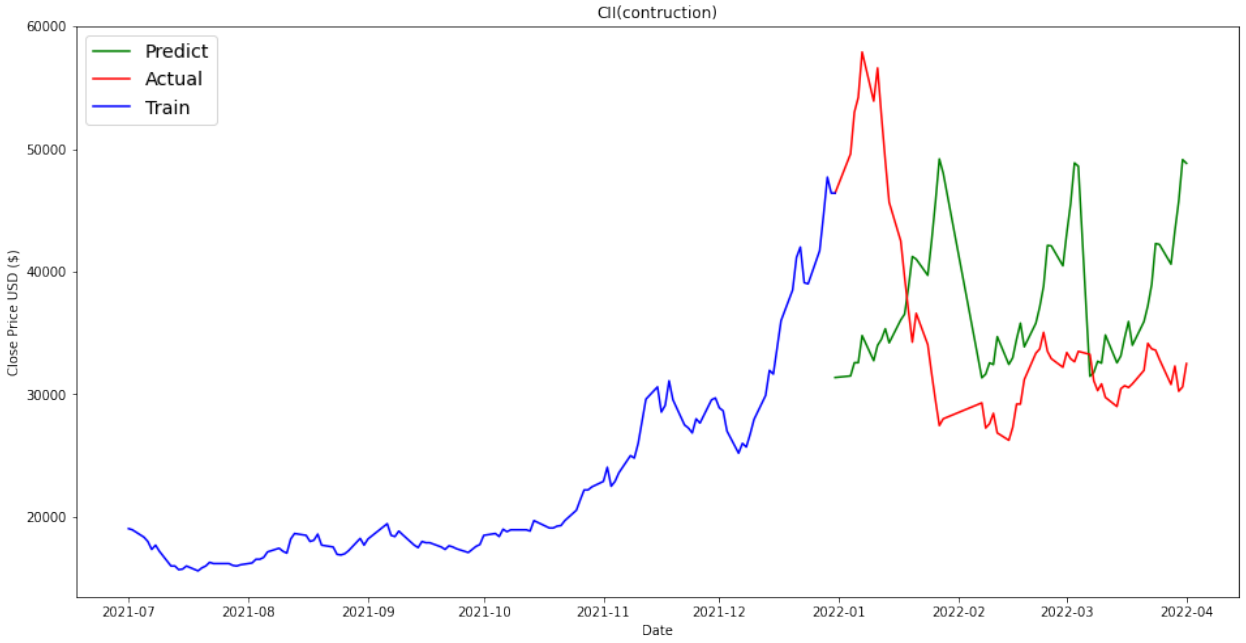

Figure 85: Predicting CII stock using LSTM

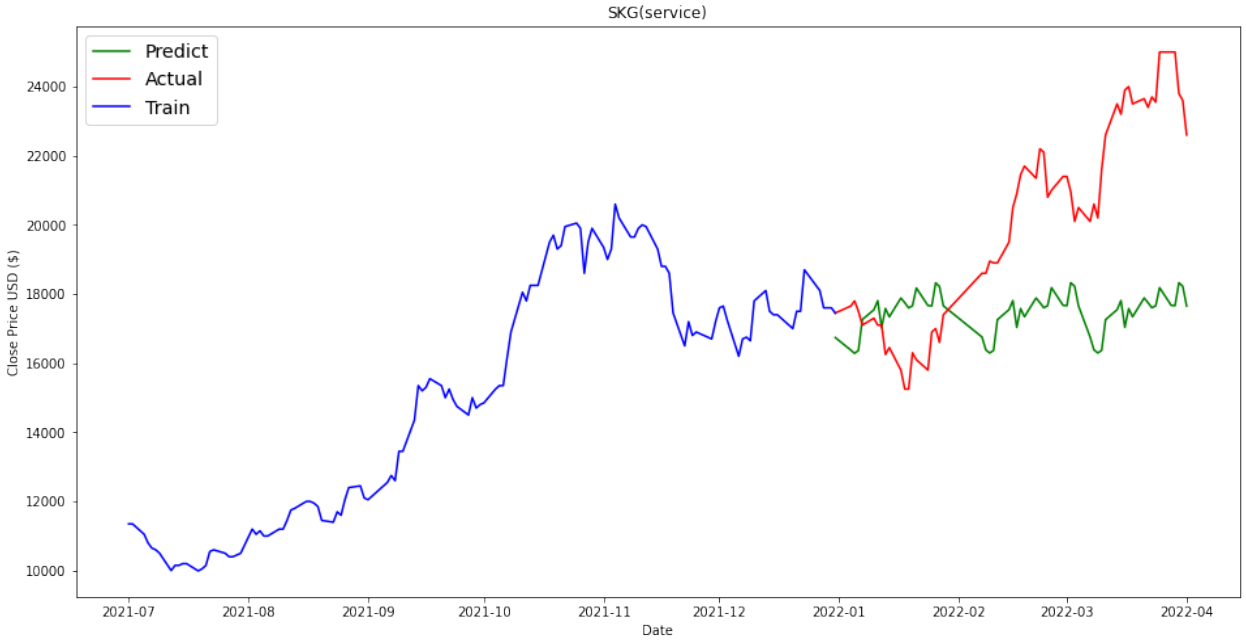

Figure 86: Predicting SKG stock using LSTM

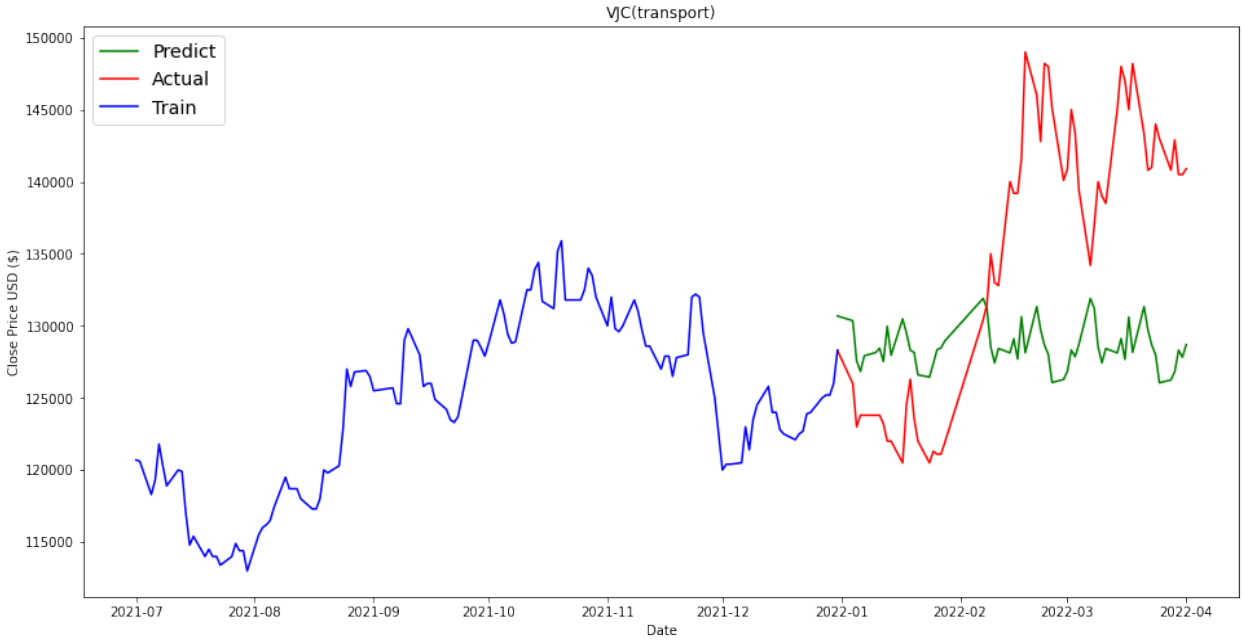

Figure 87: Predicting VJC stock using LSTM

5.5.6 Predictions of 3 stocks with the strongest drop

Figure 88: Predicting APH stock using LSTM

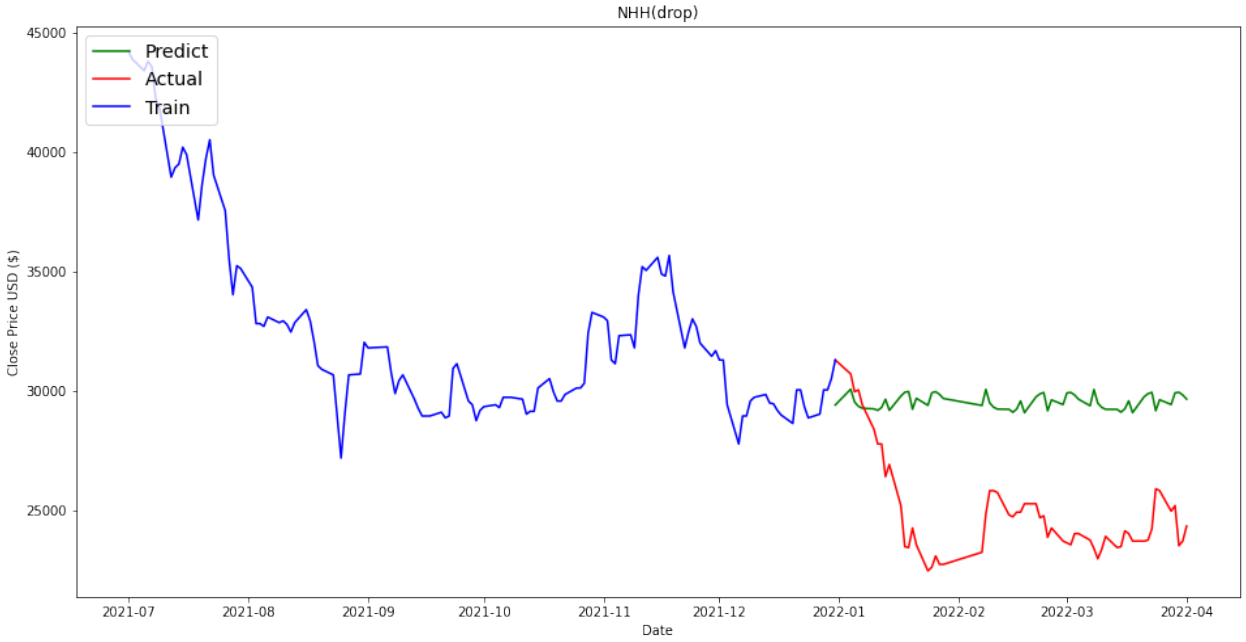

Figure 89: Predicting NHH stock using LSTM

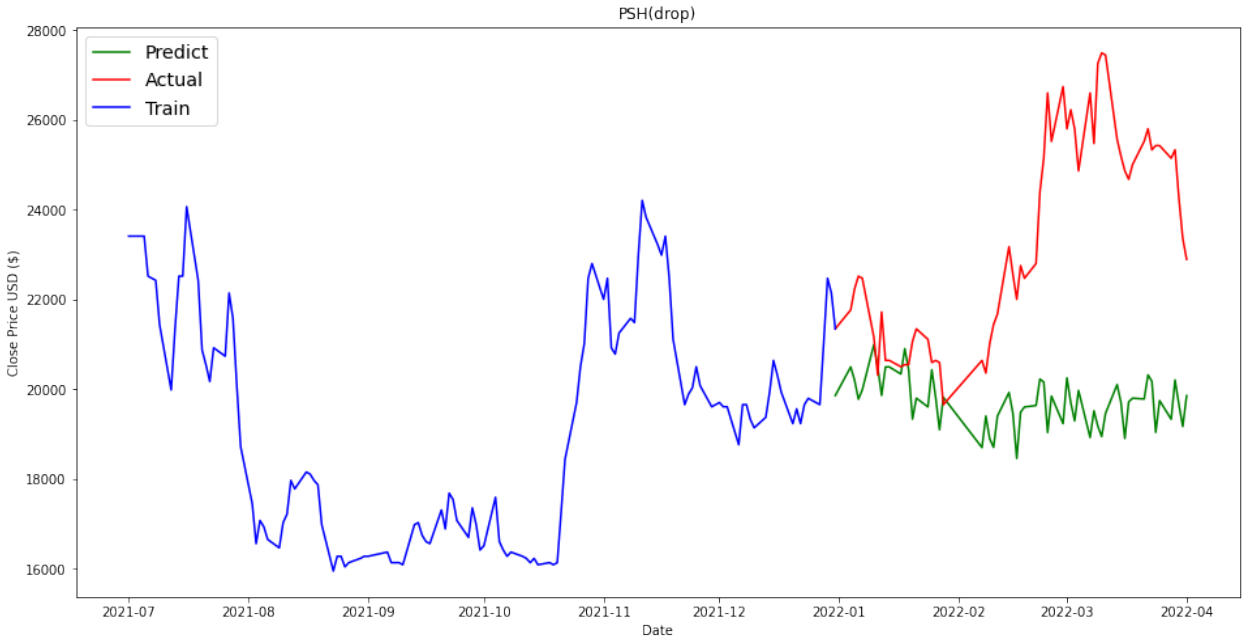

Figure 90: Predicting PSH stock using LSTM

5.5.7 Predictions of 3 stocks with the highest growth

Figure 91: Predicting TCD stock using LSTM

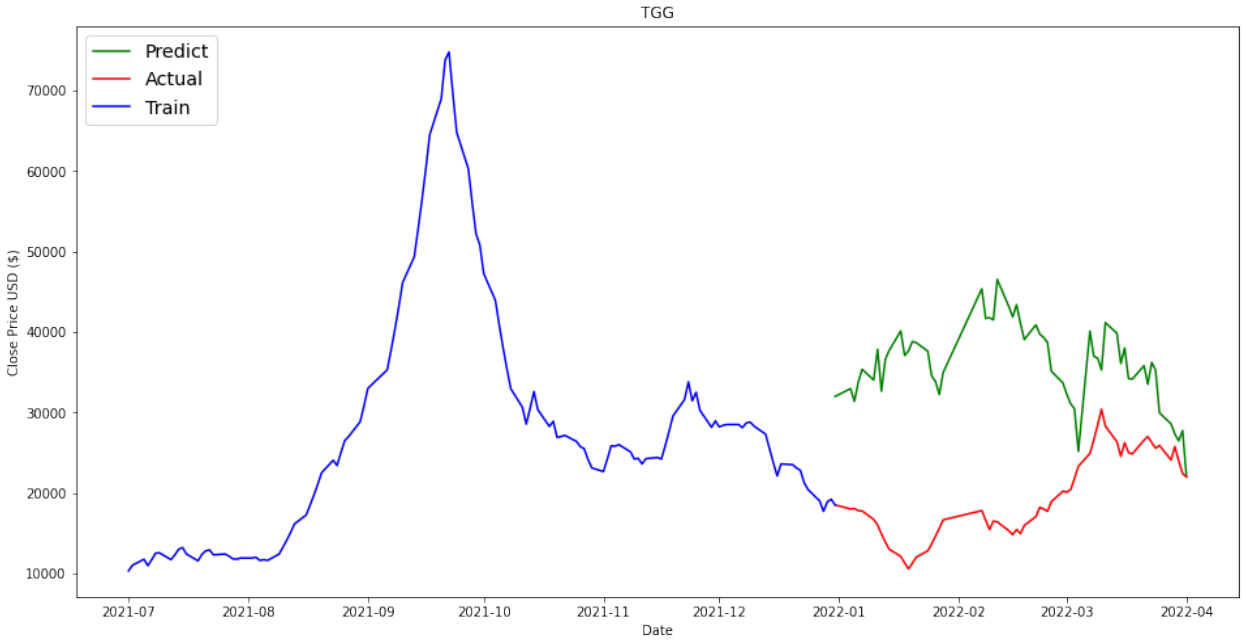

Figure 92: Predicting TGG stock using LSTM

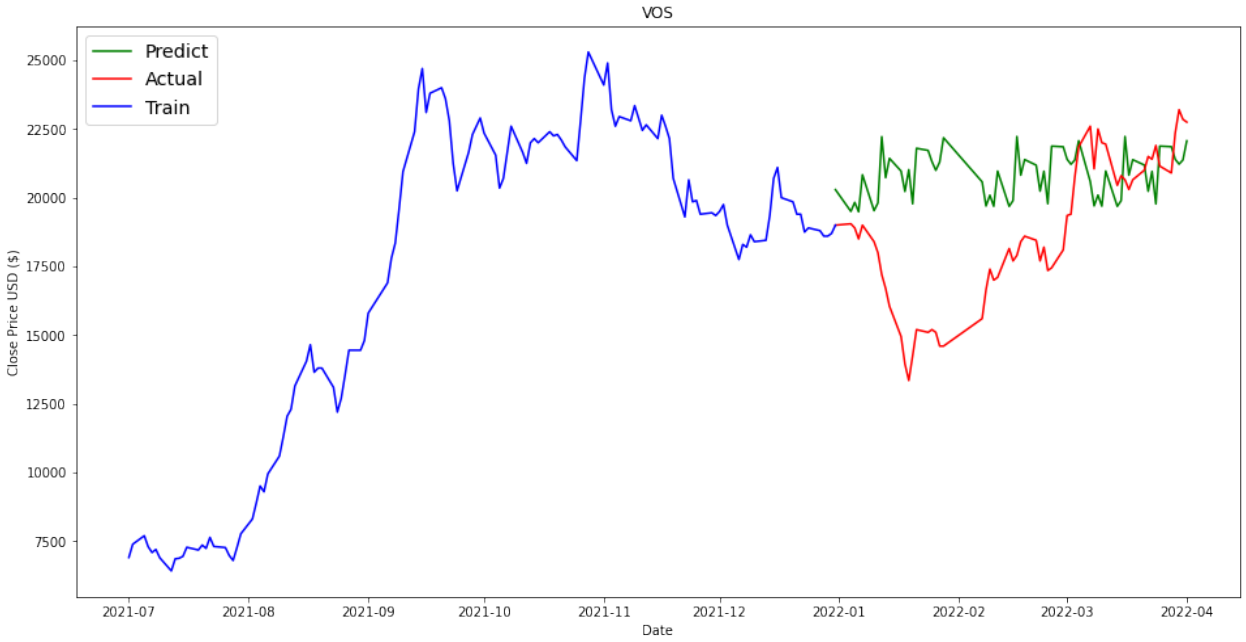

Figure 93: Predicting VOS stock using LSTM

6 Evaluating Model

This step is really necessary. It is crucial to evaluate the predictive model's performance after building a model.

6.1 The importance of choosing metrics and how to choose

Choice of metrics influences how the performance of a performance evaluation model is measured and compared. If we are not using metrics that correctly measure how accurately the model is predicting our problem, we might be fooled to think that we built a robust model. Let's take a look at an example to understand why that can be a problem and how predictive analytics can cope with it.

Consider the prediction of a rare disease that affects 1 % of people. If we use a metric that solely indicates how well the model predicts correctly, we can end up with a 98% or 99% accuracy since the model will be correct 99% of the time by anticipating that the patient does not have the disease. But it is not the model's main purpose.

Instead, we may want to employ a metric that simply assesses true positives and false negatives, assessing how accurate the model is in predicting the occurrence of the disease.

We want our model to have the same predictive evaluation across many different data sets, thus proper predictive performance model assessment is equally crucial. In the other words, the outcomes must be comparable, measurable, and reproducible.

And how we can choose reasonable metrics for our problems. All problems a performance evaluation model can solve fall into one of two categories: a classification problem or a regression problem. It is important to first determine what your goal or problems need to be solved. That will be the starting point to choose the metrics, and ultimately determine what a good model is.

A classification problem is about predicting what category something falls into. An example of a classification problem is analyzing medical data to determine if a patient is in a high-risk group for a certain disease or not.

A regression problem is about predicting a quantity. A simple example of a regression problem is a prediction of the selling price of a real estate property based on its attributes (location, square meters available, condition, etc.).

Certainly, our problem (predicting stock price) is a regression problem, so we will more concentrate on this kind of problem than the classification problem.

To evaluate how good your regression model is, you can use the following metrics:

- R-squared: It is how much better your regression line is than a simple horizontal line through the mean of the data.

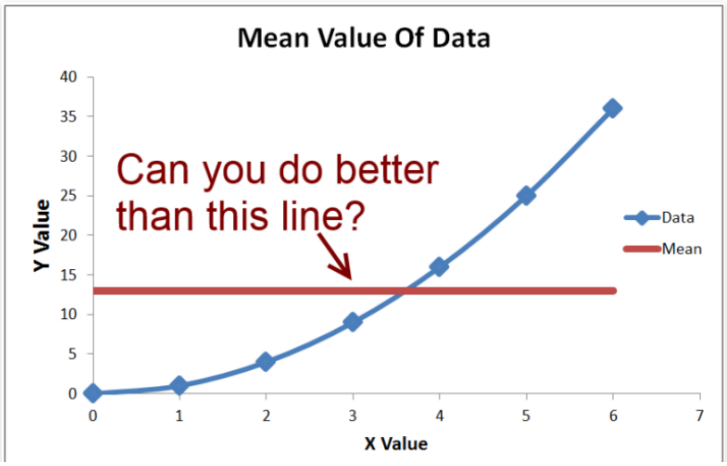

Figure 94: R-squared explanation

- Average error: the numerical difference between the predicted value and the actual value.
- Mean absolute error (MAE): It is the mean of the absolute difference between the actual value in the dataset and the value predicted by the model. The absolute values are taken, and if it's not then the negative and positive differences will cancel out each other. The smaller the MAE, the more accurate the model is. If MAE is zero it shows the model is perfect. If MAE is large then the model is not good.
- Mean Square Error (MSE): This is the mean of the squared difference between the actual value in the dataset and the value predicted by the model. It is good to use if you have a lot of outliers in the data.
- Average absolute error: similar to the average error, only you use the absolute value of the difference to balance out the outliers in the data.

In this section we will mainly use R-squared and MAE to evaluate. Let us explain more about R-squared, in most statistics books, you will see that an R squared value is always between 0 and 1, and that the best value is 1.0. That is only partially true. The lower the error in your regression analysis relative to total error, the higher the R^2 value will be. The best R^2 value is 1.0. To get that value you have to have zero error in your regression analysis. However R^2 is not truly limited to a lower bound of zero. You can get a negative R squared value.

In conclusion:

- An R^2 of 1.0 is the best. It means you have no error in your regression.
- An R^2 of 0 means your regression is no better than taking the mean value, i.e. you are not using any information from the other variables
- A negative R^2 means you are doing worse than the mean value. However maybe summed squared error isn't the metric that matters most to you. (For instance, maybe you care most about mean absolute error instead)

For evaluation, we will use sklearn.metric (from sklearn library) to execute MAE and R-squared of each model.

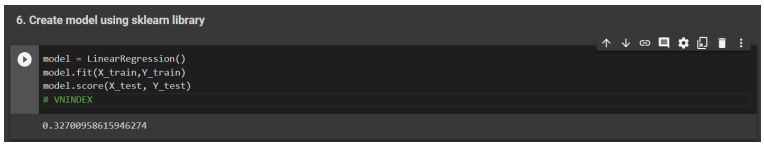

Figure 95: Using SKlearn to execute R-squared

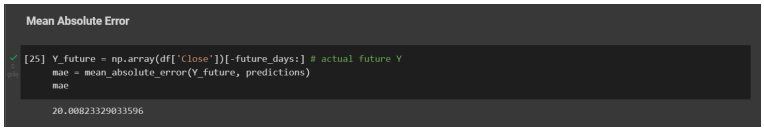

Figure 96: Using SKlearn to execute MAE

6.2 Linear Regression and Decision Tree

	Linear Regression		**Decision tree**	
	R-squared	MAE	R-squared	MAE
VNINDEX	0.33	20.01	0.07	4.45
CII	0.01	5668.65	0.27	1158.89
SKG	0.1	2749.42	0.38	849.58
VJC	0.01	8730.52	0.08	4250.28
TCD	0.44	1328.92	0.64	615.68
TGG	0.06	6175.98	0.01	1691.52
VOS	0.61	1765.08	0.32	629.24
APH	0.45	1581.52	0.4	844.92
NHH	-0.01	3367.81	-0.73	1569.31
PSH	-0.08	2433.37	-0.17	1245.37

From the table we can see the value of R-squared and MAE of linear regression and decision tree model of each stock index, decision tree predicts more precisely than linear regression. However, generally, both linear regression and decision tree did not finish their task as well as we expect. Most of R-squared values are too low and even negative and MAE values are also too high, which is unacceptable.

This huge deviation error due to some factors. Firstly, linear regression has some obstacles:

- Non-Linearity of the response-predictor relationships: It is the assumption for linearity, which states that the relation between the predictor and response is linear. If the actual relation between response and the predictor is not linear, then all the conclusions we draw become null and void. Also, the accuracy of the model may drop significantly.
- Correlation of error terms: A principal assumption of the linear model is that the error terms are uncorrelated. The "uncorrelated" terms indicated that the sign of error for one observation is independent of others. The correlation among error terms may occur due to several factors. For instance, if we are observing the weight and height of people. The correlation in error may occur due to the diet they consume, the exercise they do, environmental factors, or they are members of the same family.
- Collinearity: In linear regression, we assume that all the predictors are independent. But often the case is the opposite. The predictors are correlated with each other.

Secondly, there are also some disadvantages in using decision tree model:

- Method of overfitting: If we discuss overfitting, it is one of the most difficult methods for decision tree models.Overfitting refers to the condition when the model completely fits the training data but fails to generalize the testing unseen data. Overfit condition arises when the model memorizes the noise of the training data and fails to capture important patterns. The overfitting problem can be solved by setting constraints on the parameters model and pruning method. As you know, a decision tree generally needs overfitting of data. In the overfitting problem, there is a very high variance in output which leads to many errors in the final estimation and can show highly inaccuracy in the output. Achieve zero bias (overfitting), which leads to high variance.
- Reusability in decision trees: In a decision tree there are small variations in the data that might output in a complex different tree is generated. This is known as variance in the decision tree, which can be decreased by some methods like bagging and boosting.

Therefore, If you want to overcome the limitations of the decision tree, then you should use the random forest method, because it does not depend on a single tree. It creates a forest with multiple trees and takes the decision based on the number of majority of votes.

6.3 LSTM and ARIMA

	ARIMA	**LSTM**
	MAE	MAE
VNINDEX	35.74	34.58
CII	18247.5	12821.261
SKG	3173.33	3129.19
VJC	10361.67	10671.05
TCD	4546.40	1623.33
TGG	12743.53	11805.95
VOS	2229.78	3347.81
APH	2574.22	4236.29
NHH	6370.75	6783.53
PSH	2395.57	2762.28

The value MAE value of LSTM and ARIMA model are out of the range that can be accepted. There are many reasons explaining for the incorrect in prediction of ARIMA and LSTM model:

- In both model, when we use the previous data for predicting the future data, we ignored many other factors affecting to the stock price, especially in the covid pandemic. Moreover, the stock price in a short period don't have the seasonality so when we using data in only 5 month it's not enough data for predict in the future.
- For ARIMA model, in the code, we use auto arima and let the model choose its parameter. The model evaluate the best set of parameters for model using AIC index so sometimes when the set of parameters is (0,0,0), the output value will be a constant.
- For LSTM model: Because of the size of the training data, so we had to build a model that receives data of 40 data and then give the prediction of next 20 data. As the result, we had to code a function to feed again the previous predicted, so the value in long-term prediction will be significantly incorrect.

6.4 Objective reasons

There are also many reasons apart from the algorithm of each model, which lead to the deviation error of our prediction.

- Supply and demand: One of the main factors affecting the share market is the imbalance between supply and demand which leads to the increase or decrease in the price of stocks. In addition, factors such as economic data and interest rates affect the demand for stocks leading to fluctuations in the value of stocks.
- Interest rates: When the interest rate is low, the companies can borrow a considerable amount at a lower interest, resulting in their profits due to an increase in the stock price. On the other hand, higher interest rates lead to lesser profits and reduced stock prices.
- Political factors: There have been multiple political factors affecting stock markets. For instance, the price of stocks goes down in case of risk of war, weak government, public outrage against the government, etc. Budget announcements or elections significantly impact the volatility of the market, affecting the stock prices. In addition, the value of stocks is also reduced in case of riots or political turmoil in the country.
- Natural calamities: Natural calamities and pandemics such as floods, earthquakes, and pandemics such as Yellow Fever, Ebola and the recent COVID-19 one too, can drastically affect the value of stocks. Due to the stock prices are bound to fall due to the destruction of property, finances, and other assets. It affects not only a company's performance but also people's capability to spend.
- Inflation: Inflation directly affects the finances of people resulting in reduced capacity to invest. Moreover, increased inflation rates discourage people from investing, making companies suffer. Hence, inflation has a critical role in affecting one's investing power, purchasing power, and the country's overall economy.

In general, these are factors affecting directly stock price, we find it challenging to apply those to our models in order to have more accurate results.

6.5 Cross-model usability among indicators or groups of indicators

After evaluating and analyzing 4 models, cross-model usability among indicators or groups of indicators (for instance, we apply the model train by tourism data for finance inference) is nearly impossible. As we can see, the MAE values of 10 distinguished stock indexes are extremely different. As we can see, the MAE values of 10 distinguished stock indexes are extremely different(although some values are in the acceptable range, the rest is not). For instance, the range of values of VNINDEX is small enough to predict and the MAE value is quite good. However, the range of values of CII is too high to make a precise prediction. In conclusion, each indicator has a different range of values, variability and size of training data, so we need to apply the most suitable model for each stock index.

6.6 Which is the best model?

As we mentioned before, your model fit is reflective of your data set. In general, we use 4 different models to predict 10 datasets, decision tree seem to be the best and closest to the reality model(based on MAE value and the graph).

To be more detailed, Linear Regression and Decision Tree:

- Decision trees support nonlinearity, where LR supports only linear solutions.
- Where there are large number of features with less-datasets(with low noise), linear regressions may outperform decision trees. In general cases, decision trees will be having better average accuracy.
- Decision trees handles colinearity better than linear regression.

Neural networks and decision trees:

- Decision trees have an easy to follow a natural flow. They are also easy to program for computer systems with IF, THEN, ELSE statements. We can see that the top node in the tree is the most influential piece of data that affects the response variable in the model. Because these trees are so easy to understand, they are very useful as modeling techniques and provide visual representations of the data.
- The neural network is not so easy to understand from the visual representation. It is very difficult to create computer systems from them, and almost impossible to create an explanation from the model. Neural networks can handle binary data better than decision trees but cannot handle categorical values.

In conclusion, the best-fitted model is the one that most accurately fits your data. And in our perspective, the most suitable model for 10 datasets is decision trees.

7 Affected sectors by the Economic Crisis

7.1 Stock index sectors are most affected

Throughout a one-year period starting from July 2021 to July 2022, it is irrefutable that Vietnam in particular, and the world, in general, went through a rough time for economic growth due to being influenced by witnessing one of the harshest Coronavirus lockdowns all over the world during the second half of 2021. What is more, the Russia - Ukraine war in the first half of 2022 made the financial situation of the world go from bad to worse due to inflation and high-interest rates negatively affecting the supply chain and prices of raw materials, energy, food - food, and so on. The stock market is severely affected as a result, and definitely, these bring about more and more challenges for all businesses. Additionally, **Manufacturing** and **Wholesale Trade** are the two industries most substantially damaged by the Economic crisis.

Notice: Stock prices fall when interest rates rise, and vice versa, according to the inverse relationship between interest rates and the stock market. The amount of idle cash held by individuals and companies tends to flow into the banking system or be invested in government bonds when interest rates go up due to increased competition between lending institutions and the government. This contributes to the fact that the stock market's cash flows and stock price index both decline. As a result, it has a detrimental effect on investors' psychological well-being, which lowers investment and damages the stock market. Thus, it can be seen that interest rates have a detrimental influence on the stock index.

- **Wholesale Trade:**

  Our team has opted for petroleum, particularly the stock PSH (Nam Song Hau Trading Investing Petroleum Joint Stock Company) illustrated in the graph below, which is a Vietnam-based oil and gas refining and marketing corporation. Condensate and petroleum solvents, as well as gasoline, fuel, and other petroleum products, are refined and manufactured by the company. It engages in retail trading of gasoline and fuel through its network of gas stations as well as wholesale trading of petroleum products through its shops, terminals, and bulk stations. With fleets of seaborne tankers, ferries, and road tankers, the company also provides marine, river, and road transportation services for petroleum products. It also works on real estate projects such as the Vam Lang Resettlement Area Project, the Ecotourism and Resort Area Project in Can Tho City, a number of bulk station and terminal projects, refinery plant projects, a project to grow organic rice, and a project to store agricultural products.

  - Oil is an indispensable source of fuel and necessary input for transportation that cannot be substituted in the manufacturing process. In addition, oil is a widely traded commodity worldwide. This means that changes in rising oil prices have an impact on a variety of macroeconomic factors, including asset values and financial markets generally. These factors include inflation rates, monetary policy, national income, production costs, and business sector profits. As a result, it is anticipated that the stock market would be somewhat impacted by the change in the price of oil. Increased company expenses and sectors that depend on energy will be under pressure due to rising oil prices. It consequently raises projected future expenses, lowers cash flows, and thus reduces the value of assets.

  - It is transparently obvious that the war between Russia and Ukraine "redraws" the global oil market. The Vietnam-Russia joint venture Vietsovpetro produces over a third of Vietnam's crude oil output. Russia's state-owned oil corporation Zarubezhneft owns 49% of this joint venture while PetroVietnam owns 51%. Realizing that Russia is the world's top producer of oil and that its oil and gas reserves are tremendous. With the conflict in Ukraine, Russia has received the most international sanctions, which target the nation's economy as Western governments try to prevent

assaults. Due to a tight supply, sanctions on Russia caused prices to rise past 100 USD per barrel which leads to plunging the oil and gas industry into turmoil.

- The volatility in oil prices will increase income if the company is a net producer of oil and lower it if the company is a net consumer of oil while taking into account a certain form of security. For a country that is a net importer, a spike in oil prices will specifically cause pressure that lowers exchange rates and raises the domestic inflation rate; the predicted inflation rate rising results in a corresponding rise in the discount rate. Based on the information from Lao dong newspaper, the price of gasoline went through 16 adjustments that there are 13 increases, 3 decreases during the half of the year 2022. On the other hand, for a country that is a net exporter of oil, a rise in oil prices will provide a beneficial effect on the stock market. Hence, there are two possible outcomes for the link between oil prices and stock prices.

⇒ **Rising oil will exert a huge impact on earnings from stocks.**

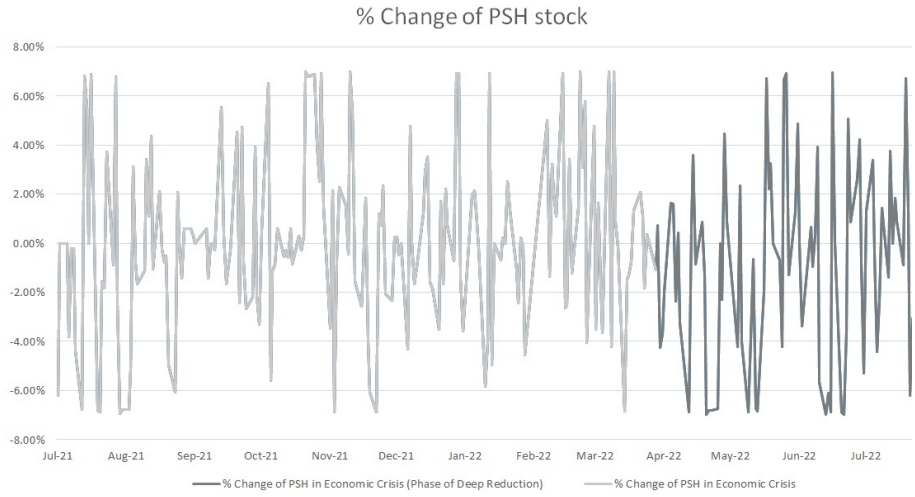

Figure 97: The graph demonstrates the change of PSH stock from July 2021 to July 2022

• Overall, it is plainly evident from the given line graph that the given course from July 2021 to July 2022 saw a sharp decline in the figure for the stock prices of PSH, which was equivalent to -63.80%.

• PSH stock was reported to have a definite sign throughout 9 months starting from July 2021, standing at 12.96%. An adverse tendency could be witnessed in the following over 3 months later, which registered a massive drop of -76.76%.

Such fluctuations in the industry of oil and gas have brought about numerous consequences, especially the extensive alterations in stock market. Due to the outbreak of COVID-19, this industry dropped quite hard, not so dramatically as opposed to the first half of the year 2022, which saw a number of violent oscillations but almost hit the low throughout the period affected mostly by the war, specifically from April to July as can be known as the phase of deep reduction in the demonstrated graph aforesaid.

- **Manufacturing:**

Having 2 stocks in the manufacturing industry that are 2 of the 3 strongest drop stocks that we have collected above. APH: An Phat Holdings Joint Stock Company (Plastics Product Manufacturing) and NHH: HaNoi Plastics Joint Stock Company (Industrial Machinery Manufacturing) are demonstrated in the graphs below. Notably, Hanoi Plastics Joint Stock Company, also known as HPC, has become a part of the High-tech Plastic Group of An Phat Holdings Group since the

end of 2018. HPC is considered a plastic manufacturing company with the most cutting-edge and contemporary in Vietnam, which is the leader in the North of our country in terms of technology for manufacturing all kinds of high-quality industrial engineering plastic products according to international standards. Moreover, An Phat Holdings is not only a leading group operating in the field of high-tech and environmentally friendly plastic production in Southeast Asia but also a leading group operating in the field of high-tech and environmentally friendly plastic production in Southeast Asia. Therefore, the two stocks share the trait of being associated with the plastic sector.

Plastic resin costs go up along with the price of oil

- Experts explained the issue of stocks decreasing significantly in the plastic industry by stating that the challenge faced by plastic firms was caused by the ongoing increase in the price of key raw materials. Because plastic resins make up between 60 - 70% of the cost structure of companies that make plastic, when their price is high, as the period 7/2021 - 7/2022, it dramatically raises the input costs for such companies, which has an impact on the output of the manufacturing process and business outcomes. That has an influence on the allure of plastic stocks as well.

- According to information from the international financial portal Investing, the cost of PE resin rose by 10.4% during the course of three months, from December 9, 2021, to March 8, 2022. Analogously, from December 6, 2021, to March 8, 2022, the price of PP resin climbed by almost 10%.

- The reason is that petroleum products, most often PVC, PP, and PE, are used to make plastic resins. Based on the current context of the world, the war between Russia and Ukraine has produced catastrophic effects on gasoline, consequently leading to the huge rise of plastic productions as mentioned.

⇒ **Low investor expectations are a result of the business's difficulty.**

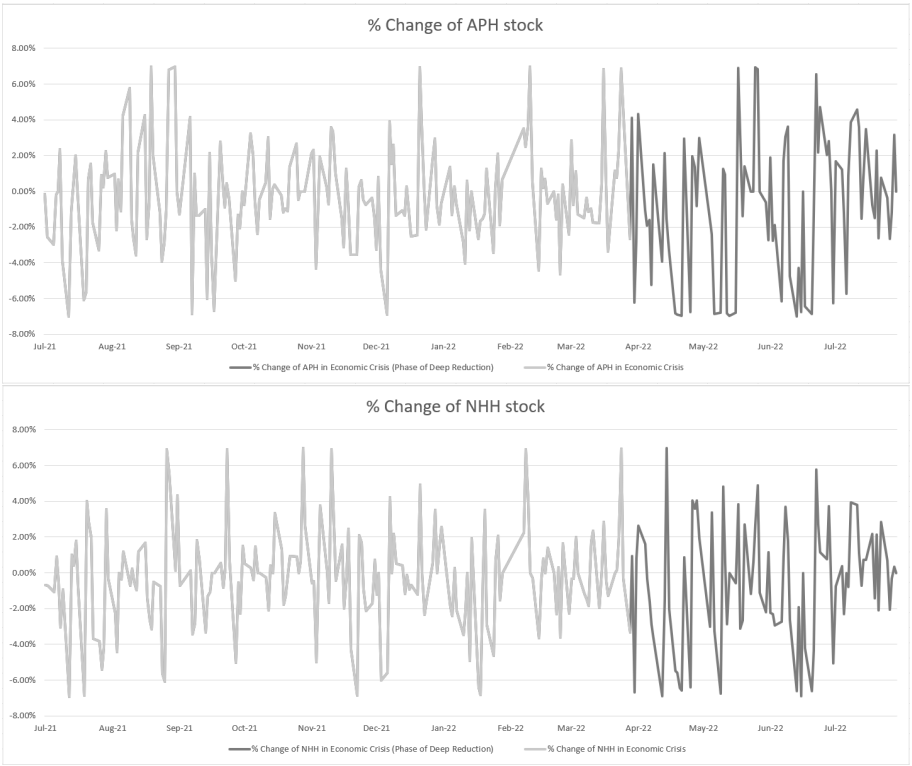

Figure 98: The graph demonstrates the change of APH and NHH stock from July 2021 to July 2022

• As can be seen from the charts, the stock prices of APH and NHH both dropped on a significant basis, by a collective -114.16% and -102.19% within a year. (7/2021 - 7/2022)

• APH stock had a precipitous loss of -76.84% in merely 3 months, upwards of double its -37.32% decline during the prior 9 months. (4/2022 - 7/2022)

• An analogous trend could be seen in the figure for NHH stock, which suffered a steep fall of -54.70% for only 3 months, which was 7% more than the -47.49% decline over the previous 9 months. (4/2022 - 7/2022)

Previously, the demand for oil and gas grew as the economy began to recover from the Covid-19 outbreak, throwing off the market's equilibrium. Causing a rise in the cost of raw materials for plastics manufacturing, which in turn led to declines in stock prices (specifically, starting in 2021, the cost of raw materials began to rise rapidly, rising 1.6 times during one year). And the culmination is the formal dispute between Russia - Ukraine that started in March 2022 and coincided with the dramatic decrease in the prices of APH and NHH stocks since the end of March 2022, as seen in the aforementioned chart.

7.2 Potential growth after Economic Crisis

When the global and domestic economic context changes (inflation increases with the risk of global recession), the global financial-monetary market tends to tighten, and interest rates and exchange rates increase. Investors adjusted their assessment and expectations, and a substantial alteration of the global stock market (including Vietnam) is inevitable, particularly from April 2022.

Considering the state of the global economy at the moment, it is universally acknowledged that Vietnam's stock market still has the prospect of recovering in the last month of 2022 and expanding more favorably in 2023 with a gain of roughly 15 - 20% by several analysts. Therefore, the 2 aforementioned industries indeed have the ability to resurge in December of 2022. Meanwhile, the prior months' growth potential is unfavorable, August through November specifically.

- **Wholesale Trade:**

  After the course of deep reduction, the price of analyzed stock market of PSH in particular, and petroleum in general was still negative due to the long-lasting military – based situations. According to the news from Thanh Nien, the price of Brent crude oil rised to 2.3% with $95.69 per barrel while the price of US WTI crude oil jumped dramatically to 3% with $87.91 per one on October 26, and its price still increased significantly till the first half of November 2022. Additionally, from the second half of November 2022 on, we have clues that the stock market will resurge and recover once again.

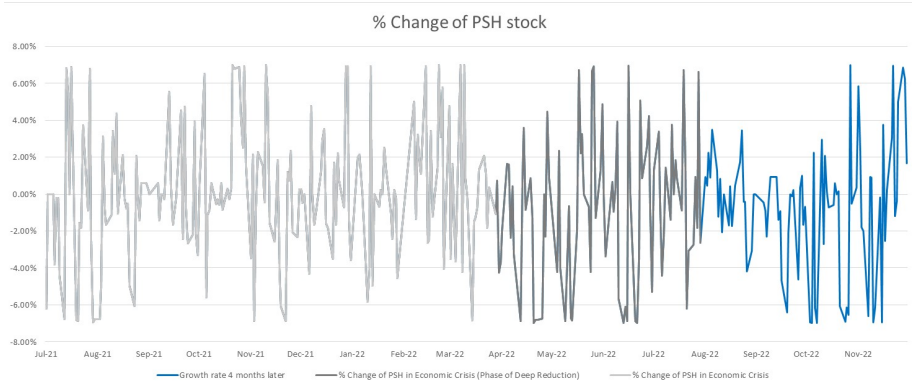

Figure 99: The graph demonstrates the overall change of PSH stock from July 2021 till now

• Compared to the previous time of massive decrease, PSH stock jumped slightly to -63.36% in the next 3 months between August 2022 and October 2022, which was still not a positive indication

for this industry. In addition, during the period of the first half of November 2022, the change of the stock was also on the decline of -20.64%, which was less than three times as opposed to the mentioned course of the huge drop.

• Such violent oscillations would not be witnessed in the figure for the stock in the remainder of November 2022, which turned out to be 29.61% and became the highest sum of change throughout the analyzed course.

⋆ According to VTC news on December 9, 2022, a representative of a leading petroleum trading company supposed that the world oil price has continuously dropped sharply recently; if the downtrend continues, the regulator will certainly reduce the retail price of domestic gasoline and oil in the next operating period. At least, Vietnamese Minister of Industry and Trade will assure the oil prices will stabilize the market before-during-after Tet Holidays. Moreover, if the war does not spread to NATO member countries around Russia, several analysts assumed that the demand for oil will decrease because of the economic downturn, the supply will increase because of high oil prices till the end of this year. This will restore the balance of Supply and Demand in the world market, and oil prices will fall again.

⇒ **The figure for PSH stock in the second half of November 2022 provides us a firm indication for a new chapter of this industry and its potential growth in the next period of time.**

- **Manufacturing**

  In general, 4 months after the drastic reduction, the Economic Crisis, the Russia-Ukraine war going on is still having an impact on the plastics industry. NHH and APH stocks both plummeted, with APH shares falling quite hard. However, the rising signals of these 2 equities at the end of November provided a promising indication.

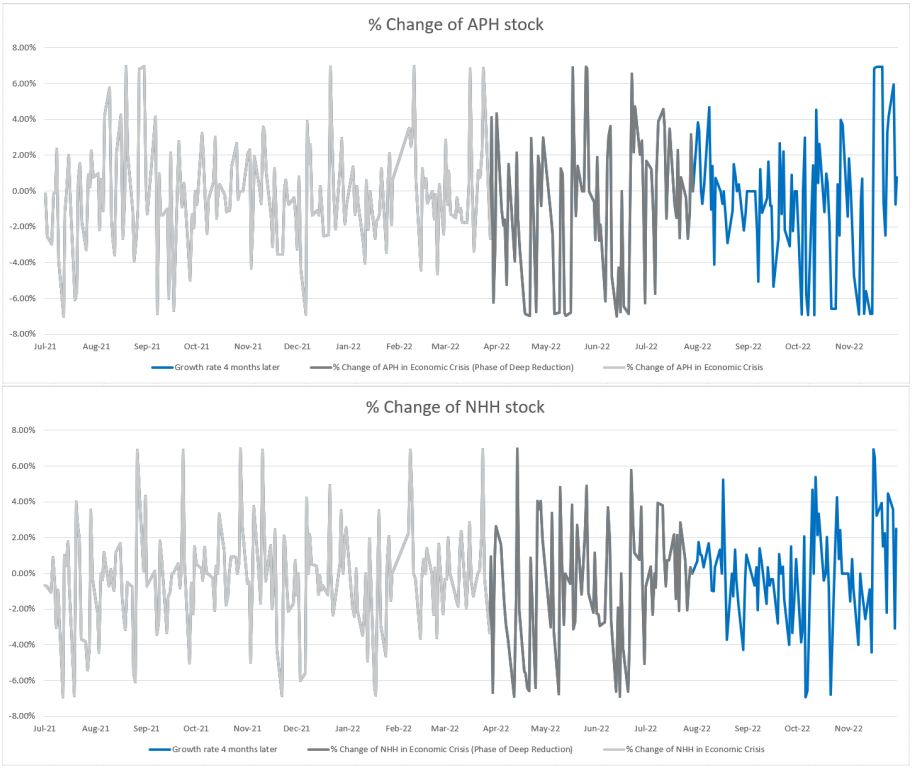

Figure 100: The graph demonstrates the overall change of APH and NHH stock from July 2021 till now

• The stock prices of APH continued to decline within 3 months after the phase of deep reduction,

falling by -42.24%. (8/2022 - 10/2022)

APH stock experienced a sharp decline of -38.79% in just the first half of November, approx its (-42.24%) decline during the prior 3 months. (1/11/2022 - 15/11/2022)

On the other part, APH stock had a quick increase, rising 37.83% in just a half-month after that. (15/11/2022 - 30/11/2022)

Similar to APH stock, NHH stock also declined, but modestly.

• After 3 months of the phase of steep decrease, the stock prices of NHH still plummeted, lowering by -9.15%. (8/2022 - 10/2022)

In the first half of November, NHH stock had a loss of -15.93%, which is almost 7% greater than the decline (-9.15%) it underwent throughout the prior 3 months. (1/11/2022 - 15/11/2022)

On the other hand, NHH stock immediately rose, gaining 29.63% within a half-month after that. (15/11/2022 - 30/11/2022)

⋆ The rise shown above illustrates the point at which the issue of raw resources starts to be resolved. Thanks to large enterprises' expansion plans, the US supply shortage has improved and the global supply has the intent of increasing. Large manufacturers in China, India, and the US invest to expand the production of plastic products right in 2022, with a capacity of millions of tons.

⇒ **The growth of APH and NHH stocks in the second half of November 2022 gives us an obvious picture of the future and upside opportunities of these two stocks in particular as well as the plastics industry in general.**

8 Conclusion

All in all, based on the current context of global economy and several adverse situations triggered by COVID-19 and military – related endeavors, our team has channeled all-out efforts in order to research the overall changes of the market recently and find out numerous approaches to carry out this assignment, *"Vietnamese Stock Market Analysis"*, and also with a view to discovering and clarifying the questions we have got initially that "Who will be the millionaire? Is this the game for us?". By analyzing scientifically and accumulating reliable sources of data in the required period to work out the task of the project, we have produced various predictive models from *Linear Regression*, *Decision Tree*, *ARIMA*, *LSTM* for the next 3 months after the second half of 2021. The questions mentioned above could be challenging for Sophomore students like us in 4 weeks to figure it out or merely release any firm statements; however, we suppose that those models which have been run and proved definitively throughout the project would be functional and practical to some extent when they are applied in the actuality.

If you have any further questions, please contact us via this email: nam.nguyenolkmphy@hcmut.edu.vn

9 Our code for analysis, model prediction and collected data

9.1 Distribution

LINK 1

9.2 Linear Regression model

LINK 2

9.3 Decision Tree model

LINK 3

9.4 ARIMA model

LINK 4

9.5 LSTM model

LINK 5

9.6 Collected data

LINK 6

References

[1] **Introduction**
https://www.investopedia.com/terms/s/stockmarket.asp
https://www.nerdwallet.com/article/investing/stock-market-basics-everything-beginner-investors

[2] **Tools and environment**
https://www.w3schools.com/python/python_intro.asp
https://www.w3schools.com/python/numpy/numpy_intro.asp
https://scikit-learn.org/stable/
https://pandas.pydata.org/
https://matplotlib.org/
https://scikit-learn.org/stable/
https://www.mathworks.com/matlabcentral/fileexchange/36000-fbd-find-the-best-distribution-tool
fbclid=IwAR1Aa3NSII_qNKpwGYCUAANWcr3nJ2l9DWO-ENrSIWIB07MmM0RD2t5IH2E
https://www.geeksforgeeks.org/introduction-machine-learning/
https://dev.to/petercour/machine-learning-classification-vs-regression-1gn
https://www.sciencedirect.com/topics/computer-science/deep-learning-model
https://colab.research.google.com/

[3] **Distribution**
https://otexts.com/fpp2/arima.html
https://byjus.com/normal-distribution-formula/
https://en.wikipedia.org/wiki/Gamma_distribution
https://byjus.com/maths/exponential-distribution/
https://en.wikipedia.org/wiki/Normal_distribution
https://en.wikipedia.org/wiki/Exponential_distribution
https://en.wikipedia.org/wiki/Probability_density_function
https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
https://www.statistics.com/glossary/continuous-vs-discrete-distributions/
https://www.mathworks.com/matlabcentral/fileexchange/36000-fbd-find-the-best-distribution-tool
https://www.programsbuzz.com/article/differentiate-between-discrete-and-continuous-probability

[4] **Linear Regression**
https://www.youtube.com/watch?v=hOLSGMEEwlI&fbclid=IwAR2ZezARbEWvk_
TVEK96YbfCylXHMVb9mYvA-2uI8YEcs2I-OfHrCefvJb8
https://www.mygreatlearning.com/blog/what-is-regression/#what-is-regression
https://www.topcoder.com/thrive/articles/introduction-to-linear-regression#:~:
http://datapandas.com/index.php/2016/09/17/machine-learning-hypothesis-function-cost-function-

[5] **Decision Tree**
https://www.youtube.com/watch?v=hOLSGMEEwlI&fbclid=IwAR2ZezARbEWvk_
TVEK96YbfCylXHMVb9mYvA-2uI8YEcs2I-OfHrCefvJb8
https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
https://corporatefinanceinstitute.com/resources/data-science/decision-tree/

[6] **ARIMA**
https://otexts.com/fpp2/non-seasonal-arima.html
https://www.analyticsvidhya.com/blog/2021/07/stock-market-forecasting-using-time-series-analys

[7] **LSTM**
https://otexts.com/fpp2/nnetar.html
https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e
https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9

[8] **Evaluation**
http://www.312analytics.com/decision-trees-vs-neural-networks/?fbclid=
IwAR2MVrB4SJeT0qQ5i7P7598FfjaZQQ8_x-Z7ZRFA4H6OlBCUxmuvzvHI8JA
https://www.linkedin.com/pulse/pros-cons-decision-trees-aashima-yuthika
https://indatalabs.com/blog/predictive-models-performance-evaluation-important
https://www.topcoder.com/thrive/articles/metrics-to-evaluate-a-regression-model
https://towardsdatascience.com/five-obstacles-faced-in-linear-regression-80fb5c599fbc

[9] **Affected sectors and Potential Growth**
https://nhadautu.vn/vi-sao-chung-khoan-lao-doc-d71159.html
https://vnbusiness.vn/co-phieu/ky-vong-co-phieu-nganh-nhua-but-pha-1085291.html
https://diendandoanhnghiep.vn/chien-su-nga-ukraine-tac-dong-tieu-cuc-den-nganh-dau-khi-viet-n
html
https://www.tinnhanhchungkhoan.vn/gia-nguyen-lieu-giam-doanh-nghiep-nhua-xay-dung-de-tho-hon-
html