FORECASTING OF BANKING SECTOR SECURITY PRICES IN KENYA USING MACHINE LEARNING TECHNIQUES

Marwa Hassan Chacha

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Statistics of Murang'a University of Technology

October, 2022

DECLARATION

I hereby declare that this thesis is my original work and to the best of my knowledge has not been presented for a degree award in this or any other university.

_____ _____

Marwa Hassan Chacha Date

AS400/5083/2019

APPROVAL

The undersigned certify that they have read and hereby recommend for acceptance of Murang'a University of Technology a thesis entitled **"Forecasting of Banking Sector Security Prices in Kenya Using Machine Learning Techniques".**

_____ _____

Dr. Ayubu Anapapa, PhD Date

Department of Mathematics and Actuarial Science

Murang'a University of Technology

_____ _____

Dr. John Mutuguta, PhD Date

Department of Mathematics and Actuarial Science

Murang'a University of Technology

DEDICATION

To my beloved family.

ACKNOWLEDGEMENT

I want to express my gratitude to the Most High God for making it possible for me to finish my thesis by providing me with the abilities, grace, and tenacity I needed. Throughout the entirety of the research, Dr. Ayubu Anapapa and Dr. John Mutuguta provided me with wonderful assistance and direction from a variety of perspectives. I am profoundly grateful to both of them. In addition, I would like to extend my most heartfelt gratitude to my parents for the unwavering support they have shown me throughout my time spent pursuing my education. I would also like to express my gratitude to my lovely fiancee Cynthia Akaliche as well as to everyone else who assisted me financially or emotionally while I was pursuing my degree.

ABSTRACT

Financial analysts should be able to make accurate predictions about the direction of the stock market. First and foremost, investors should understand how the stock market works before making any form of investment in a company. An investment in a stock that is trading at a cheap price at the right moment can result in profits, but an investment in a stock that is trading at a high price at the wrong time might result in low results. When deciding whether to buy, sell, or do nothing at all, experienced traders typically use a combination of indicators to make their assessment. Less experienced traders, on the other hand, may not be able to recognize the appropriate patterns of market elements when using the available indicator baskets. Throughout the course of history, there have been a wide variety of strategies and methods that may effectively predict the behaviour of stocks. On the other hand, more resources have been devoted to the study of the stock market since Machine Learning (ML) was developed, and it has been shown that accurate stock market prediction is feasible. Although research in this area has been done, there hasn't been any that compares the various machine learning algorithms for predicting the securities of the Kenyan banking sector. As a result, the purpose of this study was to make projections on the pricing of banking sector securities in Kenya by utilizing Machine Learning techniques and fit Autoregressive Integrated Moving Average (ARIMA), Artificial Neural Network (ANN), and Support Vector Machine (SVM) models for forecasting banking sector security prices. The study made its predictions based on the random walk theory of predicting stock market movements. The research aimed to look at all of the banks that are traded on the Nairobi Securities Exchange, and a representative sample was picked from three of those institutions, using purposive sampling technique: the Kenya Commercial bank, the Equity bank, and the Co-operative bank. The correctness of each model was evaluated by applying the Root Mean Squared Error (RMSE) criterion, the Mean Squared Error (MSE) criterion, the Mean Absolute Percentage Error (MAPE) criterion, and the Mean Absolute Error (MAE) criterion. The confusion matrix criterion was used to choose the most effective model. With an RMSE of 0.1217, SVM beat the other models when using the accuracy metrics criteria and the Confusion Matrix. In comparison, ANN and ARIMA had RMSEs of 0.1477 and 0.1743 respectively. It was also clear, based on the confusion matrix, that SVM performed better than ANN since it had an accuracy of 0.6171, which is equivalent to 61.71 percent, whereas ANN only had 0.5959, (59.59%). The study recommended that SVM should be used by financial experts for stock price predictions. For further research, historical data can be used in conjunction with studies of financial and political events when forecasting.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

ACF	Autocorrelation Function
AF	Activation Function
AIC	Akaike Information Criterion
ANNs	Artificial Neural Networks
AR	Auto-regressive
ARIMA	Auto-regressive Integrated Moving Average
BP	Back Propagation
CO-OP	Cooperative
IIR	Infinite Impulse Response
KCB	Kenya Commercial Bank
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multi-layer Perception
MSE	Mean Squared Error
NSE	Nairobi Securities Exchange
NLP	Natural Language Processing
PCA	Principal Component Analysis
RBF	Radial Basis Function
RMSE	Root Mean Squared Error
SVM	Support Vector Machine

DEFINITION OF TERMS

Stock Market – A stock market is a type of public market in which investors and traders may buy and sell shares of companies as well as derivatives of those shares using either electronic or physical means of processing and exchange (Pahwa and Agarwal, 2019).

A **time series** is defined as "a collection of observations made about a process at various intervals of time throughout the course of time" (Fuller, 2009).

Machine Learning - A process known as machine learning (ML) is one in which a computer program is claimed to study from information pertaining to some group of activities and valuation metric. As a consequence of this practice, robots will eventually mimic human behavior (Nti et al., 2019).

The **Autoregressive Integrated Moving Average (ARIMA)** model is applied to the problem of forecasting non-stationary time series in which it is assumed that the variables are linear. The forecasting of non-stationary time series is the purpose of this model (Box and Jenkins, 2013).

A **Support Vector Machine,** more commonly referred to as an SVM, is a form of supervised learning model that analyzes data in order to recognize patterns by making use of various learning strategies. SVM models can be used for a variety of tasks, including classification and regression analysis, respectively (Horak et al., 2020).

The term "**Artificial Neural Network" (ANN)** refers to any mathematical or computer model that makes an effort to replicate the structure and actions of a biological neuron.

Estimating or approximating the functionality of real-world systems is one of xiii its use (Kukreja et al., 2016).

CHAPTER ONE

INTRODUCTION

1.1 **Background Information**

The ability to accurately anticipate time series is essential in many disciplines, including but not limited to economics, finance, business intelligence, meteorology, and telecommunications (Bontempi et al., 2012). Since the 1950s, one of the most popular research topics has been the use of time series data to make predictions about future events. Since then, a number of other methods of prediction have come into existence. Techniques such as technical and fundamental analysis, traditional methods for analyzing time series, and machine learning are among them (Weigend, 2018).

The purpose of technical analysis is to forecast future prices by analyzing historical price data, including price, volume, and price movement (Edwards et al., 2018). In the process of forecasting, technical indicators are applied to recognize trends and patterns, which are subsequently put to use in order to estimate the price direction for the future. In order to forecast an overall direction for the sake of making future investment decisions, technical analysis also considers the way that prices in the past have moved. When determining a company's potential for the future, this method takes into account not only the business activities of the organization but also its existing financial situation (Lin, 2018).

The study of the relationship between a company's many different financial measures and other parts of the business, such as the rate of rise in inventories and sales, is the primary objective of fundamental analysis (Wanjawa and Muchemi, 2014). The evaluation of essential statistical indicators inside firms, such as the balance sheet, profitability, and

future plans, is what is meant by the term "fundamental analysis." The purpose of fundamental analysis is to gain an understanding of a company's performance by looking at the company's financial performance and to come to conclusions about the potential for the company's future (Bartram and Grinblatt, 2018).

When making forecasts about the performance of a stock's future based on the performance of a stock in the past, traditional time series methodologies typically make use of a time series scale. The fact that time series data is dependent on time is one of the distinguishing characteristics of this type of data. This means that in order to make sense of the data, a current observation needs to be reliant on an observation that occurred at an earlier point in time. However, by the late 1970s and early 1980s, linear models were not as well adapted to applications as their nonlinear counterparts, which led to the development of nonlinear models (Khedmatia et al., 2020).

In addition, throughout the course of the previous two decades, in the realm of market forecasting, machine learning (ML) models have made a name for themselves as rivals to classical statistical models. By using historical data, ML models — which are non-parametric and non-linear — learn the stochastic dependence between past and future variables (Bontempi et al., 2012). These models are sometimes referred to as datadriven models and black-box models (Mitchell and Mitchell, 1997). They can also make better predictions by combining data from many sources, and many of them do not require data pre-processing. ML approaches have been used because of their greater performance and ease of implementation as compared to traditional statistical methods.

Stock market prediction is a method for estimating the value of the shares of a company or of another financial instrument that is traded on an exchange (Sharma et al., 2020). There has been a significant amount of work done in this area using a variety of algorithms. In order to make an estimate of the values of securities in the banking sector in Kenya, this study examined forecasting systems that were based on the ARIMA and ML methodologies that are typically used. The machine learning methodologies of Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) were utilized in this study.

1.1.1 Autoregressive Integrated Moving Average (ARIMA)

The Autoregressive Integrated Moving Average (ARIMA) model is applied in situations where linearity between variables is required in order to make predictions about non-stationary time series (Box and Jenkins, 2013). The approach is adaptable to non-stationary and continuous data, much like stock market prices. ARIMA consists of Autoregressive (AR) functions that are regressed on the process's previous values, Moving Average (MA) functions that are regressed exclusively on the random process with mean zero (0) and variance (σ^2), and an integrated (I) component that differentiates and stabilizes the data series.

The theory behind an autoregressive model, which is also referred to as an Infinite Impulse Response (IIR) Filter, Box and Jenkins (2013), is that a present value can be projected using prior p values. Here, p refers to the number of steps back in time that are necessary in order to predict the present value. The autoregressive model of order p, sometimes known as $AR(p)$, is represented mathematically as follows:

$$Y_t = \mu + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t \quad (1.1)$$

where $\alpha_1, \alpha_2, \cdots \alpha_p$ are constants, μ is the model's average, ε_t is a white noise that has an average of 0 and a variance of σ^2.

The current value of, Y_t is expressed by AR (p) as a linear mixture of previous values plus some random variation generated externally. As a result, a first-order AR model, can be written as:

$$Y_t = \mu + \alpha_1 X_{t-1} + \varepsilon_t \quad (1.2)$$

A Moving Average model, MA is very useful in identifying the trend of the time series. A MA of order q, denoted MA (q) is of the form:

$$Y_t = \mu + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \cdots + \beta_p \varepsilon_{t-p} + \varepsilon_t \quad (1.3)$$

where $\beta_1, \beta_2, \cdots \beta_p$ are constants, μ is the model's average and ε_t is a white noise that has an average of 0 and a variance of σ^2 and calculates each term of the time series from the MA of the last terms of the error sequence.

The model expresses Y_t as a linear combination of white noise. Therefore, the MA is also a linear function and a stationary process too. A MA (1) can be written as:

$$Y_t = \mu + \beta_1 \varepsilon_{t-1} + \varepsilon_t \quad (1.4)$$

The ARIMA's integrated portion is obtained by reverse differencing a stationary ARMA (p,q) process, which combines AR and MA. Therefore, integration is the process of inverse differencing. An ARIMA (p, d, q) is defined as:

$$\Delta^d Y_t = \alpha_1 \Delta^d X_{t-1} + \alpha_2 \Delta^d X_{t-2} + \cdots + \alpha_p \Delta^d X_{t-p} + \sum_{j=1}^{q} \beta_j \varepsilon_j \qquad (1.5)$$

Where p denotes the order of the AR model, d denotes the number of times the data is differenced to ensure stationarity, and q denotes the order of the MA model.

1.1.2 Artificial Neural Networks (ANNs)

An Artificial Neural Network, also known as an ANN, is a collection of interconnected nodes that are referred to as neurons (weights). The functioning of an ANN is modeled after that of an animal neuron (Neal, 1996). Neural networks are an excellent tool for predicting stock prices because they are able to identify nonlinear relationships hidden inside data.

There are three layers in an ANN: input, hidden layers, and output. The input layer, which is made up of multiple formats, receives the data for training the neural network, while the hidden layers are made up of convolutional and pooling layers. The convolutional layers identify regional patterns and characteristics in the data collected by the layers that came before them, while the pooling layers logically combine features that are conceptually similar into a single feature (Zhang, 2018).

In the output layer, which is typically controlled by Activation Functions (AFs), the network's outputs are shown. The function of an AF will decide where it sits in the organizational hierarchy of a network (Nwankpa et al., 2018). When the AF is used after

the hidden layers, the linear mappings end up being transformed into non-linear forms for propagation as a consequence. The output layer is where the predictions are made.

Figure 1.1 illustrates the neural network's structural layout.

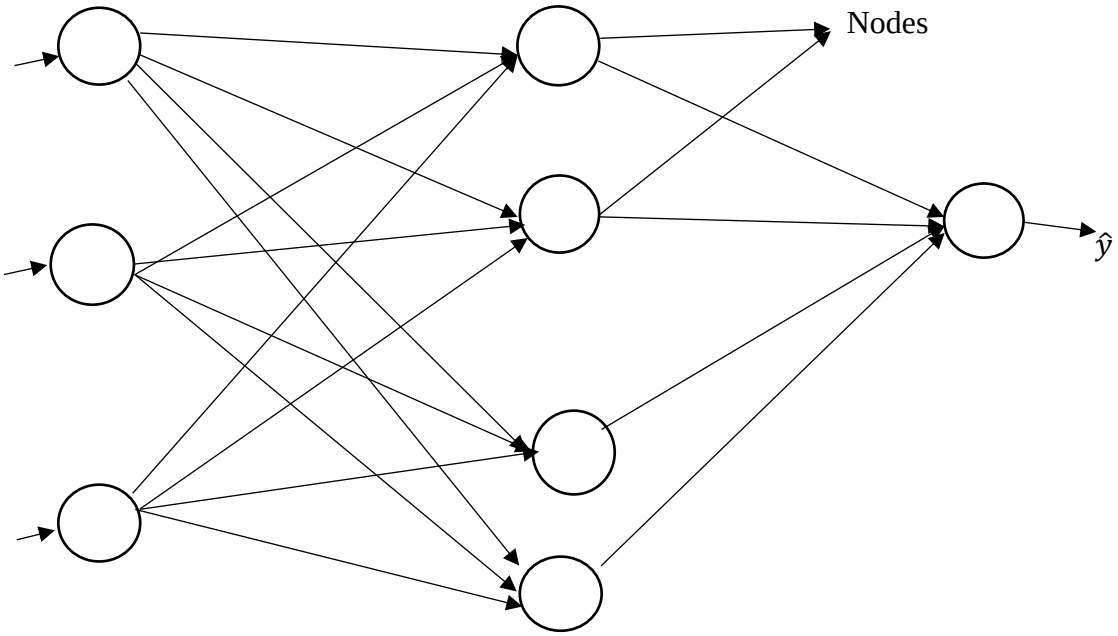


Figure 1.1: An Architecture of Artificial Neural Networks

Mathematically, an ANN can be represented as:

$$\hat{y} = g\left(w_0 + \sum_{i=1}^{m} x_i w_i\right) \tag{1.6}$$

where $\hat{y}$ is the result, g is an activation function, w_0 is a bias, and $\sum_{i=1}^{m} x_i w_i$ is a linear combination of inputs or weights.

Activation Functions, also known as Transfer Functions, Nwankpa et al. (2018), are responsible for determining whether or not a neuron should fire by computing the weighted total of input and biases. They modify the input using a technique called gradient processing, and then construct a neural network output that contains the parameters of the data.

The Sigmoid Function, the Hyperbolic Tangent, and the Rectified Linear Unit (ReLU) are the three most typical types of AFs (Nwankpa et al., 2018). The Sigmoid Function is a bounded differentiable real function with positive derivatives that is used to anticipate probability-based output for real input values. This is possible due to the fact that it reduces the inputs to a range that is contained inside 0 and 1 (Pratiwi et al., 2020). The Sigmoid function is written as follows:

$$S(x) = \left(\frac{1}{1 + e^{-x}}\right) \tag{1.7}$$

The smoother function that has a centre of zero and a range that goes from -1 to 1 is called the hyperbolic tangent (tanh) function (Sharma et al., 2017). The function is written as:

$$tanh(x) = \left(\frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}\right) \tag{1.8}$$

In multi-layer neural networks, the tanh function is chosen over the sigmoid function due to its superior performance during the training process. The tanh function is beneficial because it provides an output that is zero-centred, which aids in the process of back-propagation.

A threshold operation is carried out on each of the input elements by the ReLU AF, which results in the setting of all values that are less than zero to zero (Agarap, 2018).

The ReLU is determined as;

$$f(x) = \max\ (0, x) = \begin{cases} x, & if\ x\ \geq 0 \\ 0, & if x\ <\ 0 \end{cases} \tag{1.9}$$

In order to get rid of the gradient problem that might occur in sigmoid and tanh AFs, ReLU activation function resets the values of any inputs that have a value that is lower than zero to zero.

The Back Propagation (BP) algorithm is one of the most well-liked techniques for training neural networks. The outputs of the hidden layer are shifted to the output layer in the BP technique. There, the output for a particular input is calculated and compared to the desired result (Cilimkovic, 2015). Through the hidden layer, the fault is sent from the output layer back to the input layer. As a direct consequence of this, the weights that are shared between the neurons shift giving a neural network a set of known input data and then asking it to produce an output that is previously known is the network training process (Nandakumar et al., 2019). The process of switching from input to output and back is repeated numerous times until the error, which is defined as the discrepancy between the planned output and the actual output, is within a predefined tolerance. The weights that are derived from a trained network are what are utilized to determine how the network will respond to data that is unknown.

1.1.3 Support Vector Machine (SVM)

Support Vector Machine, or SVM, is an approach to self-supervised learning that is utilized in data mining (Das and Padhy, 2018). Vapnik et al. (1997) laid the ground work for support vector machines (SVMs), which are widely used due to their many useful

features and excellent empirical performance. SVMs were first developed to solve classification challenges, but in recent years, their applications have been broadened to include regression challenges as well.

Take a look at a linear SVM classification method where each input data point is represented in a n-dimensional space, where n is the total number of input dimensions. The hyper-plane that divides the two classes is then used to classify the data. A hyperplane that results from classification is depicted in Figure 1.2.

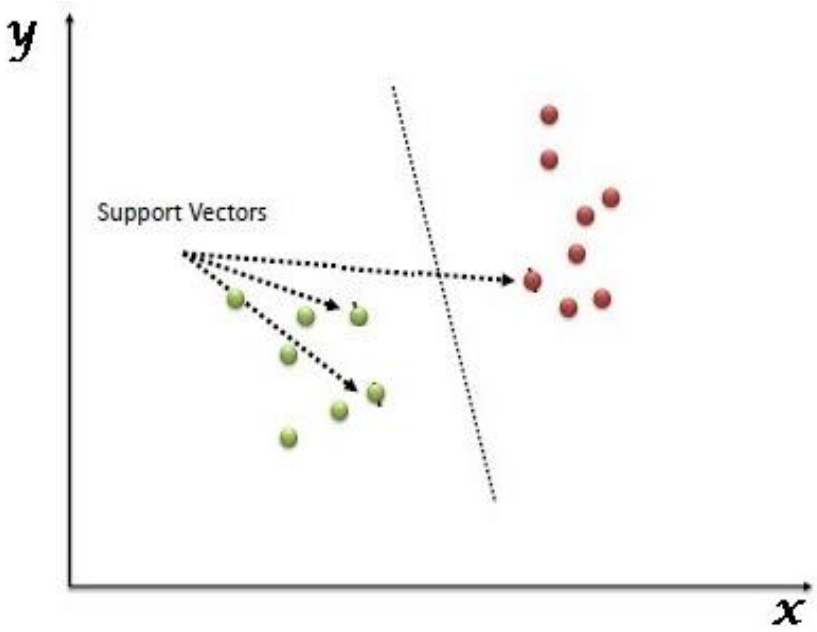


Figure 1.2: Support Vectors Hyper-Plane (Rajput, 2019)

For regression, consider a given training dataset, $D = \{(x_i, y_i) | x_i \in \mathbb{R}^d, y_i \in \mathbb{R}, i = 1, 2, \cdots, n\}$ where x_i is a multidimensional input with all independent variables, n is the number of training samples, and y_i is the scalar output. Support Vector Regression is defined as follows (Awad and Khanna, 2015);

$$y_i = f(x, \omega) = \sum_{i=1}^{n} \omega . \Phi_i(x) + b \qquad (1.10)$$

where ω is the associated weight vector to $\Phi_i(x)$, $\Phi_i(x)$ is a mapping function for non-linear transformations and b is a level that never changes.

Both ω and b must be estimated. A linear SVR may be represented as:

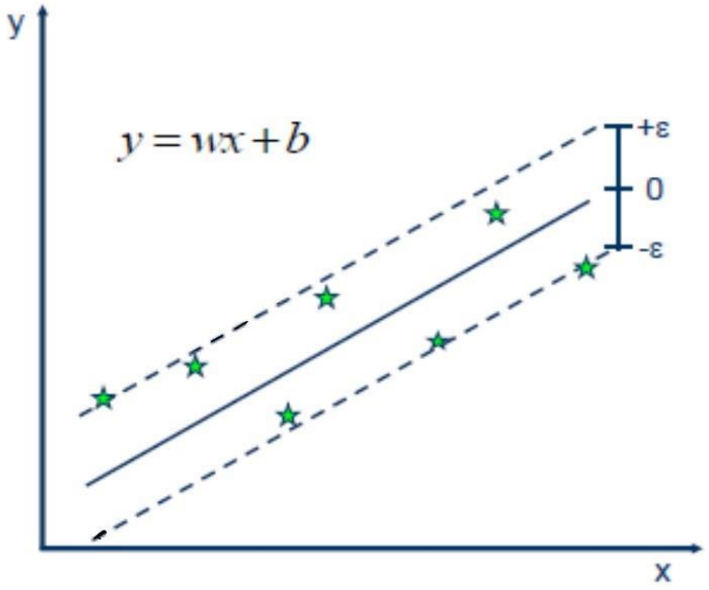


Figure 1.3: Setup of a linear SVR. (Lagat et al., 2018)

The following expression represents the formula for a linear line in a linear hyperplane:

$$H: \bar{\omega} \cdot \bar{x} + b = 0 \qquad (1.111)$$

where $\bar{\omega} = (\omega_1, \omega_2, \cdots, \omega_d) \in \mathbb{R}^d$ and $\bar{x} = (x_1, x_2, \cdots, x_d) \in \mathbb{R}^d$ then $\bar{\omega} \cdot \bar{x} = \sum_{i=1}^{d} \omega_i \cdot x_i$.

Assume that $y_i \in \{+1, -1\}, i = 1,2, \cdots, n$ where y_i has been designated as the output class. After that, the outputs will be defined as:

$$y_i = \begin{cases} 1, & if\ \omega \cdot x_i + b \geq 0 \\ -1 & if\ \omega \cdot x_i + b < 0 \end{cases} \quad (1.122)$$

where ω is a vector, $x_i \in \mathbb{R}$ and $b \in \mathbb{R}$.

The margins between classes are at their greatest on the hyperplane H. This indicates that SVMs will need to construct two additional hyperplanes as: H_1: $\bar{\omega} \cdot \bar{x} + b = 1$; and H_2: $\bar{\omega} \cdot \bar{x} + b = -1$where there is no existing x_i in H_1 and H_2, and the margin between H_1 and H_2 is maximum.

The distance of x in H_1 and H_2 to H can be calculated as:

$$\frac{|\bar{\omega} \cdot \bar{x} + b \pm 1|}{\| \omega \|} = \frac{2}{\| \omega \|} \quad (1.133)$$

Therefore, the margin can be calculated as: $\frac{2}{\|\omega\|}$ where $\| \omega \| = \sqrt{\omega^2} = \sqrt{\omega_1^2 + \omega_2^2 + \cdots + \omega_d^2}$.

Alternatively, the maximum margin will be referred to minimum of $\frac{\omega^2}{2}$ subject to:

$$y_i = \omega \cdot x_i + b \geq 1, i = 1,2, \cdots, n \quad (1.144)$$

In other cases, wrongly classified occurrences in the data may prevent SVM from finding any separating hyperplane at all. In this particular scenario, SVM makes use of a flexible margin, which allows for some of the training cases to be misclassified. By utilizing Lagrange multipliers $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n) \geq 0$, then equation (1.14) will be seen as:

$$\mathcal{L}(\omega, b, \alpha) = \frac{1}{2}\omega^2 - \sum_{i=1}^{n} \alpha_i \cdot y_i(\omega \cdot x_i + b) + \sum_{i=1}^{n} \alpha_i \tag{1.155}$$

Or the langrage dual problem as:

$$\max_{\alpha} \mathcal{L}(\omega, b, \alpha) \tag{1.166}$$

with constraints of:

$$\frac{\partial \mathcal{L}}{\partial \omega} = \omega - \sum_{i=1}^{n} \alpha_i \cdot y_i \cdot x_i = 0\,, \alpha \geq 0 \tag{1.177}$$

$$\frac{\partial \mathcal{L}}{\partial \omega} = -\sum_{i=1}^{n} \alpha_i \cdot y_i = 0\,, \alpha \geq 0 \tag{1.188}$$

From equations (1.17) and (1.18); $\omega = \sum_{i=1}^{n} \alpha_i \cdot y_i \cdot x_i$ and $\sum_{i=1}^{n} \alpha_i \cdot y_i = 0$.

Therefore $\mathcal{L}(\omega, b, \alpha)$ will be:

$$\mathcal{L}_D = \mathcal{L}(\omega, b, \alpha) \equiv \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum\nolimits_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot x_i \cdot x_j \tag{1.199}$$

Each $\bar{x} \in \mathbb{R}^d$ will be classified by:

$$
\begin{aligned}
f(\bar{x}) &= (\bar{\omega} \cdot \bar{x} + b) \\
&= sign\left(\left(\sum_{i=1}^{n} \alpha_i \cdot y_i \cdot x_i\right) \cdot \bar{x} + b\right) \\
&= sign\left(\sum_{i=1}^{n} \alpha_i \cdot y_i \cdot (x_i \cdot \bar{x}) + b\right)
\end{aligned}
\tag{1.20}
$$

where $sign\,(x) = \begin{cases} 1, & if\ x \geq 0 \\ -1 & if\ x < 0 \end{cases}$

Transferring the data to a higher dimensional space (the transformed feature space), where a linear hyperplane may be used, will solve the inseparability problem for a non-linear hyperplane in a non-linear training dataset. This will allow data to be separated into its component parts.

$$
\begin{aligned}
&\Phi: \mathbb{R}^d \longrightarrow \mathbb{R}^{d'} \\
&\bar{x} \longrightarrow \Phi(\bar{x})
\end{aligned}
\tag{1.21}
$$

The equation (1.19) will be:

$$
\begin{aligned}
\mathcal{L}_D &= \mathcal{L}(\omega, b, \alpha) \\
&\equiv \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\nolimits_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \Phi(x_i) \cdot \Phi\left(x_j\right)
\end{aligned}
\tag{1.22}
$$

where $\Phi(x_i) \cdot \Phi(x_j) = K(x_i, x_j)$ is the Kernel function. Whenever a hyperplane is produced, it maps new points into the feature space for classification, and it does this automatically.

1.1.4 Banking Sector Stock Market in Kenya

The history of banking may be traced back to the early days of society, when expanding empires needed a method to make payments for products and services purchased from beyond their borders (Colvin, 2016). The use of coins made of a variety of metals and sizes eventually superseded paper currency, which brought about the requirement for secure storage. According to the World History Encyclopedia, wealthy people in ancient Rome would store their coins in the basements of temples for the sake of protection since priests and other temple personnel created a sense of safety (Michail, 2021).

According to Beck et al. (2010), developments in Kenya's banking sector have led in the expansion of financial products, activities, and organizational structures. As a result of these modifications, the competency of the financial framework has significantly enhanced. The progression of innovation has been helpful in assisting with the transitions. The growth of the international banking industry has been helped along by many innovations and changes. Research into stock forecasting has grown increasingly relevant as a result of the expanding role that the financial services industry plays in today's economies.

In 2020, NSE-listed securities performed similarly to global capital market assets, which were marked by price volatility, irregular demand, and forecast issues. Corona virus uncertainty caused investors to reallocate funds to more protective assets like bonds. NSE

reported an increase in profit after tax of Kshs. 87.7 million or over 100% for the year ended December 31, 2020 from Kshs. 80.2 million for the similar period in 2019. Revenue decreased marginally by 5% from Kshs. 577.1 million in 2019 to Kshs. 548.2 million in 2020. This was primarily attributable to a 3% decrease in equity turnover from Kshs. 154 billion in the fiscal year ending December 31, 2019 to Kshs. 149 billion in the fiscal year ending December 31, 2020. Consequently, equities trading charges were reduced by 3% for the fiscal year that ended on December 31, 2019, from Kshs. 369.2 million to Kshs. 356.8 million for the same period in 2020. The decline in the equity turnover was as a result of reallocation of capital towards the fixed income assets as well as reduced inflow from international investors (NSE, 2020).

In the Kenyan stock market, there are a number of different participants. Capital markets, the Central Depository and Settlement Corporation, the Nairobi Securities Exchange (NSE), and dealers (stockbrokers) are all considered to be actors in this industry (Wanjawa and Muchemi, 2014).

1.2 Statement of the Problem

The market's behavior is frequently known to certain investors. An investor must be informed of the stock trend before investing in any stock. The market's trend is defined as the day-to-day swings in the stock market. The market works in such a way that as new market information becomes available, it is adjusted, rendering earlier predictive criteria useless.

Investing in a low-cost stock at the appropriate moment can provide profits, while investing in a high-cost stock at the wrong time can yield disastrous outcomes. Less

experienced traders may not be able to discern the proper patterns of market forces utilizing the baskets of indicators accessible, whereas experienced traders tend to employ a combination of indications to arrive at a buy, sell, or do-nothing conclusion.

Many studies on stock market prediction have resulted in the creation of a variety of predictive approaches, including machine learning-based algorithms, which are appropriate in today's world where a significant volume of stock market data is available. Machine learning algorithms have performed admirably in applications such as weather forecasting, fraud detection, currency exchange, disease forecasting, and pollution forecasting.

Studies have been conducted in the area of time series application. With the rising of technology and rising of more appropriate models, more research ought to be undertaken. Although machine learning models like ANN have been used to predict stock prices, there has not yet been established an appropriate model to conduct such. This study helps realize that as it applies ANN and SVM in conducting the forecast. As a result, the goal of this research is to apply Machine Learning methods to forecast the behavior of Kenya's banking sector stock market and to advise investors on the implications of investing in these companies.

1.3 Objectives

General Objective

The overall goal of the study was to use machine learning techniques to forecast banking sector security prices in Kenya.

Specific Objectives

The specific objectives of the study were to;

i. Develop an Autoregressive Integrated Moving Average (ARIMA) model for forecasting banking sector security prices in Kenya.
ii. Create an Artificial Neural Network (ANN) model for forecasting banking sector security prices in Kenya.
iii. Develop a Support Vector Machine (SVM) model for forecasting banking sector security prices in Kenya.
iv. Determine the most appropriate model for forecasting the banking sector security prices in Kenya.

1.4 Research Questions

The research questions of the study were;

i. Which ARIMA model can be used to forecast the banking sector security prices in Kenya?
ii. Which ANN model can be created to forecast the banking sector security prices in Kenya?
iii. Which SVM model can be developed to forecast the banking sector security prices in Kenya?
iv. Which of the models is considered the most appropriate for forecasting the banking sector security prices in Kenya

1.5 Hypotheses

The research hypotheses were;

i. There is no ARIMA model to forecast banking sector security prices in Kenya.
ii. There is no ANN model to forecast banking sector security prices in Kenya.
iii. There is no SVM model to forecast banking sector security prices in Kenya.
iv. There is no appropriate model to forecast banking sector security prices in Kenya.

1.6 Significance of the Study

Money supply is a market contributing element and therefore, this research may be critical in determining how the Central Bank of Kenya (CBK) will manage money supply. The CBK may gain from this research in its efforts to control Kenya's banking sector, as well as the country's money supply and spending. The Nairobi Securities Exchange (NSE), which is concerned with stock market results, may also benefit from the study. The results of this study will contribute to our knowledge and comprehension of the factors that are responsible for the wavelike motions in the returns of the Kenya's stock market.

Financial managers, policy makers, and investors operating within the investment industry will find this study valuable when it comes to deciding how they will allocate their investment capital. This research may help entrepreneurs and other types of managers better predict stocks that will have a big impact on their stock market performance. The study's findings will be helpful to academics, students, and researchers

since they will expand the body of knowledge already available on the use of machine learning in stock forecasting.

1.7 Justification of the Study

Time series forecasting is a field that researchers, particularly statisticians and economists, are interested in learning more about. In a market that is always expanding, stock shares are a very important instrument. When investors in the market buy and sell their shares in various markets both domestically and internationally, they play a significant part in the overall expansion of the market. Despite the expansion of market shares in Kenya, some investors are unsure whether it is the best time to make investments. This is because investing in a weak market at the right moment might result in profits, whilst investing in a strong market at the wrong time can result in losses.

Numerous methodologies for forecasting the future have been utilized in the process of making predictions about the behavior of the market. As a result of the progression of technology, more precise models have been constructed to carry out activities like these. In order to select the best model for predicting stock prices in the market, the aim of this study was to compare how well two distinct machine learning models — Artificial Neural Networks and Support Vector Machine models — predicted the prices of securities in Kenya's banking industry.

1.8 Scope and Limitation of the Study

Study limitation refers to factors that the researcher does not have any influence on, but which nevertheless have an effect on the conclusion and results of the study (Theofanidis

and Fountouki, 2018). This study utilized secondary data that was provided by NSE. Additionally, it did not include all of the banks that were listed by NSE; rather, out of a total of 12, only three banks were chosen to participate in the study

1.9 Contributions of the Thesis

In this study, the following are the contributions;

i. The study identified that SVM Model was appropriate for forecasting banking sector security prices in Kenya.
ii. This study contributed to the researches that exist in machine learning prediction techniques.
iii. The study further assists stock brokers to be able to predict accurately the share prices.

1.10 Organization of the Thesis

The five chapters of this thesis are as follows;

Introduction is the first chapter and explains statement of the problem, research objectives, research questions, research hypotheses, significance of the study, justification of the study, limitations of the study and finally the contributions of this research work.

The second chapter reviews existing literature on ARIMA, ANN and SVM models and gives a summary.

Chapter three covers the research methodology describing the research design used, population, sample and sample size and data analysis techniques.

Chapter four presents results and discussions for the three models and gives a summary.

The last chapter gives the conclusion of the study and recommendations.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter's main focus is a review of the significant prior research. In this chapter, there are three areas that are particularly important. This chapter's first section provides a theoretical analysis of the available research, while the second part presents an empirical analysis of the existing research with a particular emphasis on (i) an ARIMA model, (ii) an ANN model, and (iii) an SVM model. The third part is where the conclusion is laid forth.

2.2 Theoretical Review of Literature

The Random Walk Theory was used to predict security prices in this research. According to the hypothesis, the fluctuations in stock prices follow a normal distribution and are not reliant upon one another (Abdul Hamid, 2018). Prices here take a random walk and can be influenced by irrelevant data. According to this hypothesis, the only way for investors to achieve a higher return than the market average is for them to expose themselves to a greater degree of risk.

The theory was first introduced by Burton Malkiel in his book "A Random Walk Down Wall Street," published in 1973. In the book, the value of stocks is compared to a "drunk man's steps," which are unexpected (Malkiel, 1973). Investors are encouraged, under this line of thinking, to put their money into passive funds such as mutual funds. A random walk is a path that is known to have been formed by a series of random steps (Jay et al., 2020).

Given the starting point;

$$y_t = y_0 + \sum_{i=1}^{t} \varepsilon_i \tag{2.1}$$

where ε_i is IID at time step i and y_0 is the time to start.

Alternatively, the bootstrapping method can be utilized to discover a path exposed by a random walk that takes into account the system's most recent state y_{t-1};

$$y_{t-1} = y_0 + \sum_{i=1}^{t-1} \varepsilon_i \tag{2.2}$$

The next state of the system is obtained by taking one step forward, as demonstrated;

$$\mathrm{y_{t-1}} + \varepsilon_t = \mathrm{y_0} + \sum_{\mathrm{i=1}}^{\mathrm{t-1}} \varepsilon_i + \varepsilon_t \tag{2.3}$$

$$\mathrm{y_{t-1}} + \varepsilon_t = \mathrm{y_t} \tag{2.4}$$

For extremely large t, equation (2.4) is more efficient since it is more computationally efficient. This idea was applied in the research.

2.3 Empirical Review of Literature

2.3.1 Autoregressive Integrated Moving Average (ARIMA) Model

The ARIMA model was utilized by Abdullah Lazim in order to forecast gold bullion coin values. This indicated that gold bullion coin selling prices were on the rise and could be regarded a feasible investment (Abdullah, 2012). As' ad (2012) calculated peak day

electricity consumption by using half-hourly demand dates. For forecasting two to seven days in advance, the ARIMA model based on data from the prior three months was deemed to be the best, while the ARIMA model based on data from the prior six months was deemed to be the best for forecasting one day in advance. It asserted that the best model for forecasting two to seven days out was the ARIMA model, which was based on data from the previous three months.

Deepika et al. (2012) made an effort to investigate gold price forecasting using the ARIMA model and regression, but its findings indicated that a suitable model for gold price forecasting using the ARIMA model had not yet been identified. As a result, regression analysis was used in the later part of the study. Banerjee (2014) used the ARIMA model to predict how future stock indices would perform because they had a big influence on how the Indian economy was doing. The article began with the construction of the ARIMA model, continued with a forecast of the Sensex based on model validation, and concluded with recurrence validation.

The ARIMA and ANN models were applied to examine the impacts of traffic noise. In order to identify which was better for the modeling and forecasting of traffic noise, the research evaluated the ARIMA and ANN time series modeling methodologies. Regarding data processing and time series modeling, ARIMA models were flexible. The most catastrophic flaw in the model was its presumed linear shape (Garg et al., 2016). In addition, Gao et al. (2017) found that in their analysis for energy price forecasting in the UK electricity market, the ARIMA (4,1,2) model performed better than the ANN model. It used ANN with a variable number of delays and neurons along with ARIMA with

different AR and MA orders. It concluded that the ANN model was inferior to the ARIMA (4,1,2) model.

Almasarweh and Alwadi's research on an ARIMA model in predicting banking stock market data revealed the advantages of the Autoregressive Integrated Moving Average (ARIMA) model's forecasting accuracy. Data obtained from Jordan's Amman Stock Exchange were utilized in order to illustrate ARIMA's capacity to accurately forecast banking data. In order to generate the forecast, the authors looked at daily data from 1993 all the way up to 2017 and collected over 200 observations. The optimal ARIMA model was picked using the Mean Squared Error (MSE) metric as the deciding factor. Because of this, the ARIMA model was able to offer significant short-term prediction results (Almasarweh and Alwadi, 2018).

ARIMA, Kriging, Artificial Neural Network (ANN), Bayesian method, Support Vector Machine (SVM), and Random Forest were some of the machine learning models used in the Khedmatia et al. (2020)'s study to forecast the Bitcoin price. ANN, SVM, and RF models were among the multivariate models that were discussed, as were a variety of univariate models like ARIMA and ANN as well as Bayesian and Kriging models. It was utilized on the Bitcoin price from December 16, 2017 to June 1, 2018 in a multivariate model with the Bitcoin price as well as the maximum, minimum, and closing values for the same time period. They found that ARIMA outperformed other univariate models because it had lower RMSE and MAPE values when comparing the performance of the proposed models using RMSE and MAPE measurements.

Chaudhuri and Pandit (2021) examined how successfully time series models predicted earnings for six companies from January 2010 to December 2020. National Stock Exchange supplied monthly average stock figures for HCL, TCS, Infosys, Reliance, Tech Mahindra, and Wipro for 11 years. Every organization plotted and tested 132 values for stationarity. Each company's data series was non-stationary. After differentiating each series, graphical charting resumed. The best ARIMA model for each stationary time series was then chosen using goodness-of-fit statistics. The residuals were random and had no external impacts after picking the best ARIMA model. In terms of mean percentage errors, AIC, and average ranks, superior ARIMA family models beat naive time-series models. The findings showed investors should use the ARIMA model.

2.3.2 Artificial Neural Network (ANN) Model

Chavan and Patil looked at a wide range of model input parameters that were identified in nine different published studies in order to gain a deeper comprehension of ANN stock market prediction. They were searching for the most important input factors that led to a more accurate forecast from the model. They found that the bulk of machine learning systems estimate stock prices by using technical aspects rather than fundamental ones. On the other hand, it is common practice to estimate the values of stock market indices using microeconomic factors. Additionally, using hybridized parameters produced better outcomes compared to the practice of using only one type of input variable (Chavan and Patil, 2013).

Kihoro and Okango used the Equity bank share prices historical data to estimate future values in the study to predict stock market prices using ANN. They believed that only

past prices affected future prices, so it applied ARIMA models to stock prices to determine the appropriate input delays for the ANN model. In the study, the greatest results in terms of the least Mean Squared Error (MSE) between the projected values and the test data were obtained by selecting the best combination of input lags. It also discovered that employing lags with little correlations lowered the MSE. The research provided the best results when it used the 3-3-1 network architecture (Kihoro and Okango, 2014).

Wanjawa and Muchemi investigated the possibility of utilizing a Neural Network model in order to make predictions on the values of stocks traded on the Nairobi Securities Exchange (NSE). It came up with a feed-forward multi-layer perception (MLP) artificial neural network model that included error back-propagation. The end result was a configuration with the numbers 5:21:21:1, which indicates that it has five inputs, two hidden layers, each with 21 neurons, and just one output. They developed a prototype using the C# programming language and put it through its paces by feeding it data from daily NSE trades collected over a period of five years, starting in 2008 and ending in 2012. The algorithm was able to accurately forecast the long-term movements of three different equities, achieving a Mean Absolute Percentage Error of less than 1 percent in the process (Wanjawa and Muchemi, 2014).

Rasel et al. (2016) made the discovery that the ANN model offers greater benefits than other SVM or LR models do. One of the benefits was an increased level of precision enabled by a diverse set of features. It was successful despite the fact that there did not appear to be any clear connection between the quality and the production. There were some drawbacks that needed to be taken into consideration as well. The amount of time

that was necessary for prediction was longer than the amount of time that was necessary for other methods. It is possible that one can experience difficulties due to excessive fitting.

Deep learning networks were investigated by Chong et al. (2017) for the purpose of stock market analysis and forecasting. It gave a frank assessment of the advantages and disadvantages of using deep learning algorithms for stock market research and forecasting. It looked into how the network's performance in terms of its capacity to forecast future market behavior was affected by the inclusion of high-frequency intraday stock returns as input data. It looked into three unsupervised feature extraction techniques: limited Boltzmann machine, auto-encoder, and principal component analysis. Data from 38 companies that traded on the KOSPI stock exchange between January 4, 2010, and December 30, 2014 were used in the study. According to the findings of empirical research, deep neural networks have the potential to enhance prediction accuracy while simultaneously extracting more information from the residuals of autoregressive models.

Sivasamy et al. (2017) devised a straightforward approach for discovering the best times to make trades in the stock market by combining the time series and trading rule methodologies. Using ARIMA and ANN methods, they fitted the best model to the observed time series dataset corresponding to closing prices of a single stock. After that, the effectiveness of each model was assessed by calculating its rate of return over a predetermined amount of time. It found that the ANN technique not only produced the greatest fit to the initial series, but ANN also provided the best average future returns for the data set that was employed. When compared to the use of moving averages just on the original technical data, the ARIMA approach and the ANN method were found to be more

effective. ANN was superior to methods based on time series models that relied on a number of assumptions despite the fact that it did not make any parametric assumptions, which was one of the reasons why it was so appealing.

2.3.3 Support Vector Machine (SVM) Model

Schumaker and Chen conducted ground-breaking research that used textual analysis and a SVM to look at how news stories affect stock prices. Using a variety of textual representations, such as Bag of Words, Noun Phrases, and Named Entities, among others, it built a predictive machine learning technique for the analysis of financial news stories. It examined a sizable number of financial news articles and stock quotes from the S&P500 over the course of five weeks, from October 26, 2005 to November 28, 2005. It anticipated a new stock price twenty minutes following the publication of a news item. It was able to show that the model that took into account the terms of the article and the stock price at that time offered the most accurate estimate of the future stock price, the most accurate prediction of the future price movement, and the highest return on their investment. It did this with the aid of a SVM derivative for discrete numeric prediction (Schumaker and Chen, 2009).

Das and Padhy employed BP and SVM to forecast future stock market prices in India. These researchers used data from the National Stock Exchange that was gathered between January 1, 2007, and December 31, 2010, to assess a machine learning model. When comparing the two methods, it was found that SVM performed better than BP. The implementation made use of MATLAB and SVM Toolkits (LS-SVM Tool Box). According to the experiment's performance criteria, the predicted price and the actual

price were quite close to each other when the SVM method was used. The Normalized Mean Squared Error ranged from 0.9299 to 1.1521 for all futures stock indices. Directional Symmetry requirements ranged from 0.2379 to 0.3887, while the MAE requirements ranged from 55.17 to 91.2512 (Das and Padhy, 2012).

Rajput and Kaulwar employed ANN, PCA, and SVM to predict the future of an Indian stock. It used SBI and Larsen & Toubro time series data. State Bank of India's models were trained and tested in 2012 and 2013, whereas Larsen & Toubro's were trained and tested in 2015 and 2016. PCA reduces highly correlated, high-dimensional data to linearly uncorrelated, low-dimensional data. According to the PCA, the first four primary components of the study reflected 98.24% of State Bank of India data and 99.27% of Larsen & Toubro Limited data. State Bank of India data had a 0.03 Normalized Mean Squared Error and a 2.12% error, whereas Larsen & Toubro Limited data had a 0.09 NMSE and a 2.41 error when using the Nonlinear Autoregressive with External Input prediction model using PCA data. Similar to State Bank of India, SVM gave Larsen & Toubro Limited data a Normalized Mean Squared Error of 0.04 and a 1.91 percent error rate respectively. SVM outperformed ANN with PCA in computational tests (Rajput and Kaulwar, 2019).

Vijh et al. (2020) suggested that a SVM model has the potential to accurately check the stock closing price of both Tesla Inc., a technology company, and Reliance Ltd., a public corporation. The SVM Kernel models for both companies were trained and evaluated using historical stock data, which encompassed the time period from November 2019 through November 2020. It came to the conclusion that the performance of every Support Vector Regression kernel approach was affected differently by a variety of data points.

The Radial Basis Function (RBF) Support Vector Regression kernel was the best possible solution for both Tesla Inc. and Reliance Ltd. RBF anticipated the stock closing prices that were closest to the original values on the same days they occurred, as compared to the original values.

Shaheen and Arshad tackled some of the problems that crop up when trying to anticipate stock market values using support vector regression while simultaneously juggling kernel function hyperparameters. The study was successful in incorporating the positive aspects of a number of different hyperparameter settings into the system, which led to improved performance across the board. It developed a two-stage multiple-kernel learning strategy by employing sequential minimum optimization and the approach of gradient projection. The improved strategy outperformed earlier techniques significantly when datasets from the Taiwan Capitalization Weighted Stock Index were used in comparisons (Shaheen and Arshad, 2020).

Dai and Zhao developed a forecasting model to foresee changes in the stock market using a support vector machine and a hybrid feature selection method. A hybrid feature selection methodology, F-score and Supported Sequential Forward Search, is advised to select the best subset of features from the initial feature collection. They evaluated the precision of this SVM-based model integrated with F-score and Supported Sequential Forward Search using paired t-tests against a backpropagation neural network and three commonly used feature selection approaches (information gain, symmetrical uncertainty, and correlation-based feature selection). The SVM outperformed the backpropagation neural network in terms of forecasting stock trends. Experiments revealed that the proposed SVM-based model coupled with F-score and Supported Sequential Forward Search had the highest

degree of prediction accuracy and generalization when compared to the other three feature selection procedures (Dai and Zhao, 2020).

2.4 Summary

In conclusion, after conducting a literature analysis, it is clear that there have been effective attempts made to predict stock prices using machine learning algorithms. This can be deduced from the fact that these attempts have been successful. However, rather than using Kenya's banking sector, the majority of these studies have built their analysis of stock market activity using share indices and other markets. The machine learning methods described in the literature review for forecasting the securities in the banking industry have not been compared in any research.

Therefore, the goal of this work was to predict the values of Kenyan banking sector securities using two distinct machine learning approaches, namely ANN and SVM, assess the predictive abilities of both models, and then suggest a model that would be suitable for doing so.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter outlines the suggested format for the examination in light of the study that was conducted. It discusses the research design, the population that was targeted, the sampling strategies that were employed, the sample size, as well as the procedure and equipment for collecting data.

3.2 Research Design

A research design is a method which ensures a problem statement is satisfactorily addressed while also allowing for the integration of a variety of different aspects of an investigation in a logical and cohesive manner (Lai, 2018). The methods of data collection, measurement, and analysis are broken down even further and more specifically in a research design.

A descriptive approach was taken for this study's research design. The goal of descriptive research is to provide a comprehensive picture of the subjects of the study, be they people, events, or situations (Ros et al., 2020). The descriptive research approach was an ideal choice for this study as it permitted the researcher to provide the material as it was observed rather than providing statistical data. The study made use of quantitative data collection strategies, as well as data analysis techniques that included the utilization of graphs to either produce or make use of numerical data.

3.3 Target Population

A collection of people or things that will be the subject of an investigation is referred to as a research population (Tang et al., 2018). According to Westlake et al. (2017), the people, services, components, and groups of items that are the focus of the research make up the population. A collection of individuals or objects from whom a statistical sample is drawn for the purpose of research can also be referred to as a "population." A population is represented in mathematics by the letter N. The securities of the banking sector that are listed on the Nairobi Securities Exchange (NSE) served as the population of interest for this study.

3.4 Sampling Techniques

Saunders et al. (2009) defines a sampling frame as "an exhaustive description of all relevant data from the target population from which a sample will be picked." A sampling frame was required because it was the factor that determined whether or not the sample accurately reflected the target population. For this research, the sample size was determined via purposive sampling. A random selection of sampling units within the population segment with the greatest amount of information on the characteristic of interest is referred to as "purposive sampling" (Campbell et al., 2020). This study concentrated on twelve banks based in Kenya that are members of the Nairobi Securities Exchange (NSE).

3.5 Sample Size

When used for study, a sample is a carefully chosen subset of a larger population that offers details about the larger population as a whole. It is also possible to define it as a part of a population that is statistically significant and denoted by the symbol n (de Souza et al., 2022). It is possible, based on the information obtained from the statistical sample, to draw inferences about the full population. In order to determine the appropriate size of the sample for the research project, the study applied purposive sampling technique.

By taking a look at 12 banks that are listed on the NSE, the purpose of this study was to analyze several forecasting approaches that are based on the traditional ARIMA and ML methods in order to estimate the stock market values of the banking sector in Kenya. According to the findings of a study that was carried out by Kenyan Magazine 2022, the following three NSE-listed banks were regarded as the most successful banks in Kenya in terms of their overall performance: Kenya Commercial Bank (KCB), Equity Bank, and Cooperative Bank of Kenya (Kenyan-Magazine, 2022). The three banks were used as the sample size for the study by use of purposive sampling technique.

3.6 Data Collection Procedure

Collection of data refers to the process of accumulating information for the purpose of answering the questions that sparked the idea for the study (Bilsborrow, 2016). This study made use of secondary data, which Saunders et al. (2009) define as information that was initially obtained for another purpose and that can be further analyzed to produce new or different knowledge, interpretations, and conclusions. The secondary data for the study came from the historical financial reports that the NSE had published on the banks' daily

market performance over the course of a period of six years, starting in January 2016 and ending in December 2021.

3.7 Research Procedures

The research process, also known as research procedures, refers to the steps done by the researcher in order to gather the data that is necessary for the study (Crotty, 2020). In light of the fact that this study relied on secondary data, the information was obtained from the Nairobi Stock Exchange. The NSE provided the study with a sufficient quantity of precise secondary data.

3.8 Data Analysis Methods

This study involved the use of descriptive and inferential statistics in analyzing data which helped to compare different ML algorithms for forecasting banking sector securities. The variables under study were the three banks which were chosen via purposive sampling technique which had high, low and closing prices. The data was coded for analysis using Python as well as R Studio software.

For this study frequencies and percentages were used as well as measures of central tendency, presented through graphs and tables to organize data for easy reference and communication. Inferential statistics refers to methods of drawing conclusions from sample data about a population (Sutanapong and Louangrath, 2015). For this study, the forecasting techniques that were used encompassed the classical ARIMA approach and the ML algorithms, that is, ANN and SVM.

3.8.1 Fitting an ARIMA Model

To fit an ARIMA model, this study incorporated the Box Jenkins methodology which was as follows (Box and Jenkins, 2013):

i. Stationary check.

ii. The ADF (Augmented Dickey Fuller) / Unit Root test was carried out. The null hypothesis of the presence of a unit root was rejected if the test's p-value was less than the (α) level of significance, implying that the series was stationary.

iii. To find the parameters of the ARIMA model, the study applied auto_arima function in "pmdarima" package.

iv. Choosing the best ARIMA model with minimum value of AIC, since AIC is a measure of goodness fit of the model. The smaller the value of AIC, the better is the fit.

v. Performing "Residual Analysis" on the chosen model. In this case, the Ljung Box test was used, and if the p-value was higher than the significance level (α), the conclusion was that the model fitted the data well.

vi. Forecasting was done based on the chosen ARIMA model.

The auto_arima function was used to determine the values of p, d and q in the ARIMA (p, d, q) model. These values were then used in the model. AIC was employed as a criterion in order to select the most appropriate ARIMA model to utilize. This is due to the fact that multiple ARIMA models are able to be constructed for a single column of data by making use of various values for the parameters p, d, and q. As a consequence of this, the ARIMA model with the lowest AIC was the one that suited the data the best. In

addition to that, the model's accuracy was measured using various metrics. The root mean square error (RMSE) is a standard method that can be used to calculate the accuracy of a model's prediction of quantitative data.

Mathematically, RMSE is determined as (Chai and Draxler, 2014);

$$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{3.1}$$

where y_i represents the i^{th} ascertained value, $\hat{y}_i$ represents the i^{th} fitted value and n represents the total number of validation data.

The first hypothesis was tested using the Ljung Box Test. The null hypothesis of the absence of an ARIMA Model was rejected if the test's p-value was greater than the (α) level of significance, implying that the series is stationary. The "pmdarima" package and the "auto_arima()" function was used in Python to execute the test.

3.8.2 Fitting an ANN Model

In this study, an ANN model was fitted with the assistance of a Multilayer Perceptron, also known as an MLP. MLPs are types of artificial neural networks (ANNs) that feature a countable continuous layer structure (Taud and Mas, 2018). An MLP consists of three layers: the input layer, the hidden layer, and the output layer. There are frequently found to be additional hidden layers within the MLP structure that are concerned with approximative solutions to challenging problems such as fitting approximation. A multi-layer perceptron, often known as an MLP, is a type of mapping that converts input vectors

into output vectors. It consists of a number of node layers, each of which is connected to the one that comes after it (the "next layer") (Ramchoun et al., 2016). In addition to the nodes that serve as inputs, each node itself is a neuron that has a nonlinear activation function.

The neural network was trained using the back propagation approach. The following is a detailed examination of the BP algorithm:

i. For any set of input variable $X = (x_1, x_2, \cdots, x_d)$, initial weight of input layer $w^{(1)}$ is randomly generated to calculate the value which is put into hidden layer, and d represents the number of cells of input layer. Therefore, the output value of input layer α_j is:

$$\alpha_j = \sum_{i=1}^{d} w_{ij}^{(1)} x_i \tag{3.2}$$

ii. Tanh function as the activation function of hidden layer is used, therefore the output result of the hidden layer h_j is:

$$h_j = g(\alpha_j) \tag{3.3}$$

where $g(x)$ is defined in equation (1.8).

iii. The linear weighted aggregative method is still applied into output layer, the weight is $w^{(2)}$, m stands for the cell number of output layer, and the final output y_k is:

$$y_k = \sum_{j=1}^{m} w_{jk}^{(2)} h_j \tag{3.4}$$

iv. The modified error of each cell can be determined using the output value, y_k, and the real value, C_k, and the formula is:

$$d_k = (y_k - C_k) * C_k * (1 - C_k) \quad (3.5)$$

v. According to such modified error the initial weight can be adjusted, repeating steps from 1 to 4, and defining target function. In other words, when the sum of squares of the difference between the output value and the real value is less than a pre-determined value, the training procedure ends. Furthermore, when the number of learners exceeds a pre-determined threshold, the training process comes to a halt.

During training, data was split into two categories: Training accounted for 80%, while testing accounted for 20%. The "train_test_split" function in python was used to train and test the model. The model with the lowest Root Mean Square Error (RMSE) was picked as the best forecasting model. Mathematically, RMSE is defined as in equation (3.1).

This study employed the Chi-Square test procedure while testing null hypothesis for the ANN model. In R Studio, the “chisq.test()” function was used for the test procedure. The Chi-Square test procedure is as follows;

i. Formulate the hypotheses; Null and Alternative hypothesis.

ii. Specify the predicted values for each cell if the null hypothesis is true.

iii. To see if the data provides strong evidence against the null hypothesis, compare the observed value from the sample with expected numbers, assuming, H_o is true.

iv. Calculate the result of the test statistic. The following is the definition of the test statistic:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \tag{3.6}$$

v. Make inference by determining whether the computed test statistic is large enough to reject H_0.

3.8.3 Fitting an SVM Model

The kernel approach was used to fit a SVM model, and it is used to classify non-linear relationships between class and sample feature vectors (Zoppis et al., 2019). As a result, SVM is an effective technique for detecting non-linear relationships. Mathematically, a kernel function is defined as;

$$K(x, y) = \varphi(x).\varphi(y) \tag{3.7}$$

The "train_test_split" function in python was used to divide the dataset into 80% for training and 20% for testing. Then SVM type 2 with the following error function was chosen:

$$\frac{1}{2}\omega^T\omega - C\left[\upsilon\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\zeta_i + \zeta_{i.}\right] \tag{3.8}$$

As a result, the entity was reduced to;

$$
\begin{aligned}
&[\omega^T \phi(x_1) + b] - y_i \leq \varepsilon + \zeta_i \\
&y_i - [\omega^T \phi(x_1) + b_i] \leq \varepsilon + \zeta_{.i} \\
&\zeta_i, \zeta_{.i} \geq 0, i = 1, \cdots, N, \varepsilon \geq 0
\end{aligned} \tag{3.9}
$$

In SVM classification, the Radial Basis Function (RBF) Kernel was used to map input space to output space and was based on the distance between the input and some of the fixed points in the data. It is defined mathematically as (Ding et al., 2021);

$$
K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2v^2}\right) \tag{3.10}
$$

where v, is the width of the RBF.

Therefore, the optimal regression function was;

$$
f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x_j) + b \tag{3.11}
$$

where $b = f(y_i) - \sum_{i=1}^{N} \alpha_i \cdot x_j \cdot K(x_i, x_j)$.

The predicted values were compared to real sample data outputs to determine the model's prediction accuracy. The RMSE was used to evaluate the model's accuracy in making predictions. The best fit was determined by the model with the smallest RMSE.

To test the hypothesis for the SVM model, this study applied the Chi-Square test. The test procedure is described in section 3.8.2.

3.8.4 Choosing the appropriate Model

The study compared the performance of each model using four accuracy metrics - Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). The model with the minimum errors was considered as the most appropriate.

In addition, for each machine learning model, a confusion matrix was developed in this research so that the appropriate model could be chosen. A confusion matrix is a table that shows how well a classification model performs on a set of test data with known true values (Visa et al., 2011). It's a two-dimensional table with actual value and predicted value columns. An error matrix is another name for a confusion matrix.

There are four dimensions to a confusion matrix.

- True Positive (TP): The data point's actual and projected classes are both 1 in this situation.
- True Negative (TN): The data point's actual and anticipated classes are both 0 in this situation.
- False Positive (FP): The real data point class is 0 in this situation, but the predicted data point class is 1.
- False Negative (FN): The real data point class is 1 in this situation, while the projected data point class is 0.

Let, C_{NXN} be the confusion matrix for N samples, with rows representing real class labels and columns representing anticipated class labels. Then, for a particular class label i

$$TP = C_{ii}$$

$$FP = \sum_{j=1}^{N} C_{ji} - C_{ii}$$

$$FN = \sum_{j=1}^{N} C_{ij} - C_{ii} \tag{3.12}$$

$$TN = \sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij} - (FP + TP + FN)$$

The Confusion Matrix can be illustrated as;

Table 3.1: An Illustration of a Confusion Matrix

	N	**PROJECTED**	
		POSITIVE	*NEGATIVE*
EXACT	*POSITIVE*	True Positive (TP)	False Negative (FN)
	NEGATIVE	False Positive (FP)	True Negative (TN)

Upon acquiring the above values, the study will evaluate each approach's performance using the five metrics listed below: (i) precision, (ii) recall, (iii) F1 Score, and (v) Matthew's correlation coefficient (MCC).

$$Accuracy = \frac{TP + TN}{N}$$

$$Recall = \frac{TP}{FN + TP}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.13}$$

$$F1\ Score = \frac{2(TP)}{2(TP) + FP + FN}$$

$$MCC = \frac{(TP\ X\ TN) - (FP\ X\ FN)}{\sqrt{(TP + FP)\ X\ (TP + FN)\ X\ (TN + FP)\ X\ (TN + FN)}}$$

3.9 Summary

The methodology involved discussion of the sampling techniques used in the study. It further discussed the sample size that was chosen for the study as well as the data analysis techniques. The Box Jenkins methodology for fitting ana ARIMA model, the BP algorithm for training an ANN and the Kernel method for fitting a SVM are also discussed in the chapter. Finally, the confusion matrix is alongside critical measures are also discussed herein.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

There are a limited number of ways to implement machine learning algorithms. MATLAB, Python, and R are the most commonly used programming languages. Python and R were utilized in this study because of their wide library of libraries and ease of implementation. The tools and techniques are discussed in detail in this chapter. Scikit-Learn, Pandas, Numpy, matplotlib, e1071, caret, and Keras were used to analyze data, build models, and appropriately visualize data. These tools were a big help. The model fitting and outcomes of the techniques utilized in this research are explained in this chapter. The results and evaluations of the three models studied were presented. Each method's results are displayed, as well as which model is the most effective among them.

4.2 Descriptive Analysis

Table 4.1 displays the descriptive statistics of the study. A total count of 1512 observations were made from the year 2016 to 2021. For Equity Bank shares, the mean of the observations was 40.22 while the standard deviation being 6.76 as observed in Table 4.1. Similarly, for KCB, the mean and standard deviation for the observations were 40.10 and 6.42 respectively. Moreover, a mean of 14.40 and a standard deviation of 2.38 were computed for Cooperative (CO-OP) bank.

Table 4.1: Descriptive Statistics

	EQUITY	KCB	CO-OP BANK
count	1512.00	1512.00	1512.00
mean	40.22	40.10	14.40
std	6.76	6.42	2.38
min	23.50	23.00	9.750
25%	36.10	36.75	12.35
50%	39.75	40.00	13.95
75%	43.75	44.41	16.36
max	56.00	55.00	21.00

Figure 4.1 represents a time series plot of the dataset.

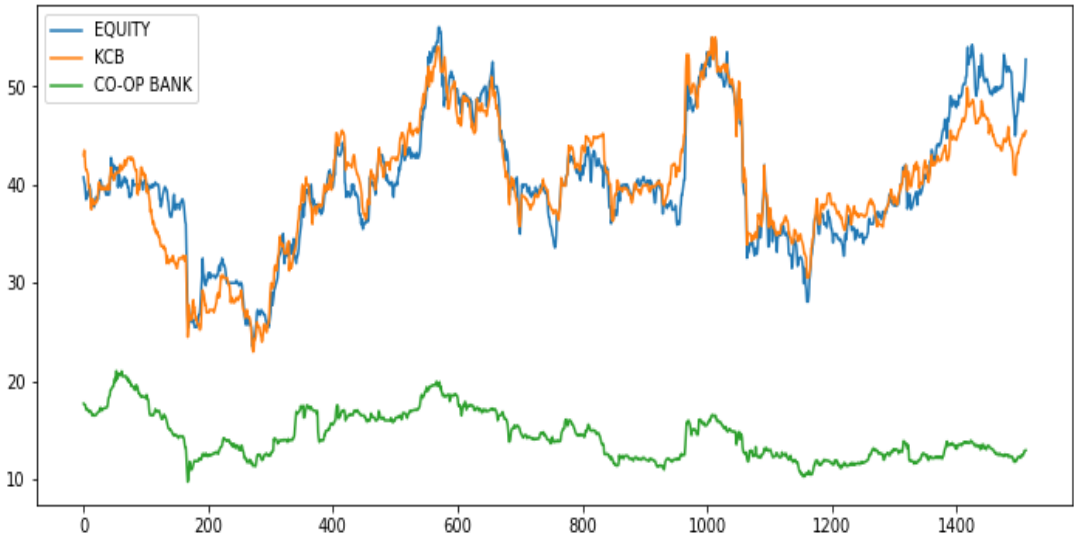


Figure 4.1: Time Series plot of the Dataset

It is evident that from Figure 4.1 above, Equity bank stock prices were higher than those for the other banks. Moreover, Equity and KCBs' shares rose from 2017 to 2018 and then depreciated in the year 2019. A steep fall was realized in 2020 due to corona virus outbreak in Kenya. Since then, the prices have risen as evident from the figure Cooperative (Co-op) bank's shares were lowest as compared to the others. These were highest in 2016 and have fallen slowly to 2021.

4.3 Autoregressive Integrated Moving Average (ARIMA) Model

From the plot in Figure 4.1, the shares prices had a specific trend. The prices of the stocks rose to higher values although for Co-op bank, there was a low trend for the securities. When the interest rate cap was affected, share prices for the banks dropped significantly and this means that interest rates are a factor that affects the decision to invest in bank shares. To ensure that the data was good for modeling, a stationarity check was conducted for both banks using Augmented Dickey Fuller (ADF) test.

Table 4.2: Results of dickey fuller test

	EQUITY BANK	**KCB BANK**	**CO-OP BANK**
Test Statistics	-2.09	-2.16	-2.33
p-value	0.25	0.22	0.16
No. of lags used	2.00	2.00	3.00

Number of observations used	1509.00	1509.00	1508.00
critical value (1%)	-3.43	-3.43	-3.43
critical value (5%)	-2.86	-2.86	-2.86
critical value (10%)	-2.57	-2.57	-2.57

The data was not stationary since no p-value from the aforementioned results was smaller than the standard value of 0.05. In order to make the data steady, differencing was required. The data's stationarity was attained by only differencing them once. A visualization of the dataset after it has been differenced is shown in Figure 4.2. The resulting dataset was therefore suitable for modeling for ARIMA.

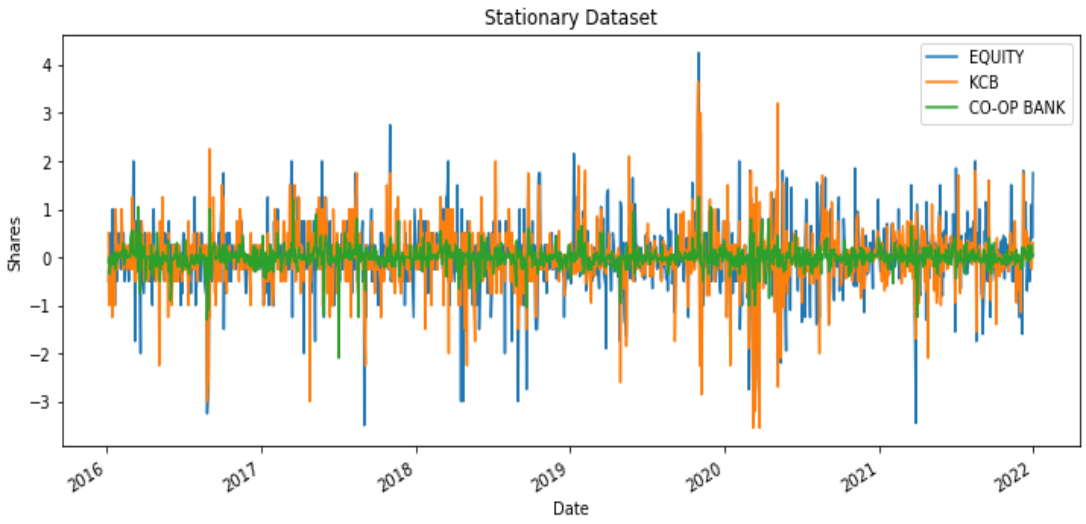


Figure 4.2: Stationary dataset

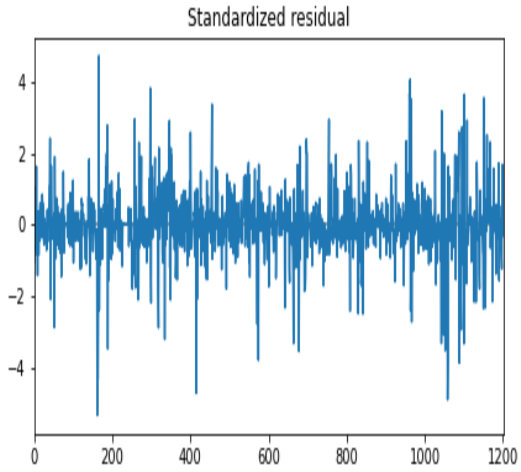
A spike at lag 1 in a stationary dataset denotes a strong correlation between each value in the series and the value that came before it. A spike at lag 2 shows a strong correlation between each value and the value that came two points earlier, and so on. Figure 4.2 shows that the model is clustered around zero for low spikes and that higher spikes represent higher lags.

The data was split to 80% for training and 20% for testing and modeling for ARIMA model was developed and trained using the shares from the train data.

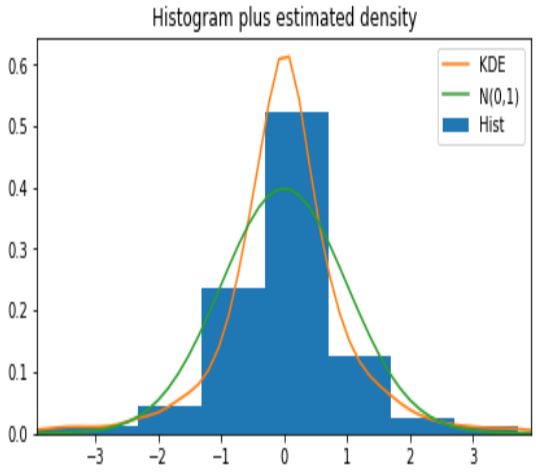
4.3.1 ARIMA Model for Equity Bank

Using the auto_arima function, it was possible to select the best ARIMA model's $p, d,$ and q parameters without glancing at the Autocorelation Function (ACF) and Partial Autocorelation Function graphs. Once the most optimal ARIMA model parameters had been determined, the auto_arima function delivered a fitted ARIMA model.

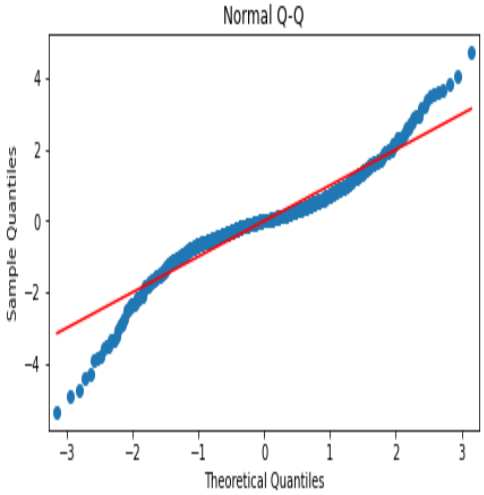
Differentiating tests (e.g., Kwiatkowski Phillips Schmidt Shin, Augmented Dickey Fuller, or Phillips Perron) were performed in order to determine d, and then models were fitted within the starting p to the maximum p and starting "q" to the maximum "q" p-ranges. If seasonal differencing was enabled, auto_arima conducted the Canova Hansen to establish the best order of seasonal differencing, d, as well as the optimal p and q hyperparameters.



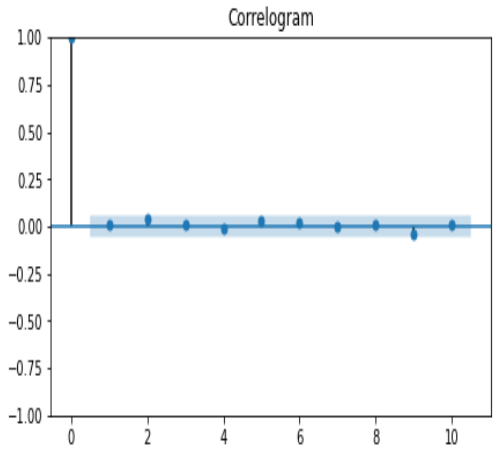
a. Standardized Residuals



b. Histogram plus Estimated Density



c. Normal Q-Q Plot



d. Correlogram Plot

Figure 4.3: Equity Bank's ARIMA Model Diagnostics

After the function processed the data, the best model generated was ARIMA (0,1,1) which implies $p = 0, d = 1, q = 1$. This was chosen since the model had the minimum AIC. However, there was no sign of seasonality in the dataset and therefore no need of examining the seasonality aspect of the data. Moreover, from the results in Figure 4.3;

Figure 4.3a: The standardized residual errors appeared to have a uniform variance and fluctuate around a mean of zero.

Figure 4.3b: The density plot suggested a normal distribution with a mean of zero.

Figure 4.3c: The red line should be perfectly aligned with all of the dots. Any significant deviations indicated a skewed distribution.

Figure 4.3d: The Correlogram, often known as the ACF plot, demonstrated that the residual errors were not autocorrelated. Any autocorrelation implied a pattern in the residual errors that the model couldn't account for. As a result, an addition of X's (predictors) could be needed in the model.

Fitting was then done to the data as the auto_arima function had assigned $p = 0, q = 1$ and $d = 1$ to the model and the results are tabulated in the Table 4.3.

Table 4.3: Equity Bank's ARIMA Model Summary Results

	Coefficient	**S.E**	**z**	**P>|z|**	**C.I (0.025)**	**C.I (0.975)**
MA (1)	0.2865	0.017	16.927	0.000	0.253	0.320
σ^2	0.0003	7.12e-06	43.335	0.000	0.000	0.000

From the above results, the ARIMA (0,1,1) model performed quite well with a probability of 0.320 for a two-sided test. The resulting model for Equity Bank share prices was therefore an MA of order 1 and the resulting equation is given by;

$$Y_t = \mu + \beta_1 \varepsilon_{t-1} + \varepsilon_t \quad (4.1)$$

where $\mu = 40.22$, as obtained in Table 4.1 and $\beta_1 = 0.2865$ as obtained in Table 4.3.

Therefore;

$$Y_t = 40.22 + 0.2865\varepsilon_{t-1} + \varepsilon_t \quad (4.2)$$

From the above results, the model's coefficient had a standard error of 0.017 which is significant compared to the standard p-value of 0.05. This therefore implied that the model was significant for the Equity Bank Shares.

Moreover, Ljung Box test gave the results in Table 4.4 below;

Table 4.4: Ljung Box test results for Equity Bank

Ljung-Box (L1) (Q):	0.11
Prob(Q):	0.74
Heteroskedasticity (H):	1.25
Prob(H) (two-sided):	0.03

From the results above, Prob(Q)=0.74 is greater than Q=0.11. Therefore, the null hypothesis was rejected implying that there exists an ARIMA model for forecasting Equity bank security prices.

A forecast for the Equity Bank Shares was then performed with a 95% confidence interval which was then made on the testing dataset and yielded the results in Figure 4.4 below.

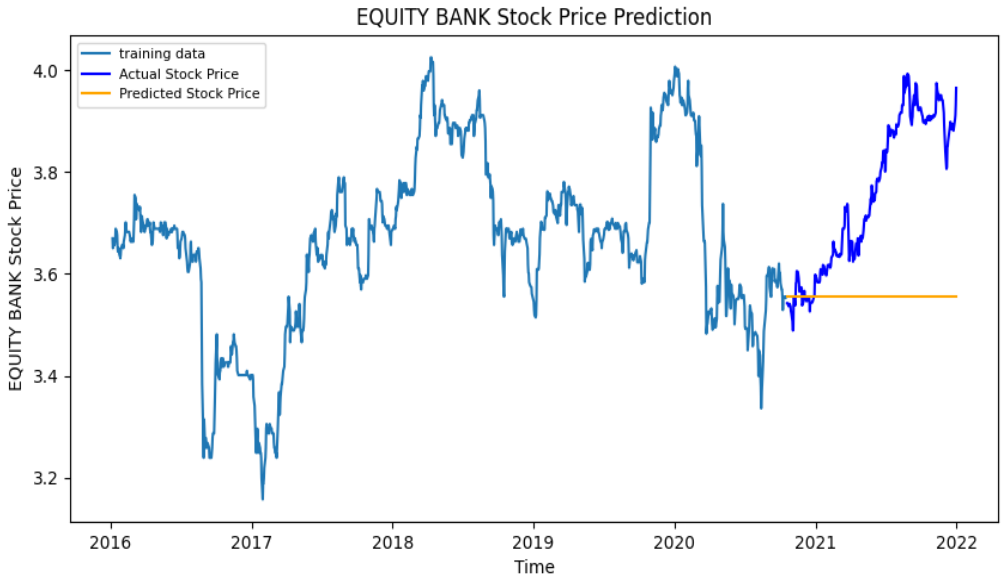


Figure 4.4: Equity Bank Stock Prediction Using ARIMA Model

The outcomes that were attained following accuracy measurements for the dataset in order to assess the model's capacity for prediction are as follows:

MSE: 0.06259 MAE: 0.2044 RMSE: 0.2501 MAPE: 0.0529

According to the RMSE number, it is possible to conclude that the performance of the model was generally satisfactory. Furthermore, according to a general rule of thumb, a RMSE value that falls within the range of 0.2 and 0.5 indicates that the model is able to relatively accurately forecast the data.

4.3.2 ARIMA Model for KCB Bank

Similar to Equity Bank, determining the appropriate model for the dataset was done by the auto_arima function in python and yielded an ARIMA model of order (0,1,1) implying that it was a Moving Average model of order 1.

The summary results of the model are presented in the Table 4.5 below.

Table 4.5: KCB's ARIMA Model Summary Results

	Coefficient	**S.E**	**z**	**P>|z|**	**C.I (0.025)**	**C.I (0.975)**
MA (1)	0.2858	0.014	20.621	0.000	0.259	0.313
σ^2	0.0003	5.26e-06	50.149	0.000	0.000	0.000

From the results in Table 4.5 above, the fitted model for the KCB dataset is given as;

$$Y_t = \mu + \beta_1 \varepsilon_{t-1} + \varepsilon_t \quad (4.3)$$

where μ= 40.1, as obtained in Table 4.1 and β1=0.2858 as obtained in Table 4.5.

Therefore;

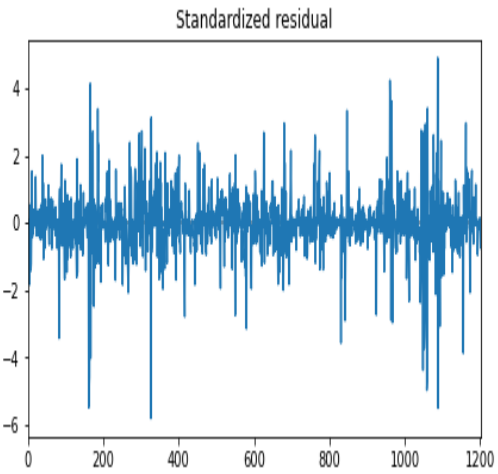
$$Y_t = 40.1 + 0.2858\, \varepsilon_{t-1} + \varepsilon_t \quad (4.4)$$

Moreover, the KCB's ARIMA model diagnostics and Ljung Box test gave the results presented in Table 4.6 and Figure 4.5 respectively;

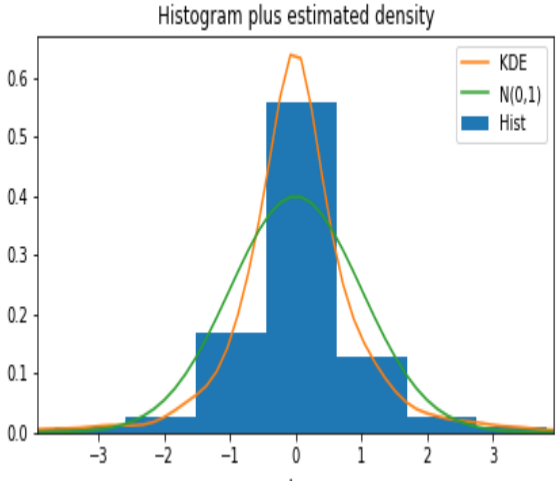
Table 4.6: Ljung Box test results for KCB

Ljung-Box (L1) (Q):	0.00
Prob(Q):	0.96
Heteroskedasticity (H):	1.04

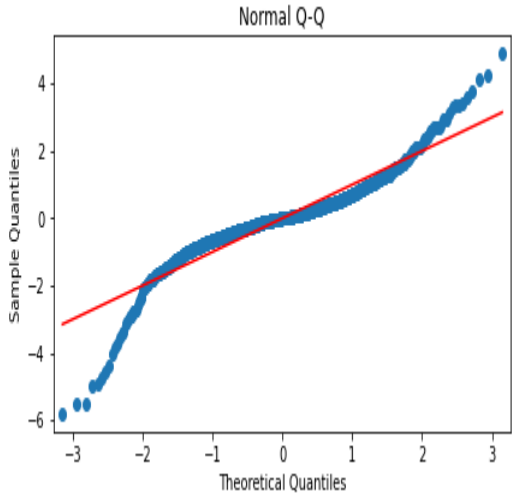
Prob(H) (two-sided):	0.67



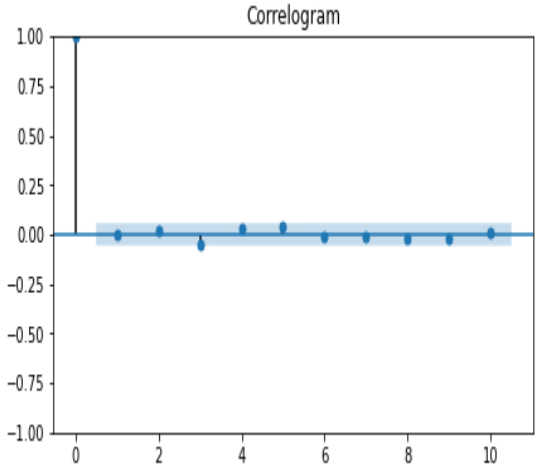
a. Standardized Residuals



b. Histogram plus Estimated Density



c. Normal Q-Q Plot



d. Correlogram Plot

Figure 4.5: KCB's ARIMA Model Diagnostics

From the results in Table 4.6 above, Prob(Q)=0.96 is greater than Q=0.00. Therefore, the null hypothesis was rejected implying that there exists an ARIMA model for forecasting KCB security prices.

In addition, from the results in Figure 4.5;

Figure 4.5a: The standardized residual errors appeared to have a uniform variance and fluctuate around a mean of zero.

Figure 4.5b: The density plot suggested a normal distribution with a mean of zero.

Figure 4.5c: The red line should be perfectly aligned with all of the dots. Any significant deviations indicated a skewed distribution.

Figure 4.5d: The Correlogram, often known as the ACF plot, demonstrated that the residual errors were not autocorrelated. Any autocorrelation implied a pattern in the residual errors that the model couldn't account for. As a result, an addition of X's (predictors) could be needed to the model.

Moreover, forecasting was then done for the dataset and these were the results.

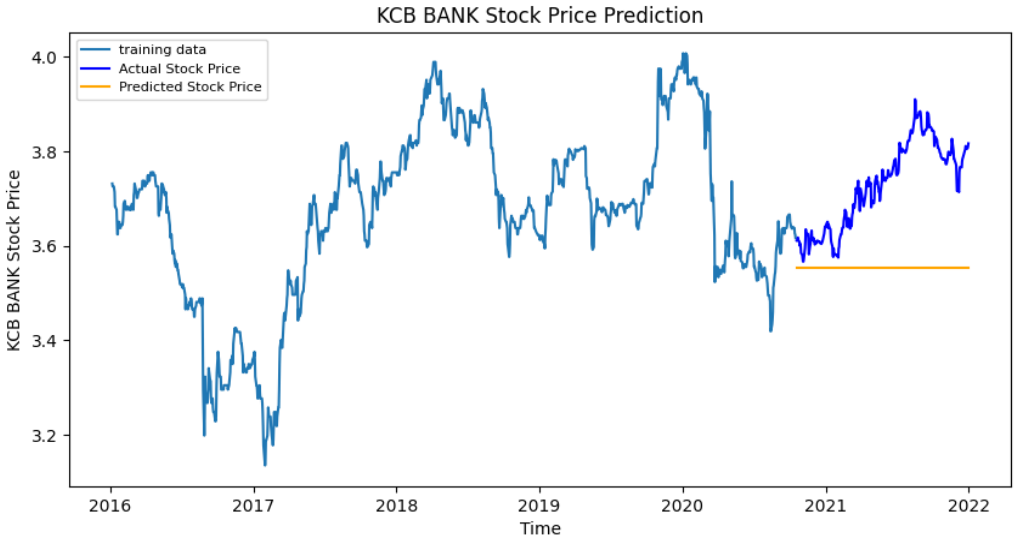


Figure 4.6: KCB Stock Prediction Using ARIMA Model

The accuracy metrics yielded these results;

MSE: 0.02101 MAE: 0.11901 RMSE: 0.1449 MAPE: 0.0314

From the RMSE of the dataset, it is possible to conclude that the performance of the model was generally satisfactory. Furthermore, according to a general rule of thumb, a RMSE value that falls within the range of 0.2 and 0.5 indicates that the model is able to relatively accurately forecast the data.

4.3.3 ARIMA Model for Co-op Bank

For Co-op Bank, the best model was of order (3,1,0). This implied that the best fitted model was an autoregressive model of order 3.

Results of the model's summary are indicated in Table 4.7 below.

Table 4.7: CO-OP Bank ARIMA Model Summary Results

	Coefficient	**S.E**	**z**	**P>|z|**	**C.I (0.025)**	**C.I (0.975)**
AR (1)	0.2251	0.015	15.249	0.000	0.196	0.254
AR (2)	0.0511	0.020	2.544	0.011	0.012	0.090
AR (3)	-0.1121	0.021	-5.329	0.000	-0.153	-0.071
σ^2	0.0003	4.2e-06	61.804	0.000	0.000	0.000

Mathematically, the model can be written as;

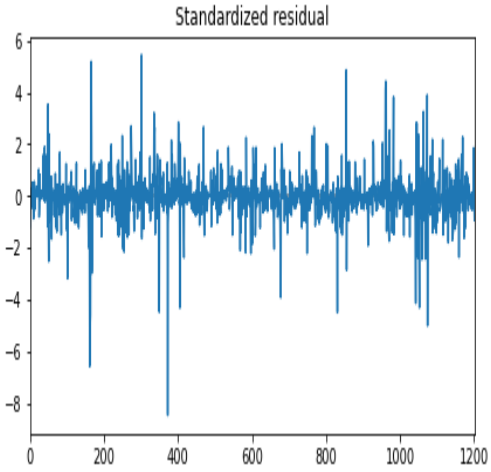
$$Y_t = \mu + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \varepsilon_t \quad (4.5)$$

where $\mu = 14.402315$ as obtained in Table 4.1, $\alpha_1 = 0.2251$ as obtained is Table 4.7, $\alpha_2 = 0.0511$ as obtained is Table 4.7, $\alpha_3 = -0.1121$ as obtained is Table 4.7.

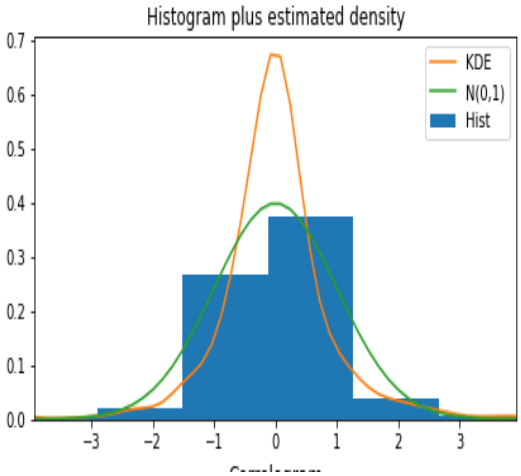
Substituting the values in equation 4.5 yields;

$$Y_t = 14.402315 + 0.2251 X_{t-1} + 0.0511 X_{t-2} - 0.1121 X_{t-3} + \varepsilon_t \quad (4.6)$$

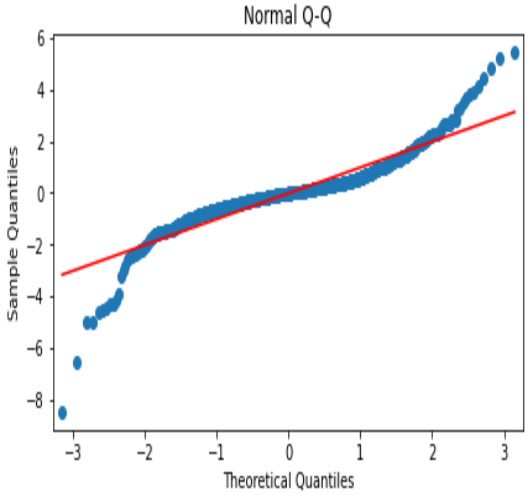
Moreover, the CO-OP Bank's ARIMA model diagnostics and Ljung Box test gave the results presented in Figure 4.7 and Table 4.8 respectively;



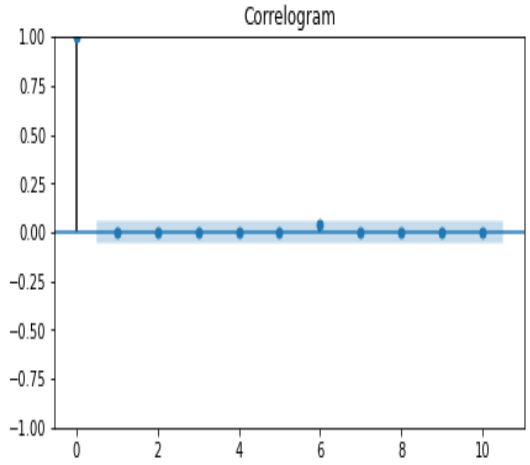
a. Standardized Residuals



b. Histogram plus Estimated Density



c. Normal Q-Q Plot



d. Correlogram Plot

Figure 4.7: CO-OP Bank's ARIMA Model Diagnostics

Table 4.8: Ljung Box test results for CO-OP Bank

Ljung-Box (L1) (Q):	0.00

Prob(Q):	0.99
Heteroskedasticity (H):	0.86
Prob(H) (two-sided):	0.13

From the results in Table 4.8 above, Prob(Q)=0.99 is greater than Q=0.00. Therefore, the null hypothesis was rejected implying that there exists an ARIMA model for forecasting CO-OP bank security prices.

Furthermore, from the results in Figure 4.7;

Figure 4.7a: The standardized residual errors appeared to have a uniform variance and fluctuate around a mean of zero.

Figure 4.7b: The density plot suggested a normal distribution with a mean of zero.

Figure 4.7c: The red line should be perfectly aligned with all of the dots. Any significant deviations indicated a skewed distribution.

Figure 4.7d: The Correlogram, often known as the ACF plot, demonstrated that the residual errors were not autocorrelated. Any autocorrelation implied a pattern in the residual errors that the model couldn't account for. As a result, an addition of X's (predictors) could be needed to the model.

Forecasting was then performed for the dataset and yielded the forecast in Figure 4.8 below;

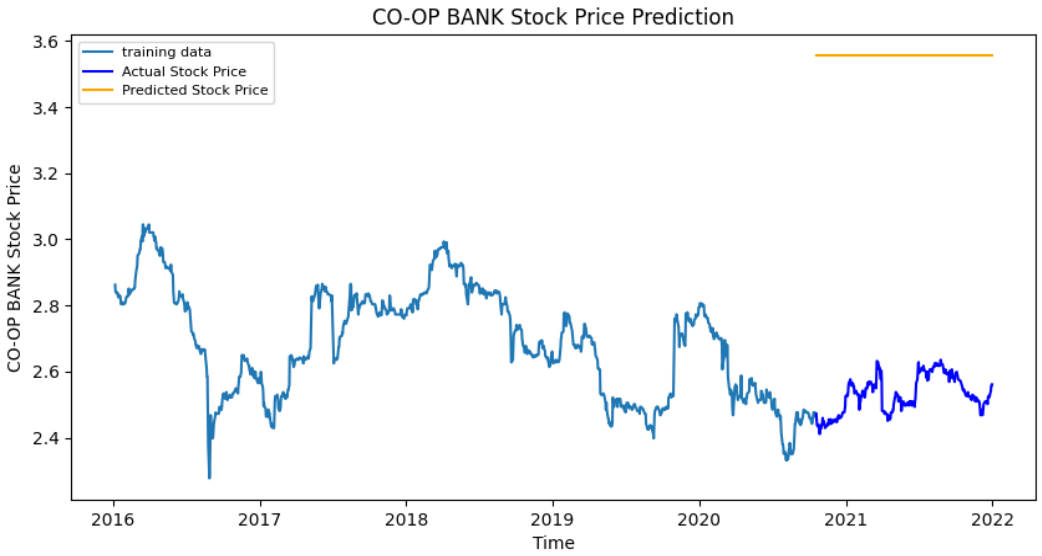


Figure 4.8: CO-OP Bank Stock Prediction using ARIMA Model

The accuracy metrics yielded these results;

MSE: 0.0075 MAE: 0.0721 RMSE: 0.0866 MAPE: 0.0281

From the RMSE of the dataset, it is possible to conclude that the performance of the model was generally satisfactory. Furthermore, according to a general rule of thumb, a RMSE value that falls within the range of 0.2 and 0.5 indicates that the model is able to relatively accurately forecast the data.

4.4 Artificial Neural Network (ANN) Model

The back-propagation method has seen a lot of use recently according to Mohammadi (2019). Numerous researchers choose to utilize it as one of the models for stock prediction due to the fact that it is so widely used and has a history of producing satisfactory results when applied to certain time series data. For the sake of clarity, the structure of the

model's Back-propagation has been laid out in Table 4.9 below. This is due to the fact that the best result was only achieved after an extensive period of trial and error.

Table 4.92: Back Propagation Training Parameters

Parameters	Values
Number of Epochs	1000
Training Method	Adam
Loss	0.0001659
Hidden Layer Activation Function	Relu
Hidden Layer Activation Function	Relu
Hidden Layer Activation Function	Relu
Hidden Layer Activation Function	tanh
Dropout percentage	0

Different values were tested and the best ones among them selected. After the model had been trained, the performance assessment approaches were used to evaluate the accuracy of the training and testing procedures separately. It is necessary to point out that the evaluation findings differed from one bank to another due to the fact that the model was evaluated with three different banks. Results for each bank are discussed below.

4.4.1 ANN Model for Equity Bank

Table 4.10 below shows Equity Bank's training and testing evaluation results for ANN model.

Table 4.10: Equity Bank's Training and Testing Evaluation Results for ANN Model

Training		**Testing**	
Parameter	**Value**	**Parameter**	**Value**
Number of Observations	1209	Number of Observations	303
MSE	0.04194	MSE	0.04153
RMSE	0.2048	RMSE	0.2038
MAE	0.1956	MAE	0.1946
MAPE	0.0487	MAPE	0.0477

It was evident that the model performed so well for prediction purposes after looking at Table 4.10. It produced a satisfactory outcome which enables the bank to make accurate projections on the future pricing of the asset. The coefficient of determination for the model was determined to be 0.9980309 which indicates that the model explained 99.80% of the variability in the data.

The model's predictability was tested in an experiment employing a different set of equity shares using the insights gained from training the model. The Equity Bank's shares were represented graphically as shown in Figure 4.9 below.

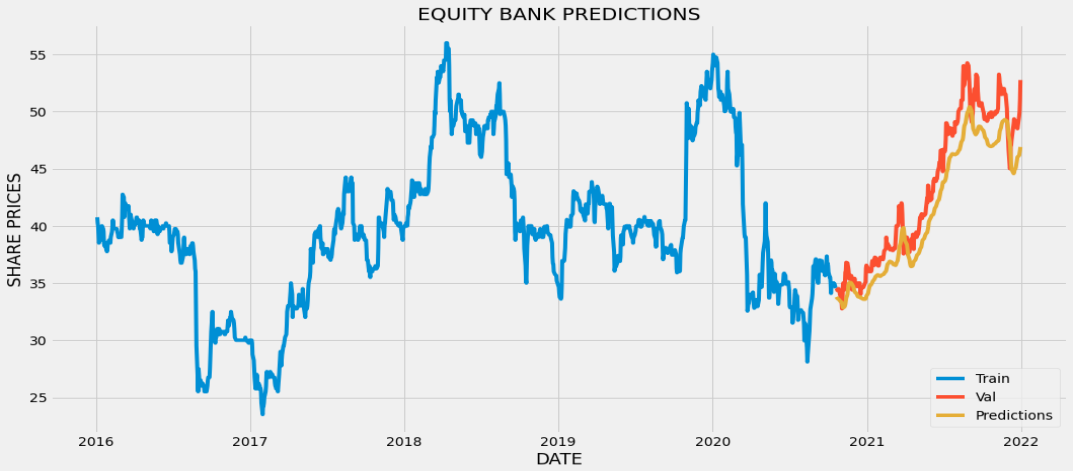


Figure 4.9: Equity Bank Stock Prediction using ANN Model

Table 4.11 shows findings that were presented to facilitate a comparison between the actual value and the expected value.

Table 4.11: Equity Bank's Prediction Results using ANN Model

Date	Actual	Predictions
2020-10-21	34.55	33.78
2020-10-22	34.50	33.72
2020-10-23	34.35	33.66
2020-10-26	34.50	33.59
2020-10-27	34.40	33.55
...	...	...
2021-12-24	48.50	46.01
2021-12-28	49.60	46.16
2021-12-29	50.00	46.37

2021-12-30	51.00	46.60
2021-12-31	52.75	46.90

The table reveals that the model's predictions were rather close to the actual values, as can be seen in Table 4.11.

To test the hypothesis regarding the existence of an ANN model, an observed Chi-Square value of 271.43 was computed. This value was compared to the standard p-value. The results of the comparison led to the conclusion that the null hypothesis should be rejected, which in turn led to the implication that an ANN Model existed for predicting stock market prices.

4.3.2 ANN Model for KCB Bank

A similar procedure was repeated for other banks and yielded these results.

Table 3: KCB's Training and Testing Evaluation Results for ANN Model

Training		**Testing**	
Parameter	**Value**	**Parameter**	**Value**
Number of Observations	1209	Number of Observations	303
MSE	0.01785	MSE	0.01758
RMSE	0.1336	RMSE	0.1326
MAE	0.1048	MAE	0.1038
MAPE	0.0256	MAPE	0.0246

Moreover, the coefficient of determination for the model was found to be 0.9985014 implying that there was 99.85% variability explained by the model. The predictions and actual values were depicted in Table 4.13 below.

Table 4.13: KCB's Prediction Results using ANN Model

Date	Actual	Predictions
2020-10-21	37.10	38.25
2020-10-22	37.25	38.11
2020-10-23	37.05	38.00
2020-10-26	36.90	37.90
2020-10-27	36.60	37.81
...	...	...
2021-12-24	45.20	45.47
2021-12-28	44.95	45.70
2021-12-29	45.00	45.85
2021-12-30	45.15	45.94
2021-12-31	45.45	46.00

For visualization, the results were presented as in Figure 4.10 below;

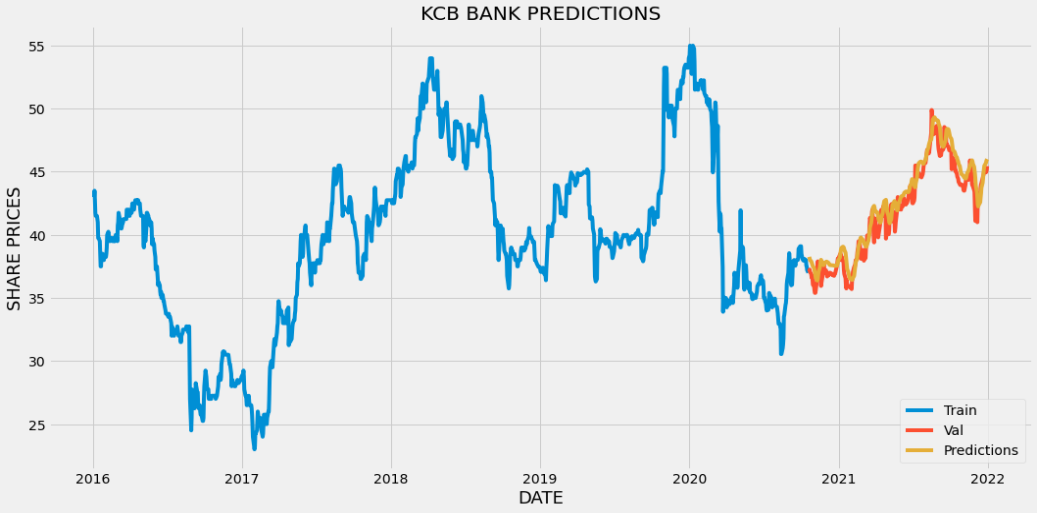


Figure 4.10: KCB Stock Prediction using ANN Model

In order to test the hypothesis regarding the existence of an ANN model, an observed Chi-Square value of 220.22 was computed. This value was compared to the standard p-value. The results of the comparison led to the conclusion that the null hypothesis should be rejected, which in turn led to the implication that an ANN Model existed for predicting stock market prices

4.3.3 ANN Model for Co-op Bank

Similarly, Table 4.14 below shows a summary of the evaluation metrics for both training and testing sections as well as the parameters for both sections.

Table 4.14: CO-OP Bank's Training and Testing Evaluation Results for ANN Model:

Training		**Testing**	
Parameter	**Value**	**Parameter**	**Value**
Number of Observations	1209	Number of Observations	303

MSE	0.00566	MSE	0.00551
RMSE	0.0752	RMSE	0.0742
MAE	0.0667	MAE	0.0657
MAPE	0.0158	MAPE	0.0148

The model observed R^2 of 0.9973062 implying that there was 99.73% variability explained by the model. Predictions was conducted and yielded results which were later presented in the Figure 4.11 and Table 4.15 below.

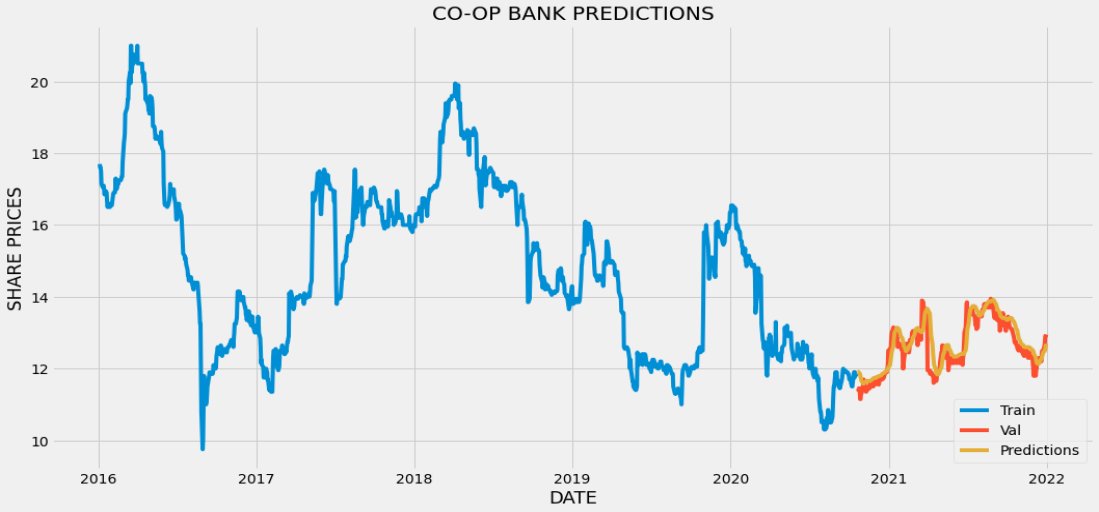


Figure 4.11: CO-OP Bank Stock Prediction using ANN Model

Table 4: CO-OP Bank's Prediction Results using ANN Model

Date	Actual	Predictions
2020-10-21	11.45	11.92
2020-10-22	11.45	11.92
2020-10-23	11.40	11.90

2020-10-26	11.45	11.87
2020-10-27	11.15	11.83
...	...	...
2021-12-24	12.45	12.45
2021-12-28	12.70	12.50
2021-12-29	12.90	12.55
2021-12-30	12.90	12.62
2021-12-31	12.95	12.69

An observed Chi-Square value of 230.58 was computed for testing the hypothesis for existence of ANN model this was compared to the standard p-value. From the comparison, it was concluded to reject the null hypothesis and imply that there existed an ANN Model for forecasting stock market prices.

4.5 Support Vector Machine (SVM) Model

SVM was also evaluated using the provided datasets, the accuracy was determined, and the results were shown at the end. SVM used an approach comparable to ANN for data processing. Here, data for forecasting was prepared using the similar approaches to ANN. SVM, on the other hand, provided a better solution for picking parameter values when it came to increasing training performance. This capability to determine values based on the information that was used for forecasting was made available via the best-estimator object function. As a result, the optimal values were determined using that function. The RBF kernel outperformed the other three SVM kernels when they were put to the test. Table 4.16 give the results of the training parameters for SVM model.

Table 4.16: SVM Model Training Parameters

Parameters	Values
Kernel	RBF
Gamma	0.005
Epsilon	0.01
C	100

The model was tested and trained based on these values, and the results for each bank discussed.

4.5.1 SVM Model for Equity Bank

A value of 0.5454 for the R^2 coefficient of determination was observed for Equity Bank. This is a clear indication that there was 54.54% variability explained by the model. Below are training and testing parameters with their accuracy metrics in Table 4.17.

Table 4.17: Equity Bank's Training and Testing Evaluation Results for SVM Model

Training		Testing	
Parameter	**Value**	**Parameter**	**Value**
Number of Observations	1209	Number of Observations	303
MSE	0.03084	MSE	0.03049
RMSE	0.1756	RMSE	0.1746
MAE	0.1855	MAE	0.1845
MAPE	0.0332	MAPE	0.0322

Forecasting was then conducted for the dataset and yielded results depicted in Figure 4.12.

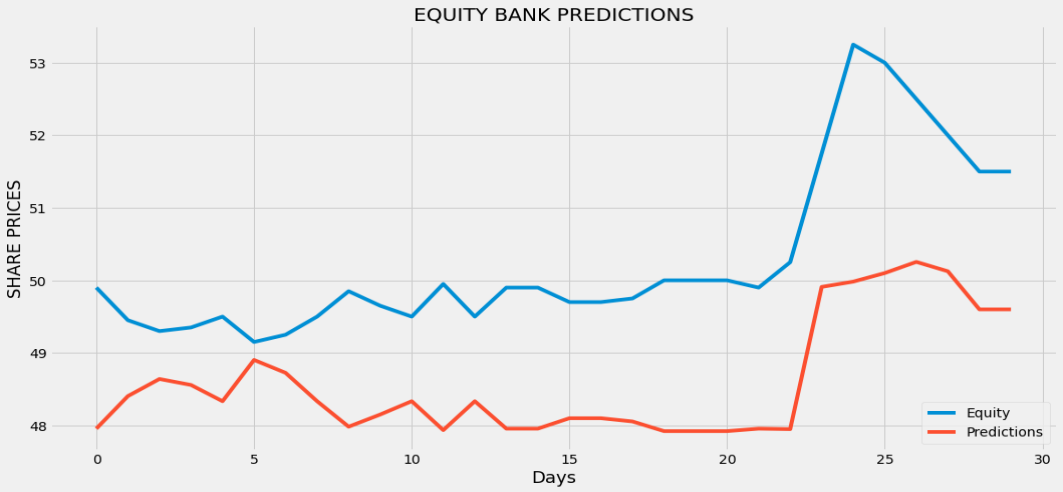


Figure 4.12: Equity Bank Stock Prediction using SVM Model

From the results in Figure 4.12 above, there is a clear indication of no much variation between the actual values and the predictions.

An observed Chi-Square value of 228.57 was computed for testing the hypothesis for existence of SVM model and this was compared to the standard p-value. From the comparison, it was concluded that we reject the null hypothesis and this implies that there existed a SVM model for forecasting stock market prices.

4.5.2 SVM Model for KCB Bank

The R^2 coefficient of determination for KCB was 0.6110. In comparison to Equity bank shares, the bank's shares had a decent performance. As a result of the realization that the bank's shares had a good R^2 value, future prices were obtained. The training and testing parameters are presented in Table 4.18 below.

Table 4.18: KCB's Training and Testing Evaluation Results for SVM Model

Training		Testing	
Parameter	**Value**	**Parameter**	**Value**
Number of Observations	1209	Number of Observations	303
MSE	0.00974	MSE	0.00955
RMSE	0.0987	RMSE	0.0977
MAE	0.0955	MAE	0.0945
MAPE	0.0146	MAPE	0.0136

The results obtained after forecasting were as in Figure 4.13 below.

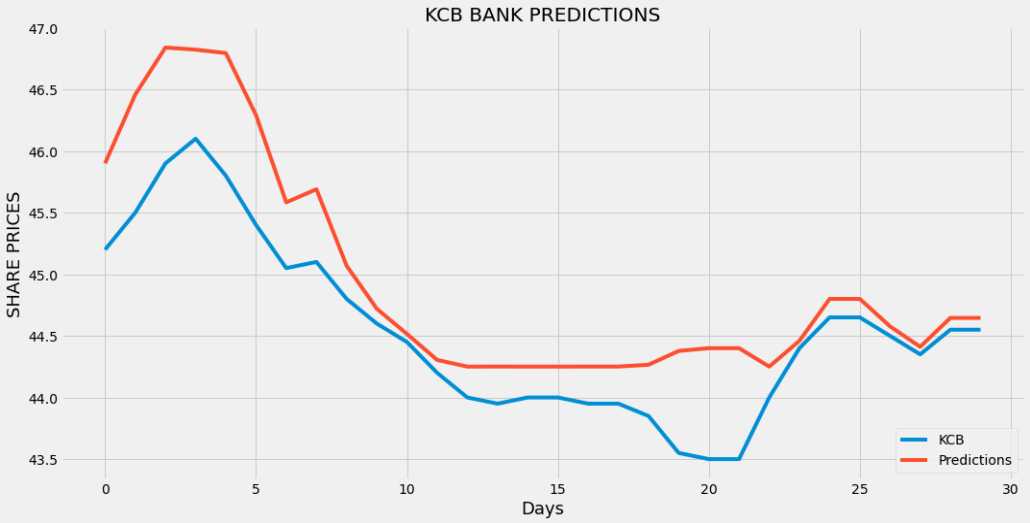


Figure 4.13: KCB Stock Prediction using SVM Model

A small variation between the actual and the predictions was realized. An observed Chi-Square value of 208.55 was obtained for testing the hypothesis for existence of SVM

model and this was compared to the standard p-value. From the comparison, it was concluded that we reject the null hypothesis and this implies that there existed a SVM model for forecasting stock market prices.

4.5.3 SVM Model for CO-OP Bank

The R^2 coefficient of determination for Co-op Bank was 0.6901. This was a higher value than the other banks. The implication of the results was that the model could be used for prediction since the performance of the model using the R^2 coefficient of determination was good as the values were positive and more than 0.5.

The Table 4.19 below indicates the accuracy metrics that were established.

Table 4.19: CO-OP Bank's Training and Testing Evaluation Results for SVM Model

Training		**Testing**	
Parameter	**Value**	**Parameter**	**Value**
Number of Observations	1209	Number of Observations	303
MSE	0.003869	MSE	0.00375
RMSE	0.0622	RMSE	0.0612
MAE	0.0532	MAE	0.0522
MAPE	0.00986	MAPE	0.00976

Forecasting was then performed for the bank and gave these results as shown in Figure 4.14.

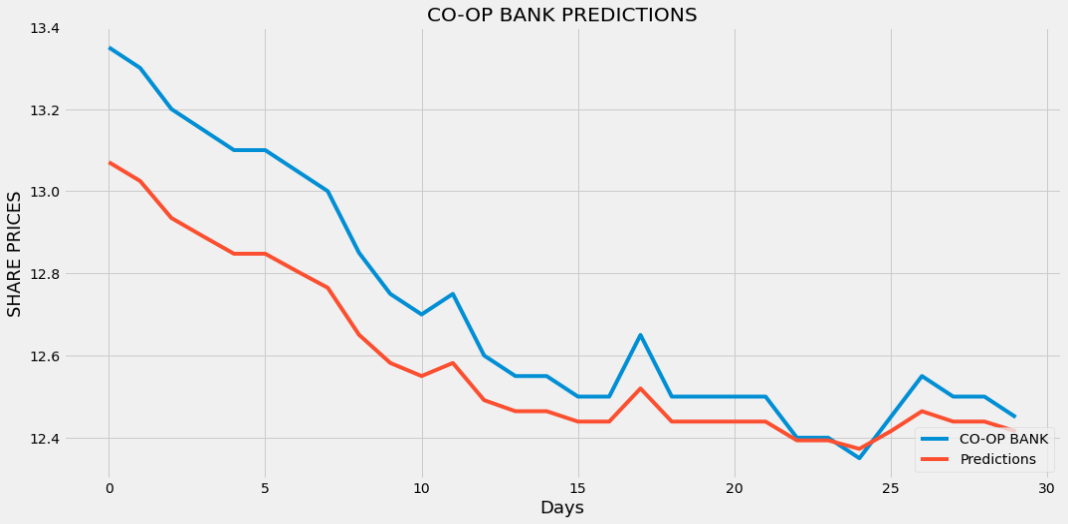


Figure 4.14: CO-OP Bank Stock Prediction Using SVM Model

The trend of the predictions and the actual results was evident as the predictions followed the same trend as the actual values. There was no doubt therefore on the performance of the model as it yielded better results in prediction.

An observed Chi-Square value of 213.13 was evident for testing the hypothesis for existence of SVM model and this was compared to the standard p-value. From the comparison, it was concluded that we reject the null hypothesis and this implies that there existed a SVM model for forecasting stock market prices.

4.6 Choosing the Most Appropriate Model

In order to get a better understanding of which model was superior, the outcomes of having models tested for three different banks were provided in detail. It was clear that the SVM performed much better than the other models when comparing the models using the error as the criteria. A graphical representation of how the errors were spread out may be seen in Figure 4.15.

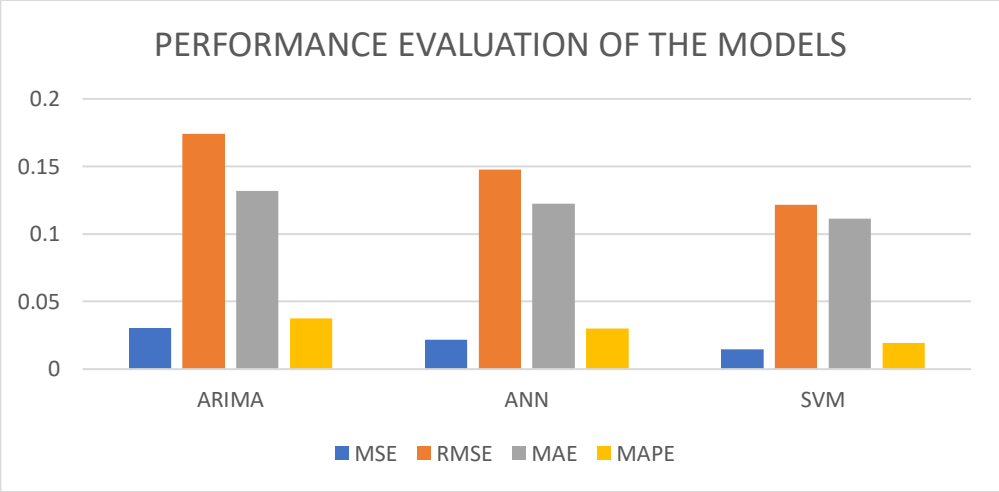


Figure 4.15: Performance Evaluation of the Models

Figure 4.15 demonstrates that SVM outperformed the other models by a significant margin. According to the results of error assessment methodologies known as MSE, RMSE, MAE, and MAPE, the support vector machine (SVM) had fewer errors than the alternatives.

Using the confusion matrix criterion, the Tables 4.20 and 4.21 below represent results of ANN and SVM models respectively.

Table 4.20: Confusion Matrix for ANN Model

	N	**PROJECTED**	
		POSITIVE	*NEGATIVE*
EXACT	*POSITIVE*	658	471
	NEGATIVE	140	243

Table 5: Confusion Matrix for SVM Model

	N	**PROJECTED**	
		POSITIVE	*NEGATIVE*
EXACT	*POSITIVE*	564	235
	NEGATIVE	344	369

For ANN,

$$Recall = \frac{658}{658 + 471} = 0.5828$$

$$Precision = \frac{658}{658 + 140} = 0.8246$$

$$Accuracy = \frac{658 + 243}{1512} = 0.5959$$

$$F1\ Score = \frac{2(658)}{2(658) + 140 + 471} = 0.6829$$

$$MCC = \frac{(658 \times 243) - (140 \times 471)}{\sqrt{(658 + 140) \times (658 + 471) \times (140 + 243) \times (471 + 243)}} = 0.1893$$

For SVM,

$$Recall = \frac{564}{564 + 235} = 0.7059$$

$$Precision = \frac{564}{564 + 344} = 0.6211$$

$$Accuracy = \frac{564 + 369}{1512} = 0.6171$$

$$F1\ Score = \frac{2(564)}{2(564) + 344 + 235} = 0.6608$$

$$MCC = \frac{(564 \times 369) - (344 \times 235)}{\sqrt{(564 + 344) \times (564 + 235) \times (344 + 369) \times (235 + 369)}} = 0.2277$$

From Tables 4.20 and 4.21 above, Accuracy, Recall, Precision, F1 Score and Matthew's Correlation Coefficient were calculated. A well optimized and well performing model will have both metrics close to one.

As evident, SVM outperformed ANN in most of the metrics. This implied that SVM had a higher true positive rate which implies that the model predicted a more accurate value for the stocks as well as being better for predicting stock prices.

4.7 Summary

This chapter laid forth the descriptive statistics of the data for the three banks chosen for the study. The chapter also looked into each specific objective with detailed results obtained from the study. Firstly, the ARIMA model's results were laid forth. Secondly, the ANN model's results were then laid forth and thirdly, the SVM model's results. Finally, results on the most appropriate model were laid forth. SVM performed better than the other two models under accuracy error metrics as well as confusion matrix criterion.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

As economies have expanded in line with the rapid development of new technologies, the practice of forecasting has become essential in virtually every industry. There are many examples, including customer-based marketing, the demand for automobiles, mobile devices, currency exchange rates, and stock prices. Forecasting the price of stocks is becoming an increasingly tough assignment for humans. Forecasting the stock market is becoming increasingly crucial in this day and age because investors need to know whether or not it is a good time to buy or sell stocks. Humans have been able to more accurately anticipate stock values with the use of machine learning and artificial intelligence in general.

The Equity Bank dataset worked well with an ARIMA model with parameters 0, 1, 1. This suggested that a Moving Average process of order 1 was the most appropriate model for the financial institution. The results obtained by KCB were comparable to those obtained by Equity Bank using an ARIMA (0,1,1). In contrast to the results of the ARIMA model obtained by other institutions, those obtained by the Co-op Bank indicated an auto-regressive model of order 3. The ARIMA model performed admirably for both institutions, yielding a root mean squared error of 0.1743, which was well within acceptable limits.

The coefficient of determination for ANN's Equity Bank was 0.9980, which indicates that the model explained 99.80 percent of the variability in the data. In addition, the model

had a mean absolute percentage error of 0.0477%, which indicated that it had an accuracy of nearly 96% while evaluating the dataset. In the case of KCB, the ANN model had an error of 0.0246 in terms of the mean absolute percentage, which indicates that the model had an accuracy of around 98 percent when testing the dataset. In a similar vein, approximately 99 percent of the evidence was there for CO-OP Bank.

For SVM, Equity Bank had approximately 97 percent testing accuracy for MAPE as compared to KCB which had about 99% MAPE accuracy for testing the bank's dataset in the testing part. Moreover, CO-OP Bank has nearly 99% MAPE accuracy for testing the dataset. This implied that all banks' predictions were close to the actual values.

This study made use of three different models in order to make stock value projections for three different banks operating in the market. These models delivered an outstanding performance. In addition, SVM performed significantly better than the other two high-error models (ANN and ARIMA). The primary objective of the study was accomplished, and the SVM Model was suggested for use in forecasting because it had the lowest error rate together with the highest accuracy and recall, which made it an efficient tool for stock prediction. This will let stakeholders make decisions about the best time to buy or sell their shares at the appropriate time.

In addition, the performance of the models was quite good, which can be attributed to the vast amount of data that was gathered during the training phase. However, there are still other forces at play, and stock prices have many characteristics; hence, it is impossible to forecast based just on time series.

5.2 Recommendations

Overcoming the challenges would greatly increase the accuracy and reliability of the study's conclusions because the study's limitations are so severe. As a consequence of this, the first piece of work that will be completed in the future may consist of an analysis of historical data as well as the daily financial and political news utilizing Natural Language Processing (NLP) models.

The second strategy for improving accuracy and dependability is to investigate the components mentioned above and determine whether or not the data from those aspects that directly affect currency prices can be integrated with the data that has been collected in the past. How much more accurate would the algorithms be if specific projections of the future of those factors were paired with data from the past while they were being trained and tested? The premise of the study suggests that the degree to which the data collection and extraction processes are accurate and timely will determine the results of the investigation.

In addition, the dataset should be expanded to include additional datasets from a variety of markets in order to determine whether or not SVM is more accurate than ANN in general and to compare with other machine learning algorithms.

REFERENCES

Abdul Hamid, N. I. (2018). Random walk hypothesis: An application in the emerging stock markets.

Abdullah, L. (2012). ARIMA model for gold bullion coin selling prices forecasting. *International Journal of Advances in Applied Sciences*, 1(4):153–158.

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv*:1803.08375.

Almasarweh, M. and Alwadi, S. (2018). ARIMA model in predicting banking stock market data. *Modern Applied Science*, 12(11):4.

Ariyo, A. A., Adewumi, A. O., and Ayo, C. K. (2014). Stock price prediction using the ARIMA model. In 2014 UKSim-AMSS *16th International Conference on Computer Modelling and Simulation*, pages 106–112. IEEE.

As' ad, M. (2012). Finding the best ARIMA model to forecast daily peak electricity demand.

Awad, M. and Khanna, R. (2015). Support vector regression. *In Efficient learning machines*, pages 67–80. Springer.

Banerjee, D. (2014). Forecasting of Indian stock market using time-series ARIMA model. In 2014 *2nd international conference on business and information management (ICBIM)*, pages 131–135. IEEE.

Bartram, S. M. and Grinblatt, M. (2018). Agnostic fundamental analysis works. *Journal of Financial Economics,* 128(1):125–147.

Beck, T., Cull, R., Fuchs, M. J., Getenga, J., Gatere, P. K., Randa, J., and Trandafir, M. (2010). Banking sector stability, efficiency, and outreach in Kenya. *World Bank Policy Research Working Paper*, (5442).

Bilsborrow, R. E. (2016). Concepts, definitions and data collection approaches. In *International handbook of migration and population distribution*, pages 109– 156. Springer.

Bontempi, G., Taieb, S. B., and Le Borgne, Y.-A. (2012). Machine learning strategies for time series forecasting. In *European business intelligence summer school*, pages 62–77. Springer.

Box, G. and Jenkins, G. (2013). Box and Jenkins: time series analysis, forecasting and control. In A *Very British Affair*, pages 161–215. Springer.

Campbell, S., Greenwood, M., Prior, S., Shearer, T., Walkem, K., Young, S., Bywaters, D., and Walker, K. (2020). Purposive sampling: complex or simple? Research case examples. *Journal of research in Nursing*, 25(8):652–661.

Chai, T. and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific Model Development Discussions*, 7(1):1525–1534.

Chaudhuri, T. and Pandit, A. (2021). Time-series Models-Forecasting Performance in the Stock Market. *Journal of Contemporary Issues in Business and Government Vol*, 27(2).

Chavan, P. S. and Patil, S. T. (2013). Parameters for stock market prediction. *International Journal of Computer Technology and Applications,* 4(2):337.

Chong, E., Han, C., and Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, 83:187–205.

Cilimkovic, M. (2015). Neural networks and back propagation algorithm. *Institute of Technology Blanchardstown, Blanchardstown Road North Dublin*, 15(1).

Colvin, C. L. (2016). The past, present and future of banking history. In *The Routledge companion to business history*, pages 103–120. Routledge.

Crotty, M. (2020). *The foundations of social research: Meaning and perspective in the research process.* Routledge.

Dai, Y. and Zhao, P. (2020). A hybrid load forecasting model based on support vector machine with intelligent methods for feature selection and parameter optimization. *Applied Energy*, 279:115332.

Das, S. P. and Padhy, S. (2012). Support vector machines for prediction of futures prices in Indian stock market. *International Journal of Computer Applications*, 41(3).

Das, S. P. and Padhy, S. (2018). A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting. *International Journal of Machine Learning and Cybernetics*, 9(1):97–111.

de Souza, R. R., Toebe, M., Marchioro, V. S., Cargnelutti Filho, A., Lúcio, A. D., Benin, G., Mello, A. C., de Lima Tartaglia, F., and Manfio, G. L. (2022). Soybean yield variability per plant in subtropical climate: sample size definition and prediction models for precision statistics. *European Journal of Agronomy*, 136:126489.

Deepika, M., Nambiar, G., and Rajkumar, M. (2012). Forecasting price and analysing factors influencing the price of gold using ARIMA model and multiple regression analysis. *International Journal of Research in Management, Economics and Commerce,* 2(11):548–563.

Ding, X., Liu, J., Yang, F., and Cao, J. (2021). Random radial basis function kernel-based support vector machine. *Journal of the Franklin Institute*, 358(18):10121– 10140.

Edwards, R. D., Magee, J., and Bassetti, W. C. (2018). *Technical analysis of stock trends*. CRC press.

Fuller, W. A. (2009). *Introduction to statistical time series*. John Wiley & Sons.

Gao, G., Lo, K., and Fan, F. (2017). Comparison of ARIMA and ANN models used in electricity price forecasting for power market. *Energy and Power Engineering*, 9(4B):120–126.

Garg, N., Sharma, M., Parmar, K., Soni, K., Singh, R., and Maji, S. (2016). Comparison of ARIMA and ANN approaches in time-series predictions of traffic noise. *Noise Control Engineering Journal*, 64(4):522–531.

Horak, J., Vrbka, J., and Suler, P. (2020). Support vector machine methods and artificial neural networks used for the development of bankruptcy prediction models and their comparison. *Journal of Risk and Financial Management*, 13(3):60.

Jay, P., Kalariya, V., Parmar, P., Tanwar, S., Kumar, N., and Alazab, M. (2020). Stochastic neural networks for cryptocurrency price prediction. *IEEE Access*, 8:82804–82818.

Kenyan-Magazine (2022). Top 15 Best Banks in Kenya 2022.

Khedmatia, M., Seifi, F., and Azizi, M. J. (2020). Time series forecasting of Bitcoin price based on ARIMA and machine learning approaches.

Kihoro, J. and Okango, E. (2014). Stock market price prediction using artificial neural network: an application to the Kenyan equity bank share prices. *Journal of Agriculture, Science and Technology*, 16(1):161–172.

Kukreja, H., Bharath, N., Siddesh, C., and Kuldeep, S. (2016). An introduction to artificial neural network. *Int J Adv Res Innov Ideas Educ*, 1:27–30.

Lagat, A. K., Waititu, A. G., and Wanjoya, A. K. (2018). Support vector regression and artificial neural network approaches: Case of economic growth in East Africa community. *American Journal of Theoretical and Applied Statistics*, 7(2):67– 79.

Lai, P. (2018). Research methodology for novelty technology. *JISTEM-Journal of Information Systems and Technology Management*, 15.

Lin, Q. (2018). Technical analysis and stock return predictability: An aligned approach. *Journal of financial markets*, 38:103–123.

Malkiel, B. G. (1973). A random walk down Wall Street. W. W.

Michail, N. (2021). Money, Credit, and Crises: *Understanding the Modern Banking System.* Springer.

Mitchell, T. M. and Mitchell, T. M. (1997). *Machine learning*, volume 1. McGraw-hill New York.

Mohammadi, W. (2019). Currency Exchange Rate Forecasting Using Machine Learning Techniques. *Graduate School of Applied Sciences. Near East University*.

Nandakumar, K., Ratha, N., Pankanti, S., and Halevi, S. (2019). Towards deep neural network training on encrypted data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Neal, R. M. (1996). Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer.

NSE (2020). NSE Integrated Report and Financial Statements.

Nti, K. O., Adekoya, A., and Weyori, B. (2019). Random forest based feature selection of macroeconomic variables for stock market prediction. *American Journal of Applied Sciences*, 16(7):200–212.

Nwankpa, C., Ijomah, W., Gachagan, A., and Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv*:1811.03378.

Pahwa, K. and Agarwal, N. (2019). Stock market analysis using supervised machine learning. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 197–200. IEEE.

Pratiwi, H., Windarto, A. P., Susliansyah, S., Aria, R. R., Susilowati, S., Rahayu, L. K., Fitriani, Y., Merdekawati, A., and Rahadjeng, I. R. (2020). Sigmoid activation function in selecting the best model of artificial neural networks. In *Journal of Physics:* Conference Series, volume 1471, page 012010. IOP Publishing.

Rajput, G. and Kaulwar, B. H. (2019). A Comparative Study of Artificial Neural Networks and Support Vector Machines for predicting stock prices in National Stock Exchange of India. In *2019 International Conference on Data Science and Communication (IconDSC)*, pages 1–7. IEEE.

Ramchoun, H., Ghanou, Y., Ettaouil, M., and Janati Idrissi, M. A. (2016). Multilayer perceptron: Architecture optimization and training.

Rasel, R. I., Sultana, N., and Hasan, N. (2016). Financial instability analysis using ANN and feature selection technique: application to stock market price prediction. In *2016 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pages 1–4. IEEE.

Ros, T., Enriquez-Geppert, S., Zotev, V., Young, K. D., Wood, G., Whitfield-Gabrieli, S., Wan, F., Vuilleumier, P., Vialatte, F., Van De Ville, D., et al. (2020). Consensus on the reporting and experimental design of clinical and cognitive behavioural neurofeedback studies (CRED-nf checklist).

Saunders, M., Lewis, P., and Thornhill, A. (2009). *Research methods for business students*. Pearson education.

Schumaker, R. P. and Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):1–19.

Shaheen, M. and Arshad, M. (2020). Use of Machine Learning in Stock Market Prediction. *European Journal of Technology*, 4(1):60–73.

Sharma, G. D., Erkut, B., Jain, M., Kaya, T., Mahendru, M., Srivastava, M., Uppal, R. S., and Singh, S. (2020). Sailing through the COVID-19 Crisis by Using AI for Financial Market Predictions. *Mathematical Problems in Engineering*, 2020.

Sharma, S., Sharma, S., and Athaiya, A. (2017). Activation functions in neural networks. *towards data science*, 6(12):310–316.

Sivasamy, R., Forcheh, N., Kgosi, P., and Mabotho, B. (2017). Best Times to Trade Stocks using ARIMA and ANN methods.

Sutanapong, C. and Louangrath, P. (2015). Descriptive and inferential statistics. *International Journal of Research & Methodology in Social Science*, 1(1):22–35.

Tang, A., Hallouch, O., Chernyak, V., Kamaya, A., and Sirlin, C. B. (2018). Epidemiology of hepatocellular carcinoma: target population for surveillance and diagnosis. *Abdominal Radiology*, 43(1):13–25.

Taud, H. and Mas, J. (2018). Multilayer perceptron (MLP). In *Geomatic approaches for modeling land change scenarios*, pages 451–455. Springer.

Theofanidis, D. and Fountouki, A. (2018). Limitations and delimitations in the research process. *Perioperative Nursing-Quarterly scientific, online official journal of GORNA*, 7(3 September-December 2018):155–163.

Vapnik, V., Golowich, S. E., Smola, A., et al. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in neural information processing systems*, pages 281–287.

Vijh, M., Chandola, D., Tikkiwal, V. A., and Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia Computer Science*, 167:599–606.

Visa, S., Ramsay, B., Ralescu, A. L., and Van Der Knaap, E. (2011). Confusion matrix-based feature selection. MAICS, 710:120–127.

Wanjawa, B. W. and Muchemi, L. (2014). ANN model to predict stock prices at stock exchange markets. *arXiv preprint arXiv*:1502.06434.

Weigend, A. S. (2018). Time series prediction: *forecasting the future and understanding the past.* Routledge.

Westlake, B., Bouchard, M., and Frank, R. (2017). Assessing the validity of automated webcrawlers as data collection tools to investigate online child sexual exploitation. *Sexual Abuse*, 29(7):685–708.

Zhang, Z. (2018). Artificial neural network. In *Multivariate time series analysis in climate and environmental research*, pages 1–35. Springer.

Zoppis, I., Mauri, G., and Dondi, R. (2019). Kernel methods: Support vector machines

APPENDICES

APPENDIX I: MUT AUTHORIZATION LETTER

MURANG'A UNIVERSITY OF TECHNOLOGY
DIRECTORATE OF POSTGRADUATE STUDIES

P.O. BOX 75 - 10200, MURANG'A **Email: bps@mut.ac.ke**

Ref: MUT/ARP/PGS/20/2020/VOL. **Date:** 3rd February 2022

Dear Marwa Hassan (AS400/5083/2019),

RE: APPROVAL OF RESEARCH PROPOSAL AND SUPERVISORS

I am pleased to inform you that the Directorate of Postgraduate Studies on 17th January 2022 considered and approved your Masters research proposal entitled ***"Forecasting of Banking Sector Security Prices in Kenya using Machine Learning Techniques"*** and appointed the following as supervisors:

1. **Dr. Ayub Anapapa -Murang'a University of Technology**
2. **Dr. John Mutuguta - Murang'a University of Technology**

You may now proceed with your data collection subject to obtaining research permit from NACOSTI, if required. You should also begin consulting your supervisors and submit through them quarterly progress reports to the Director Postgraduate Studies through your CoD and School Dean. Progress Reports can be accessed in the University Website.

It is the policy and regulations of the University that you observe deadlines. The guidelines on Postgraduate supervision can be accessed in the post graduate Handbook.
Your responsibilities as a student will include, among others;

i. Maintain regular consultation with your supervisor(s), at least once a month
ii. Submit quarterly reports on time, through your supervisors, CoD, Dean and to the Director of Postgraduate Studies;
iii. Ensure quality work all through;
iv. Present your research findings at 2- 3 seminars/conferences prior to thesis examination.
v. Publish one article from your research findings in a refereed journal prior to thesis examination

For any further clarification, please contact the undersigned.

Yours Sincerely,

DIRECTORATE OF POSTGRADUATE STUDIES
MURANG'A UNIVERSITY OF TECHNOLOGY
03 FEB 2022
P. O. Box 10200, MURANG'A
Tel: 0771463515, Email: info@mut.ac.ke

PROF. GEOFFREY MUCHIRI, PhD
DIRECTOR, POSTGRADUATE STUDIES

Cc Registrar (ASA)
Dean (SPAHS)

MUT IS ISO 9001:2015 CERTIFIED

APPENDIX II: PUBLICATION

The following paper has been published from this thesis.

Chacha, M. H., Anapapa, A., & Mutuguta, J. (2022). Forecasting of Banking Sector Securities Prices in Kenya Using Machine Learning Technique. Asian Journal of Probability and Statistics, 18(1), 19-30. https://doi.org/10.9734/ajpas/2022/v18i13043.