Application of Machine Learning Models in Forecasting Stock Market Volatility: Case Study of the Nairobi Securities Exchange

By

Gladwel Gathoni Wanjau

095401

Chapter 1: Introduction

1.1. Background to the study

1.1.1. The Investment Universe of the Stock Market

The investment universe in the stock market is divided into two major categories namely; traditional and Alternative products. Traditional investments involves trading of products such as stocks, bonds, mutual funds and cash. The remaining investments are categorized as alternative products which include Derivatives, Real Estate Investment Schemes, Hedge Funds, Commodities, Structured products, Managed Funds, Private Equity/ Venture Capital, among others. Alternative investments are basically an alternative to the traditional stock market products and they offer potential higher returns, exhibit lower volatility and can be used for capital preservation.

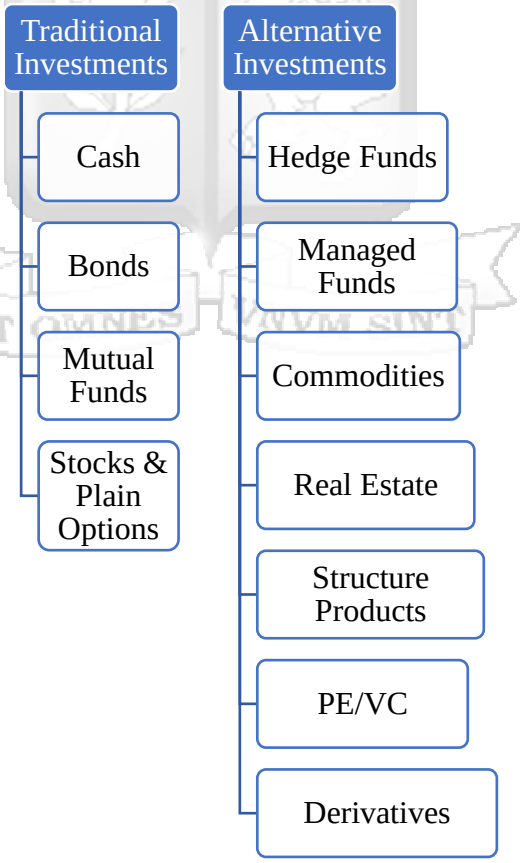

Figure 1: Investment Universe in the Stock Market

1.1.2. Volatility of the Stock market

Volatility is a measure of dispersion of stock market returns and is one of the key indices used to by investors in making investment decisions. Poon and Granger (2003) define volatility as a measure of the unpredictable changes in returns thus it follows a stochastic process and is a random variable. The performance of the stock market is affected by a myriad of factors such as macroeconomic performance, listed company's performance, political unrest, current news or events, calamities, pandemics (for example Covid-19 pandemic) among others thus making the market to be very volatile. This therefore implies that investing in this market comes with a high risk of movement and occurrence of such factors. The main concern around volatility is the lack of awareness about it especially when investors are making investment decisions. Further to this the unpredictability of the market makes investment decisions hard to make. There is therefore need to investigate the distribution of returns, variability of returns, ascertain whether periodic returns are normally distributed as well as whether the volatility is constant over time. Knowledge on these factors would lead to better informed decisions by investors. Previous works in this area has proven that this area is affected by high volatility hence the returns are unpredictable.

Over the last couple of years, the topic of volatility in the stock market has gained interest from many researchers, regulators, stock market analysts and economists as evidenced by the numerous studies undertaken in this area in different markets and countries. The main task has remained the estimation and forecasting of the volatility parameter (Brooks, 2004).

In modelling volatility, the main task lies in identifying past trends and utilization of the same in predicting future trends. According to Knight and Satchell (2017), the subject of volatility is very crucial for the financial markets since these markets are characterized by many periods of high volatility than of low volatility. Additionally, the stock returns exhibit non-normality distribution as well as excess kurtosis. Modelling volatility therefore comes in handy in risk management and evaluation of the vulnerability of the financial markets.

1.1.3. Nairobi Securities Exchange

Since its inception in 1954, The Nairobi Securities Exchange formerly known as the Nairobi Stock Exchange has been the principal Exchange in the Kenyan Capital Markets Industry tasked with the main role of connecting capital and opportunities in a bid to enhance stakeholder value. Further to this, it plays a vital role in encouraging saving and investment among the Kenyan population as well as facilitating the acquisition of the much-needed capital

for both local and international companies. NSE's products have a combined market capitalization in excess of USD 38 billion these include Equities (65+ listed companies), Bonds (76 listed securities), ETFs, REITs and Derivatives. In addition, NSE offers a platform for the trading and settlement of securities of unquoted companies called the Unquoted Securities Platform (USP). Growth and development of medium-sized enterprises are supported through an incubation and acceleration program - "Ibuka". The NSE listed its shares on the Main Investment Market Segment (MIMS) via an IPO which sought to raise Kshs.627 million. 17,859 investors applied for 504,189,700 new shares worth Kshs. 4.789 billion; a subscription rate of 763.9%, garnering an oversubscription of 663.92%. The NSE is second African Exchange after the Johannesburg Stock Exchange to be listed.

1.1.4. Machine Learning

Samuel (1959) defines machine learning as a field of study that gives computers the ability to learn without being explicitly programmed. It is the process of building computer systems that automatically improve a learning process through experience and implementation. These techniques should mimic the human reasoning in providing insights in the decision-making process. According to Ayodele, T. O. (2010) machine learning has two main objectives which are; Training the model using past data or example data and using the learned model to make inferences and predictions using the test data. Machine learning is classified into three major categories namely; Supervised learning, Unsupervised learning and Reinforcement learning according to Ayodele, T. O. (2010). Supervised Machine learning is conducted when the data provided for modelling is well labelled and the expected output is known. Some of the algorithms that fall under this category include; Decision trees, K-Nearest Neighbour, Naïve Bayes Estimator, Regressions such as Linear, Lasso, Ridge and Elastic Net, Support Vector Machines and Neural Networks. The main aim of supervised learning is the prediction of future values. On the other hand Unsupervised learning is defined as learning that is conducted on data that is unlabelled and whose main aim is to discover unknown information but useful information/classes of the data. Some of the machine learning algorithms that fall under this category include; Clustering which include K-mean, hierarchical and Density-Based Spatial Clustering, Principal component Analysis, t-Distributed Neighbour Embedding (t-SNE), Association Analysis and Anomaly detection. Reinforcement learning is learning whereby the training information is provided by an external trainer in the form of a scalar reinforcement signal that constitutes a measure of how well the system operates. In this case, the learner is not guided on the actions to take but is expected to yield great results but testing each action in

turn. The algorithms under this category are divided into policy-based, value-based and model-based algorithms.

In the recent past researchers have utilized machine learning algorithms in modelling the volatility of the financial markets. This is because machine learning models are known to overcome most of the limitations of the traditional econometrics models (Rossi, 2018). Additionally, they are known to yield higher accuracy scores, handle complex datasets, identify new patterns better than the traditional econometrics models. Further to this, unlike GARCH models which follow economic assumptions the machine learning models are more data driven thus are bound to yield better results. The problem of volatility has witnessed a lot of utilization of the machine learning algorithms both of classification models such as random forest and support vector machine as well as the regression models which include kernel factory, AdaBoost, K-nearest neighbours, logistic regression, random forest, and support vector machine which are used to forecast the future stock prices (Lapitskaya et al., 2021). In this study we will focus on four of these methods namely; K Nearest Neighbor, Support Vector Machine, Neural Network and random forest with the aim of identifying the most optimal model among these in predicting future stock market volatility.

1.1.5. Stock Market Volatility Prediction

Neuro AI (2013) outlines four main ways of making predictions namely, Machine learning, technical analysis, fundamental analysis and time series analysis. Fundamental analysis involves identification of certain factors that impact a counter such as the company's profitability, future business plans, financials, among others. The fundamental analysis provides a high-level analysis of the company and incorporates the operating environment of the company. On the other hand technical analysis focuses on the use of historical datain projecting future performance of a counter or the market. The technical indicators therefore seeks to establish the pattern of movement from past observations which come in handy in the prediction of future values. Some of the key technical utilized in predicting future stock prices include; Moving Average Convergence Divergence (MACD), rate of Change, bias, among others.

Machine Learning techniques have been utilized in the recent past in prediction of the stock market volatility and have proven to achieve better results than the other models (Jiang et al, 2020). Time series analysis differs from the rest in the system used in predictions is highly

volatile and has a lot of noise. This therefore causes the series to exhibit stylized facts such as regularity.

1.2. Statement of the Problem

The capital markets industry plays a critical role in the growth of economy of any country. In Kenya for example the market capitalization for 2021 accounted for 21.4% of the Gross Domestic Product of the country. In 2022, this figure stood at 20.16%. To this effect, the stock market is referred to as the barometer of the country's economy. Trading in the stock market therefore is an attractive venture for many people set for the risk that it carries of high volatility and uncertain returns. The market is therefore characterized by speculation on the performance of different counters and products. Investors therefore seem to shy away from participating in the market given the risks associated with participating in this market which if not taken care of can result to huge losses of money given the amount of investments involved (O' Neil, 2009).

The main concern for any investor is mainly to determine which product to invest in and the specific counter under that product. The investment advisors and stockbrokers in the Kenyan market mainly rely on past performance of different products, fundamental analysis of different counters and their experience in the industry when advising their clients. This, however, can be subjective as it is only limited to the level of research and diligence of the financial advisor (Njuguna, 2021). Having a tool that can accurately predict the future performance of different products as well as the volatility associated with these products is detrimental as it would solve many problems in the market (Adebiyi et al. 2012). Further, having an accurate estimation of future volatility and returns will come in handy in risk management, portfolio optimization and construction, pricing of derivatives and making buy or hold decisions on various stocks (O' Neil, 2009).

Various studies have been conducted on this area especially using ARIMA and GARCH models such as Gokcan (2000), Bildirici and Ersin (2009), Lelit (2017), among others. In the recent past researchers such as Anders (2021), Gerholm & Lindberg (2021), Hajizadeh et al (2012), Roh (2007), among others have also utilized machine learning algorithms in this exercise such as Support Vector Machine, Neural Network, K Nearest Neighbor, Naïve Bayes Classifier and Linear Discriminant Analysis. From these papers there is no conclusion on the best machine learning model for this task since they all have different strengths. This paper

therefore seeks to identify the most appropriate machine learning model in predicting stock market volatility. This will be done by analyzing the outputs through the out of sample method as well error metrics namely; Mean Absolute Error, Mean Squared Error and Root Mean squared Error.

Additionally, research work done in the Kenyan market such as Njuguna (2021), Chemutai (2021) have only focused on studying the performance of the NSE 20 Share Index which is the price index for the equities market capturing the top 20 companies by price. This has therefore limited the findings to the equities market thus not analyzing the volatility of other products such as derivatives, bonds, Exchange Traded Funds and Real Estate Investment Trusts which are also available in the market. Failure to analyze these products therefore fails to give a good view of the market since it is only skewed to the equities market. Alternative products such as derivatives are known to have lower volatilities and high level of returns (Antoniou and Holmes ,1995) thus this paper will seek to confirm this. The volatilities of traditional products weree compared to that of alternative products and inference made on what products one should invest in depending on their risk appetite.

1.3. Research Objectives

1.3.1. Main Objective

The main objective of the study is to determine the optimal machine learning model in predicting stock market volatility as well as compare the volatilities of different products.

1.3.2. Specific Objectives

Some of the specific objectives of the study will include;

1. To forecast the volatility of Traditional and Alternative products at the NSE using different machine learning models
2. To perform a comparative analysis of the best model in predicting future volatility of different products.

1.4. Significance of the Study

Investors- This study will provide greater insights on the different products offered in the Nairobi Securities Exchange, their volatility trend and future expectation of the same hence helping them make informed investment decisions.

Policy Makers and Key Capital Market Industry Players- This study will help the key market players such as the Capital Markets Authority, Nairobi Securities Exchange among others in understanding volatility thus informing policies around risk management, among other key aspects in the Kenyan capital market industry.

Academia: This study will also contribute to the literature available on machine learning modelling in the Kenyan industry thus the study findings will inform other future studies.

Chapter 2: Literature Reviewed

2.1. Theoretical Review

2.1.1. Efficient Market Hypothesis

The efficient market hypothesis states that stock prices should be a reflection of all the relevant information for that counter. This means that the prices are give precise signals with regards to asset allocation (Fama & Malkiel, 1970). The Efficient Market Hypothesis outlines three tests of market efficiency as follows; Strong form efficiency which investigates whether some investors have access to information that is not available to the rest of the market, Semi-strong form tests which looks at the whether the prices are able to respond effectively to market information and weak form tests which investigates historical information on prices.

Fama (1965) further postulates that in an efficient market, the price responds quickly to changes in the market hence past variables may not be relevant in predicting the trend of future price. The hypothesis therefore supports the random walk theory which states that successive price changes are independent and identically distributed variables hence the price series has no memory.

2.1.2. Random Walk Theory

The origin of the random walk theory is traced to a statement made by Maurice Kendall during a statistical conference held in London in 1970 through the presentation of his research paper titled the analysis of economic time series part one with the subject of discussion being the behavior of stock and common prices. His findings were informed by his observations of the regular stock price movements which proved to be absent (Brealey et al (1995)). This theory was further supported by Fama in 1960 who argued that in a vibrant market made up of intelligent investors, the listed counters will reflect all the available information. He concluded that in an efficient market no information or analysis can be expected to result in the outperformance of an appropriate benchmark.

2.1.3 Chaos Theory

According to Royal Society of London, Chaos is the stochastic behavior of a deterministic system. Kuchtta (2012) indicated that any chaotic system is characterized by random results on repeated experiments. One of the main characteristics of a chaotic system is that they are heavily dependent on the initial conditions. This therefore implies that any misspecification of the initial conditions will in turn tamper with the rest of the system. This will rob the system of

its predictive power as it will cause the errors to behave in randomly (Kuchta, 2012). Moreover, the errors will propagate in unpredictable ways, making forecasting impossible. In general Cohen (1997) defines a chaotic systems as a system which portrays both global determinism and local randomness. Non linearities and chaos in asset prices makes it possible to predict the future asset prices. Predictability of the prices will therefore go against the Efficient Market Hypothesis (Persaran, 1992).

2.2. Empirical Literature

Njuguna (2021), undertook a study aimed at modelling stock market volatility using random forest. She utilized five technical indicators from past prices of the Safaricom Shares and the NSE 20 Share Index for a period of 12 years in predicting the future price. These technical indicators include; Price Rate of Change (ROC), Moving Average Convergence Divergence Oscillator (MACD), Relative Strength Index (RSI), Williams % R and Stochastic Oscillator. The predictive model yielded 94.35% & 80.05% prediction accuracy rate for the NSE-20 and Safaricom daily share prices movement. Anders (2021) conducted a study of the Baltic Stock Market aimed at comparing predictive power machine learning and econometric models in predicting stock return and volatility. They were able to utilize different machine learning and econometric models namely; the support vector regression ,K-Nearest neighbours, random forest, Garch-ANN, autoregressive moving average model. He evaluated the performance of these models using six metrics namely; Mean Absolute error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage error (sMAPE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and Standardized Residuals (SR). Results from the analysis indicated that support vector regression and k-nearest neighbour, predict the returns better than autoregressive moving average models for most of the metrics, while for the other approaches, the results were not conclusive. Their analysis also discovered that training and testing sample size plays an important role on the outcome of machine learning approaches. Chemutai (2021), forecasted stock market volatility at the Nairobi Securities Exchange by comparing between Asymmetric GARCH Models and Neural Networks. She utilized the daily NSE 20 Share Index for the period between January 2021 to February 2021. She compared two GARCH models namely; EGARCH and GJR-GARCH models using AIC an BIC tests then utilized RMSE and MAE in evaluating the predictive power of the Artificial Neural Networks and the GARCH models. Results indicated that the EGARCH model was a better model in modelling volatility while in predictions of future values the Artificial Neural Networks had the highest predictive capability. Gerholm & Lindberg (2021) compared the

performance of different machine learning models in making market predictions with different time horizons. They targeted securities listed in stock exchanges based in America whereby they utilized discrete technical indicators in predicting the price movements of stocks ten and 30 days into the future. Results from this study indicate that a stock's price will increase the following trading day thus the results of this study supported the random walk theory. The technical indicators utilized included; Momentum, Moving Average, Relative Strength Index and MACD, Williams %R and Stochastic Oscillator. The models investigated included; Random forest and Support vector Machine.Bildirici & Ersin (2019) undertook a study on Moroccan single stock market return volatility with thet aim of improving the forecasts of the GARCH family models with the Artificial Neural Networks. The Back Propagation Neural Network was combined with GARCH models and utilized as an input of the Neural Network model. Results from this model proved the efficiency of performance of GARCH models through neural networks. Further, the study identified that the combination of MRS-GARCH and EGARCH with neural network outperformed predictions from other models. Through a study aimed at developing a hybrid model in forecasting the volatility of the S&P 500 Index return, Hajizadeh et al (2012) utilized two hybrid models based on EGARCH and Neural Networks models. The inputs of EGARCH hybrid model were historical values of the explanatory variables while those of the Neural Network were series of the simulated data and explanatory variables. Results from these two models were then compared to those of an EGARCH model. The second hybrid model yielded the highest accuracy in prediction of volatility. Kristjanpoller et al (2014) conducted a study aimed at forecasting volatility of energy prices using hybrid Artificial Neural Network and GARCH type models. The results from the research indicated that the EGARCH-ANN was better than other models in the Chinese energy market. Roh (2007), conducted a study with the aim of coming up with a hybrid model from time series models and Artificial Neural Network. The hybrid model achieved a lower MAE thus implying it is a better predictor of volatility. Caligiuri (2018) utilized a multiple linear regression to estimate the volatility of the Standard and Poor's 500 Index with the independent variables being GDP, Money Supply and Unemployment Index. The results of this model were compared with the Chicago Board Options Exchange Volatility Index. The model proved not to be strong in predicting future values but the results exhibited a similar trend to that of the VIX Index. Dai et al (2020) utilized a combined approach in forecasting stock market volatility. They combined stock market implied volatility and oil volatility as predictors in a regression model aimed at

2.3. Conceptual Framework

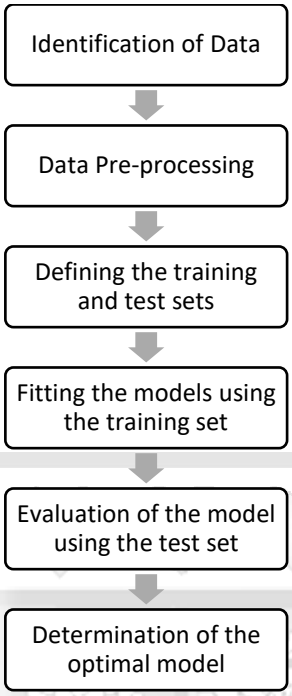

Figure 2:Conceptual Framework

2.4. Summary of Literature

The literature reviewed highlights some of the works that have been done in the area of volatility most especially the papers utilizing machine learning methods. Most of the papers have focused on comparing the performance of machine learning models versus GARCH and ARIMA models. This however, little has been done to investigate the most optimal machine learning model. Additionally, the studies conducted on the Kenyan market have only focused on the top 20 companies by price through the use of the NSE 20 Share Index. This fails to incorporate the alternative products and the bonds market present in the Kenyan market. This paper therefore seeks to compare the volatility of both the traditional and alternative products information that is very crucial for investors in making decisions.

Chapter 3: Research Methodology

3.1. Research Design

This study will utilize the descriptive research design as it is aimed at providing an in-depth analysis of the Nairobi Securities Exchange volatility and conclusions derived from the analysis.

3.2.Population

Mugenda & Mugenda (1999) defines a population as the entire group of people being studied. Further to this, a population is the group a researcher is seeking to make inferences about. In this study the population will be the NSE 20 Share Index since the establishment of the Nairobi Securities Exchange. The data to be utilized for the study will cover a 10-year period from January 2012 to March 2023. The testing set will constitute 80% of the entire dataset while the test set will be made up the remaining 20%.

3.3.Data Collection

The study utilized secondary data for the various products from the Nairobi Securities Exchange. Data available for the various products covered different time periods since the different products were introduced to the market at different times. The following data was used for the various products; Equities market, the NSE 20 Share Index which is a price index tracking the performance of the top 20 counters was utilized. The Data available on this index covered the period between January 2002 to March 2023. On the Bonds market, two bonds were sampled whose maturity dates are in 2023, an Infrastructure bond and a Fixed Income bond. The two were selected as the price of the bonds could be tracked for the whole period since the issuance to maturity thus giving insights on the movement of the clean price throughout the life cycle of a bond. The derivatives market was launched on 04th July 2019 hence the data was obtained from then covering the daily Mark to Market price for the 12 contracts available in the market. The market is mainly made up of the Single Stock Futures namely; Absa, Safaricom, KCB, Equity, British American Tobacco, EABL, NCBA, Co-Operative Bank, Standard Chartered Bank, I&M bank and two Indices namely; the NSE 25 Share Index and the NSE 25 Mini Index. Available data for the derivatives market captured the period between July 2019 and March 2023.

Exchange Traded Funds, under this category only one counter is listed the Absa New Gold ETF which was listed in September 2017, the average price for this counter was therefore

observed for this period. Under the Real Estate Investment Trusts product the NSE currently has four listed REITs namely; the Acorn D-REIT, Acorn I-REIT which are listed on the Unquoted Securities platform which is a replica of the Over the Counter Market. The Laptrust Imara I-REIT is listed under the restricted of the Main Investment Market Segment, the REIT was listed on 22nd March and is currently not trading. To this effect the ILAM Fahari I-REIT is the only REIT which is currently trading publicly hence the daily average price of this counter was utilized covering the period between 2015 to 2023.

While making comparisons of the volatilities of the traditional versus alternative products the data used covered the period between 2019 to 2023 in a bid to accommodate all the products.

3.4.Data Analysis

In this research data was analyzed using python which is a programming language mainly utilized for machine learning and software development. The returns used is defined as follows: $r_t = \log\left(\frac{P_t}{P_{t-1}}\right)$.

3.4.1. Feature Selection

In this research technical indicators were utilized in the analysis which are numerical statistics obtained from the past performance of the targeted variables. The indicators are able to analyze the trend, direction and volatility of past occurrences and utilize these findings in the prediction of future values of the targeted variables. In this research six indicators were utilized namely; Simple Moving Average (SMA), Moving Average Convergence Divergence (MACD), Exponential Moving Average (EMA), Rate of Change (ROC), Relative Strength Index (RSI) and Stochastic Oscillator.

3.4.1.1.Simple Moving Average

This indicator is used identify the trend of stock prices. It is the simplest form of Moving average and it takes the average price over a certain period of time. This average is known as moving since it involves plotting bar by bar and developing a straight line that changes as the average values changes. In interpretation of this indicator, upward movements of the SMA line indicate that the trend is increasing while downward movements are an indication of a decreasing trend. The indicator can also be used in smoothing data whereby longer period of SMA result into smoother indicators and data.

3.4.1.2.Exponential Moving Average

It is a weighted moving average indicator. The difference between this indicator and the Simple Moving Average is that EMA allocates more weight to the recent observations hence able to predict the trend better than SMA. The indicator is therefore very sensitive to price movements which is a good thing as it is able to identify future trends early enough. On the other hand, this means that the indicator will have many short term movements compared to the SMA.

The indicator is given by the following equation:

EMA = (K x (C - P)) + P

Where: C = Current Price

P = Previous periods EMA

K = Exponential smoothing constant

3.4.1.3.Price Rate of Change (ROC)

This indicator focuses on the momentum of the stock prices by analyzing the rise or fall of the prices. It is calculated as a percentage of the difference between the current price and the price n periods ago. In interpreting this indicator negative and positive values are analyzed with positive values indicating upward momentum and negative values representing downward momentum. The indicator can also be used as a trend indicator most especially to indicate divergence (a situation whereby the price change is moving in the opposite direction to stock price). It is also an indicator of overbought and oversold circumstances which values above +30 indicating overbought situation and values less than -30 being an indicator of oversold stocks.

3.4.1.4.Stochastic Oscillator

This is an indicator which focuses on the momentum of the target variable by comparing the closing price to other prices. It is made up of two lines namely; %D and percent K. %D is a measure of the three day Simple Moving Average of the indicator. On the other hand, %K represents the indicator itself. In interpreting this indicator the trend of the two lines is observed and if they cross at any point it is an indicator of an upcoming shift in terms of the trend.

3.4.1.5.Relative Strength Index

This indicator analyzes both the speed and direction of the change in stock prices. Its applicability mainly focuses on identifying whether a certain stock is overbought or oversold.

The computation of this indicator is done in two stages namely; determination of average losses and gains of the counter and secondly obtaining the ratio of the obtained values. In interpreting this indicator, the values are analyzed with values above +70 being an indicator of overbought stocks hence tis is stock has high chances of being sold in the near future. On the other hand, values below +30 indicate that the counter is oversold hence it is expected to be bought in the near future.

3.4.1.6.Moving Average Convergence Divergence (MACD)

This model is based on the evaluation of convergence or divergence of two moving averages. Convergence is a situation where by two Moving Averages are moving in a similar direction while divergence occurs when two moving averages move in opposite directions. It is calculated by taking the difference between the shorter Moving Average from the longer Moving Average as shown by the equations below:

MACD= 12day EMA – 26 day EMA

Signal line= 9 day EMA of the MACD line

MACD histogram = MACD line- signal line

3.5.Machine Learning algorithms to be utilized in the study

3.5.1. Random Forest

Random forest is an extension of Decision trees. According to Breiman et al.1984, decision trees are hierarchical structure in which every internal node contains a test on an attribute , each branch corresponds to an outcome of the test and each leaf(terminal) node gives a prediction for the value of the class variable. Decision trees can be used for both regression and classification techniques depending on the problem. The following are terms and parts of a decision tree: Node: A part of the tree which performs a test on a particular feature and applies it to a particular subset of the dataset. The outcome of this test determines the next node to be used. Root: It is a special type of node and each tree has only one root which is the origin of the model. This means that this first test is applied to the entire dataset. Leaf: A special type of node and each tree has at least one leaf. Leaves are the final nodes of a tree and contains a subset of the data that will aid in making predictions hence they do not perform tests and they do not transition to other nodes. Branch: This is the path from the root node all the way to some leaf node and serve as a representation of a series of tests and decisions that ultimately lead to a specific subset of the data. The process of utilizing decision trees in Machine Learning

involves; picking a feature in the dataset that splits the data meaningfully, splitting the data into subsets based on the feature selected which creates a node and the subsets to be used by the nodes and decide on whether to continue splitting the data or not ,if yes repeat the process above. Two metrics exists which are used in evaluating the optimal decision tree namely; the minimum sample split which aims at determining the number of entries that should be in a subset before splitting it and the maximum depth which aims which evaluates the largest number of nodes used from the root to leaf node.

Random forest is an extension of the decision trees whereby a collection of decision trees are combined and utilized in making predictions. The trees are random because each subtree trains on a random subset of the data and each subtree uses a random subset of the features. Random forests yield a better model fit as they seek to minimize on the bias and high variance. Random forest can be utilized both for classification and regression. As a classifier, the random forest model is able to classify information into broad categories such as Yes or No example shown below. This makes visualization easier of the different classes thus is very easy to use.

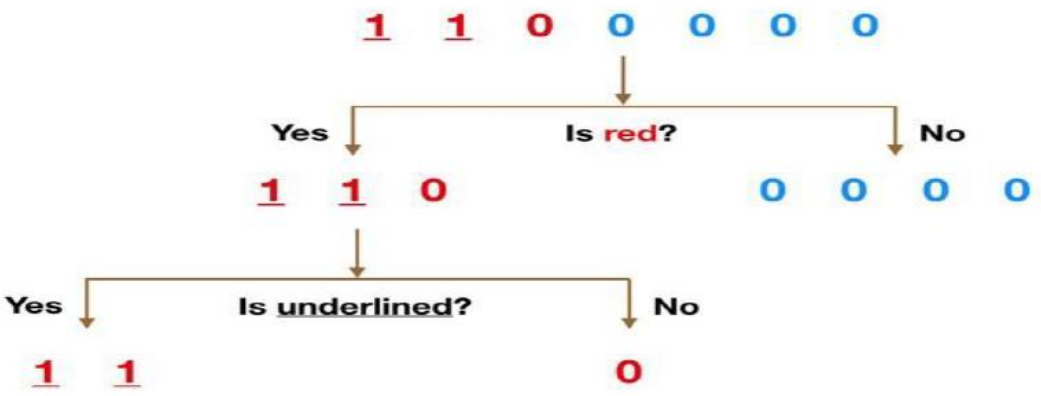

Figure 3: Implementation of a Random Forest Classifier

In regression, the random forest algorithm utilizes a number of individual decision trees which form an ensemble. Every tree in the models outputs a class of prediction and the class with the highest number of votes is taken as the prediction of the model.

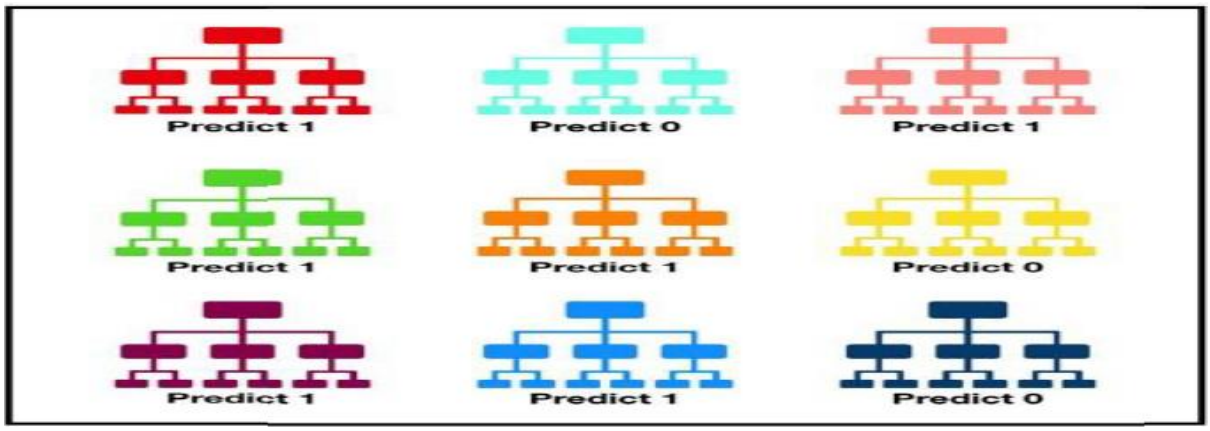

Figure 4: Implementation of the Random Forest regression

The method evaluates the predictions using the residual sum of squares which are given by the following equation:

RSS=$\sum left\,(Y_i - Y_L^*)^2 + \sum right(Y_i - Y_R^*)2$

Where: Y^*_L= mean y-value for left node

Y^*_R= mean y-value for right node

3.5.2. Neural Networks

According to Gurney, K. (1997)., neural networks are interconnected assemblies of simple processing elements whose functionality is loosely based on the animal neuron. They are sets of algorithms designed to mimic the human brain functionality and can be used for both classification and regression purposes. The key components in a Neural Network algorithm include the following ;Input Layer which is the data available, Hidden Layers which lie between the input and output layers, they are called hidden layers since the input and output are known but the operations that take place at this stage are hidden, the output layer gives us a metric for comparison with our target, weights and biases exists between all the layers and come in handy in the calculation of the different values between the layers and Activation Function which are used in converting large outputs into smaller values and in promoting non-linearity into our neural network.

Components of Neural Networks

Neural Networks are made up five key components namely; Input layer; This is the data we have. Hidden layer (s); These are the layers that are between the input layer and the output layer. they are called the hidden layers because we know the inputs and we get the outputs but we don't know happens in between as these operations are hidden from us. Don't worry if this

doesn't make sense right now as we are going to see how everything connects when we explore how neural network works. One more thing worth mentioning is that stacking layers one after the other produces a deep network. The number of hidden layers solely depends on you. Output Layer: This is the last layer of a Neural Network. It's what we want to compare our targets to. Weights and Biases: Between each of these layers (input layer, hidden layer and output layer), there are weights and biases. These weights and biases are very crucial in determining and computing the different values between each of these layers. Activation Functions**:** These are mapping functions. they are used to convert large outputs into smaller values and promote non-linearity into our neural networks

An example of a shallow neural network:

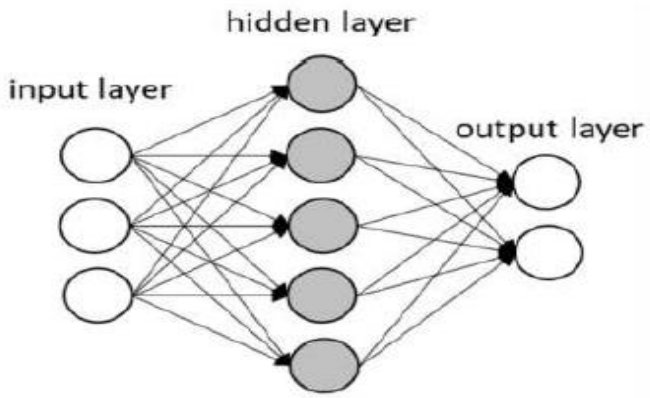

Figure 5: The theory behind a shallow Neural Network

Below is a deep learning neural network with three hidden layers.

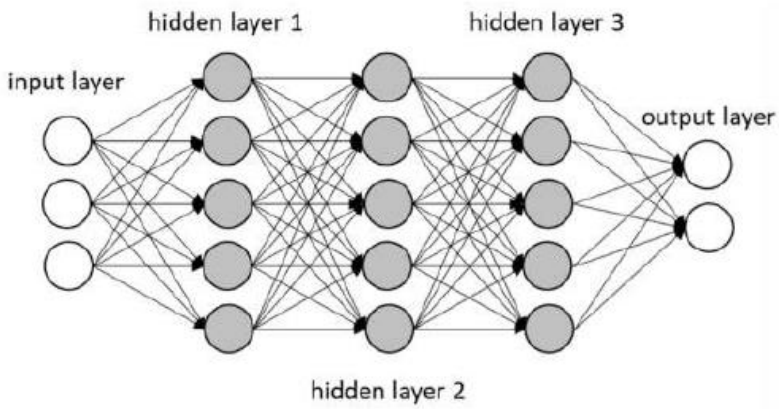

Figure 6: The theory behind Deep Neural Network

3.5.3. Linear Regression

Regression is a statistical method used to evaluate the relationship between independent and dependent variables. Linear regression allows us to predict some dependent variable - *y* - based on independent variable *x* by finding an optimal linear relationship between them. Remember from high school the formula for a line equation:

y = *m***x** + *b*

Linear regression algorithms will seek to find optimal values for **m** (the slope of the line)and **b**(The intercept, also known as the bias). This line equation should then allow us to compute a value for y, our dependent variable, so long as we have some input x.

The concept of machine learning is applied in that we come up with an optimal line equation for our model? what is the best value for **m** and **b?** A first step to optimizing **m** and **b** is to create an **error function** (often referred to as cost function as well) to assess how good our line y = mx + b is.

If you look at the graph above, our regression line does not pass through most of the points, suggesting it's fairly error-prone. To compute the error function we can perform the following steps on each data point in our data set; Compute the difference between your data for the dependent variable, and the prediction from the regression line, Square that value, we want all the differences to be positive, add up all the values you've obtained and divide by the number of data points.

In summary, your error function is the average value of $\mathbf{(y_i - (mx_i + b))^2}$ for all your data points. We want to tweak the values of **m** and **b** in order to minimize this function.

This is where the **gradient descent algorithm** comes into play. Imagine plotting the possible values of **m** and **b** against the error function's value. You'd end up with a shape like this.

The fitted model was of the form:

$Predicted\ Returns = f(\text{SMA, EMA, MACD, ROC, RSI})$

3.5.4 eXtreme Gradient Boosting (XGboost)

This is a decision tree based model that utilizes the gradient boosting framework. It can be used both for classification and regression problems. The model is given by the following equation:

$\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{f}_{\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{i}})$ given that $\boldsymbol{f}_{\boldsymbol{k}} \in \boldsymbol{F}$

Where;

$\hat{\boldsymbol{y}}_{\boldsymbol{i}}$=predicted value for the ith row

K= number of boosted trees

$\boldsymbol{x}_{\boldsymbol{i}}$**=** the ith data point

f=function in the space F

F= set of all possible Classification and Regression Trees

The boosted predictions are defined by the following equation:

$\hat{y}_i^{(k)} = \sum_{k=1}^{K} f_k(x_i)$

The method sums up the prediction of the previous trees with that of the new tree in coming up with the gradient boosted trees. This therefore follows the following equation:

$$\hat{y}_i^{(k)} = \hat{y}_i^{(k-1)} + f_k(x_i)$$

In the definition and optimization of the trees an objective is defined as follows:

$Obj^{(k)} = \sum_{i=1}^{N} L(y_i, \hat{y}_i^{(k-1)} + f_k(x_i)) + \varphi(f_k)$

Using Newton Raphson method the above equation ends up being as follows:

$$Obj^{(k)} = \sum_{i=1}^{N} [g_i f_k(x_i) + 1/2 h_i f_k(x_i)^2] + \varphi(f_k)$$

Where $\boldsymbol{g}_{\boldsymbol{i}}$ and $\boldsymbol{h}_{\boldsymbol{i}}$ are the first and second order partial derivatives for the loss function.

The regularization term is give by the following equation:

$$\varphi(f_k) = \gamma T + 1/2\rho \sum_{j=1}^{T} w_j^2$$

The learning objective function is obtained by combining the loss function with the regularization function and is given as follows:

$$Obj^{(k)} = \sum_{i=1}^{N} [g_i\, w_{q(x_i)} + 1/2 h_i w_{q(x_i)}] + \gamma T + 1/2\rho \sum_{j=1}^{T} w_j^2$$

3.6.Models Validation and Testing

3.6.1. Out- of -sample Validation

This involves testing the accuracy of the model using unseen data not involved in modelling. The data is therefore split into two main sets namely; training and test sets with the training set being utilized to train the model while the test set is utilized for testing the accuracy of the model. The data collected in this case will therefore be divided into two with 80% of the data set being taken as the train set while the remaining 20% will be utilized as the test set.

3.6.2. Error Metrics

The evaluation of the machine learning models will be done using the following metrics:

3.6.2.1.Mean Absolute Error

Chai & Draxler, 2014 define the Mean absolute error (MAE) as a statistical method of evaluating the performance of a model. It is calculated using the formula below:

MAE=$1/n \sum_{t=1}^{n} |y_t - x_t|$

3.6.2.2.Mean Squared Error

Wang & Bovik, 2009 define the Mean squared error (MSE) as a measure of the similarity between two values. The applicable range for this metric is [0, ∞] and its equation is outlined as follows:

$$MSE = 1/N \sum_{t=1}^{n} (y_t - x_t)\text{^}2$$

3.6.2.3.Root Mean Square Error (RMSE)

The square root of he Mean Squared Error is known as the Root Mean square error and is obtained through the equation below:

$$RMSE = \sqrt{1/N \sum_{t=1}^{n} (y_t - x_t)\text{^}2}$$

4.0. Data Analysis, Results and Discussions

4.1. Introduction

The focus of this chapter is the results obtained from the analysis as well as the findings thereof. The aim of the paper was to determine the most optimal machine learning model in forecasting future returns at the Nairobi Securities Exchange. Additionally, the paper sought to compare the volatility of traditional products to that of alternative products. The analysis leveraged on the technical indicators identified above which were used to fit the different machine learning models. The analysis was done in python which is a programming package used for various tasks such as fitting machine learning models, software development, among others.

4.2. Data Cleaning and Pre-Processing

Data obtained from the Nairobi Securities Exchange covered the period between 2002 and 2023 depending on when the product was introduced in the market. Missing values were dropped to ensure completeness and accuracy. Stationarity test was conducted for the returns using the Augmented Dickey Fuller test while normality test was conducted using the Shapiro Wilk test.

4.3. Feature Extraction

Technical indicators are variables obtained from historical stock data and are very useful in identifying the trend and direction of stock data. The following technical indicators were utilized in this analysis; Simple Moving Average, Exponential Moving Average, Relative Strength Index, Rate of price Change, Stochastic Oscillator and Moving Average Convergence Divergence. The indicators were then utilized in making future predictions of future performance.

Interpretation for these indicators was done as follows. Relative Strength Index Oscillates between 0 to 100 with values above +70 indicating the stocks are overbought thus stocks are expected to be sold. On the other hand, values below +30 indicate that stocks are oversold hence they are likely to be bought in the short run. Rate of Price Change: upward movements and positive values are an indication of oversold stocks thus signaling buying. On the other hand negative values are a sign of overbuying in the stock market thus stocks are likely to be sold in the short run. The MACD generates signals of possible sell when the MACD line is below the zero line. Purchases on the other hand are inferred when the MACD line is above the zero line. Stochastic Oscillator values of this indicator revolve around 0 to 100. Values above +80 are an indication of an overbought stock thus alludes to a sell in the future. On the other hand, values below +20 are an indication of oversold positions thus indicating a buy in the future. Simple Moving Average and Exponential Moving Average are interpreted by the direction of the line with upwards trend indicating a rising trend while downward movement is an indication of low trend.

4.4. Empirical Results

4.4.1. Equities Market

The NSE 20 Share Index constituted 5,298 observations covering the period between January 2002 to March 2023, a period of 21 years on a daily basis. The Index has been sporadic for the period under review recording highest value of 6,161.46 and the lowest value being 1,004.70. This is attributable to the varying price performance of the top 20 counters from time to time.

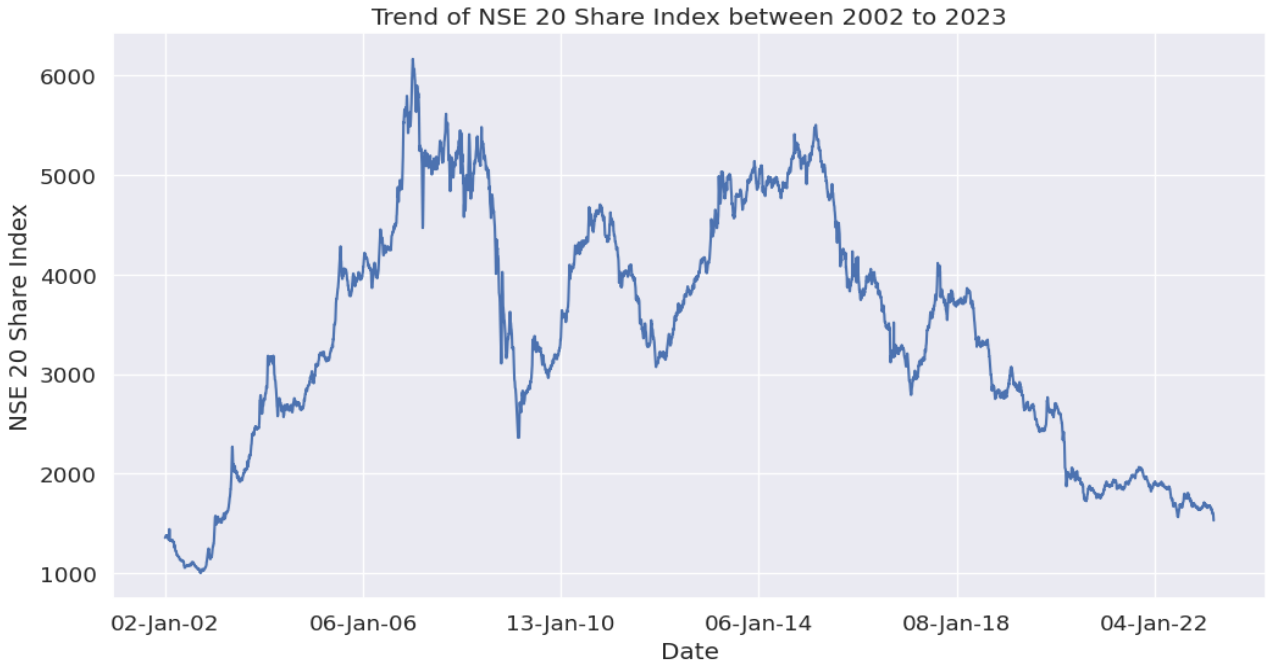

Figure 7: Trend of the NSE 20 Share Index between 2002 to 2023

The log returns for the NSE 20 Share Index are given

The daily returns were obtained by taking the log of the change in price between two consecutive days using the following equation. From the figure below, the returns are characterized by high level of volatility.

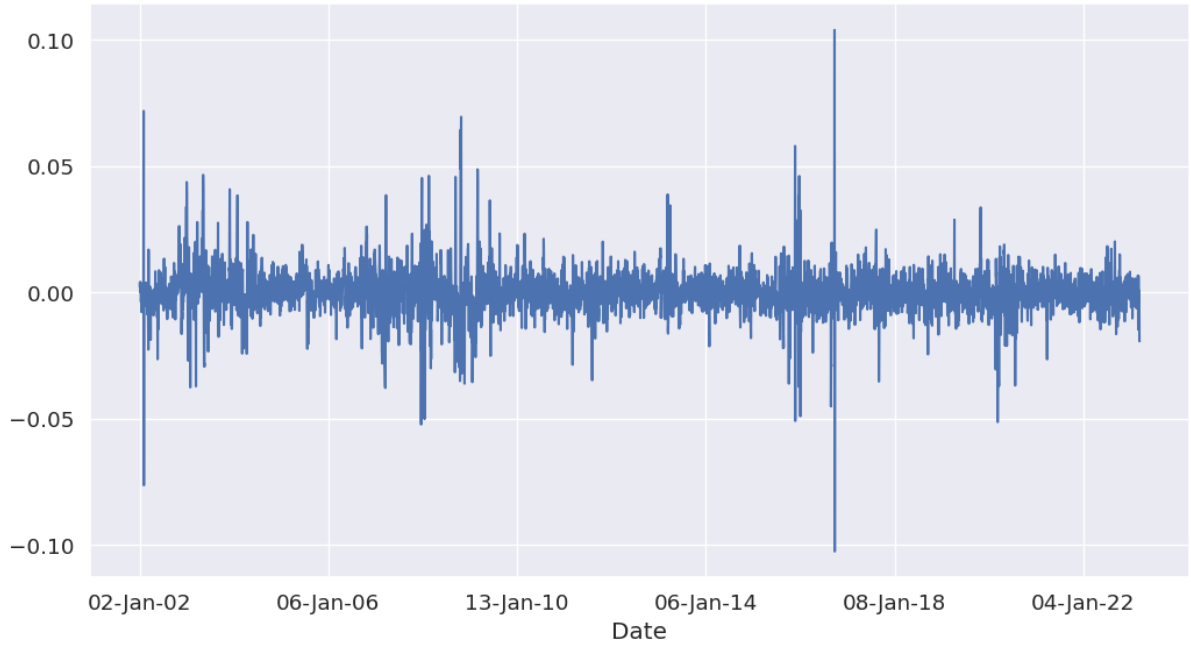

Figure 8: Trend of the NSE20 Log Returns for the period between 2002 to 2023

In forecasting the future volatility of the returns, the following methods were utilized namely; Random Forest, Linear regression, Muti-Layer Perceptron Regressor and the extreme Gboost. The evaluation of these models was evaluated using the following error metrics; Mean Squared Error, Mean Absolute Error, Root Mean Squared Error and R squared score. According to Table 1 the Linear regression model proved to be the most robust model recording the lowest error values and the highest regression value. These were as follows; Mean Squared Error at 0.00253, Mean Absolute Error of 0.03876, Root Mean Squared Error of 0.00253 and an R squared score of 0.9076 indicating that the model achieved a 90.76% accuracy in predicting future values.

Table 1: Prediction Evaluation of Various models in forecasting future returns of the Equities market

Metric	**Random Forest**	**XGboost**	**Regression**	**Multi-Layer perceptron Regressor**
Mean Squared Error	0.0046	0.0088	0.00253	0.042157
Mean Absolute Error	0.0523	0.0792	0.03876	0.08355

Root Mean Squared Error	0.00458	0.00884	0.00253	0.20532
R Squared	0.6963 or 69.63%	-0.1323	0.9076	-608.67

4.4.2. Bonds Market

Two bonds were analyzed under this category namely; the FXD1/2013/010 and the IFB1/2011/012 both of which are set to expire in 2023. The Yield Maximum yield was used as the parameter of evaluation for this market. The infrastructure bond turned out to be more volatile than the Fixed income bond recording a daily volatility of 8.2% compared to 5.7% recorded by the Fixed Bond. This is attributable to the high appetite for infrastructure bonds since they are tax free thus making their prices and the yield therefore very volatile.

Fixed Bond

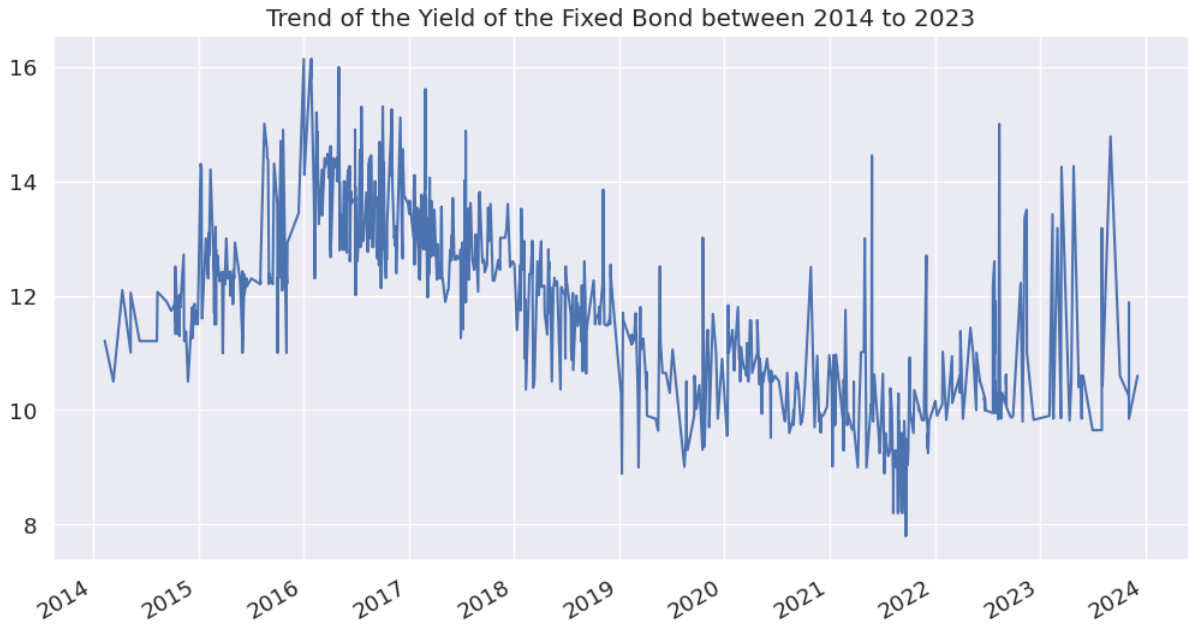

Figure 9: Trend of Fixed Income Bond between 2014 and 2023

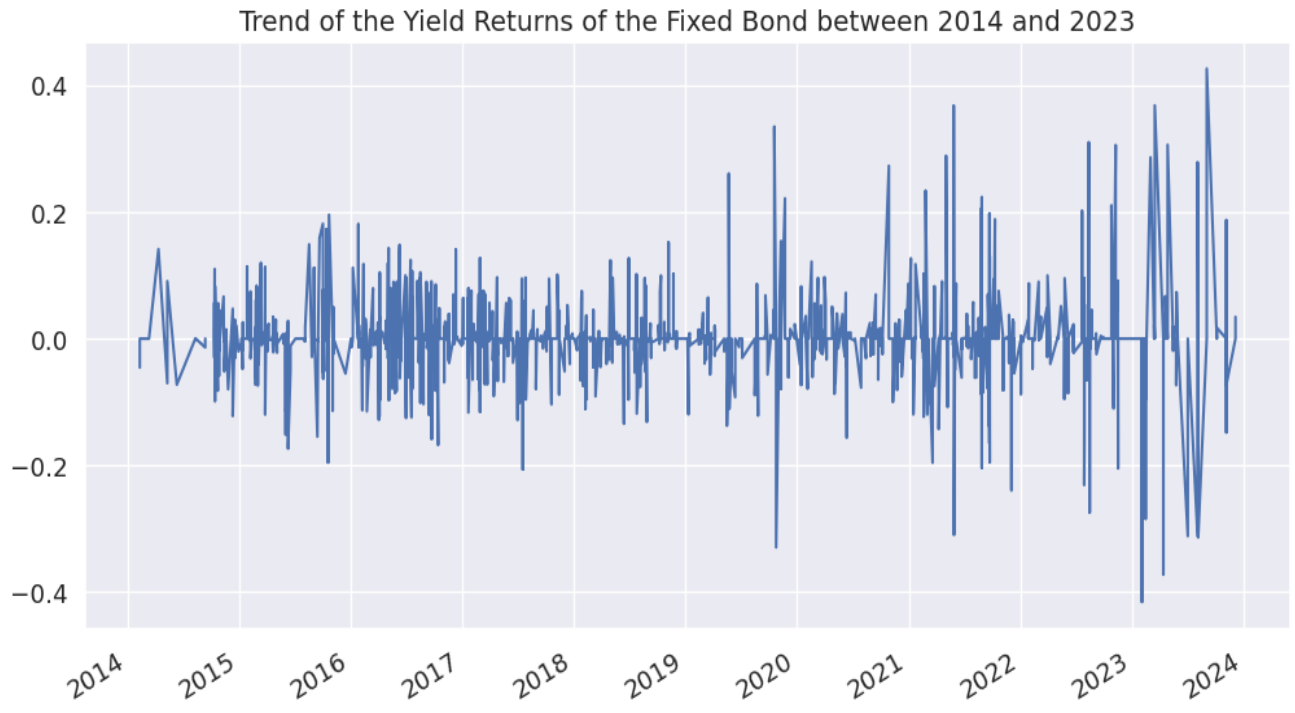

Figure 10: Trend of the Fixed Income Returns between 2014 and 2023

Infrastructure Bond

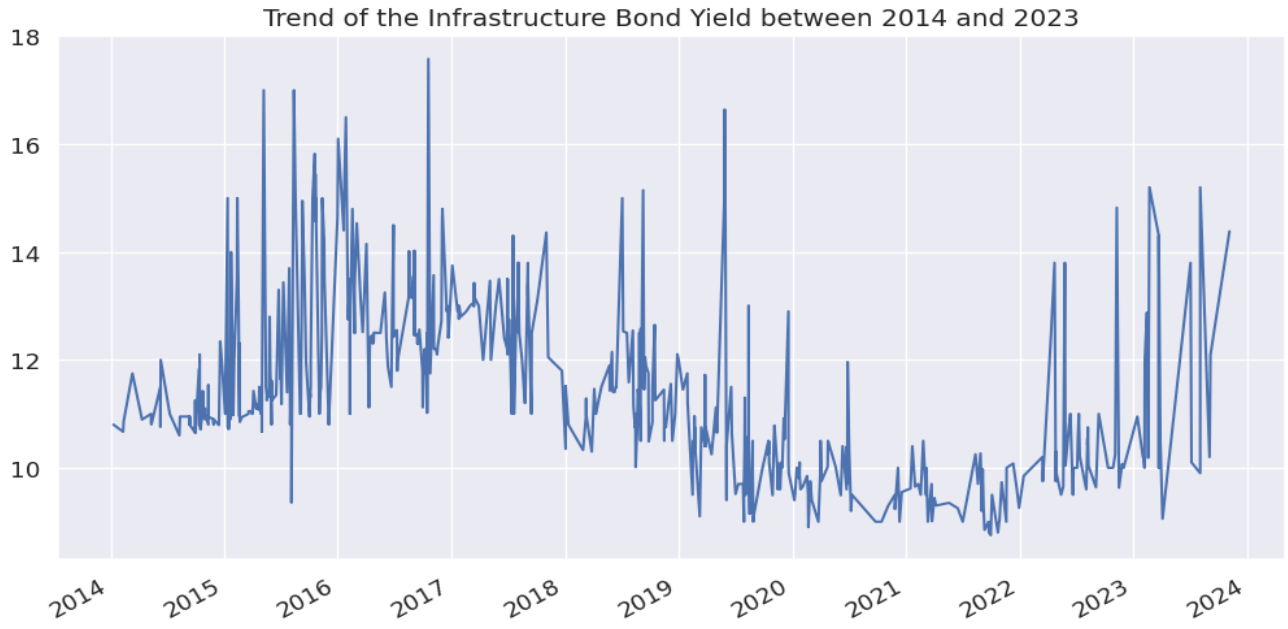

Figure 11: Trend of the Infrastructure Bond Yield between 2014 and 2023

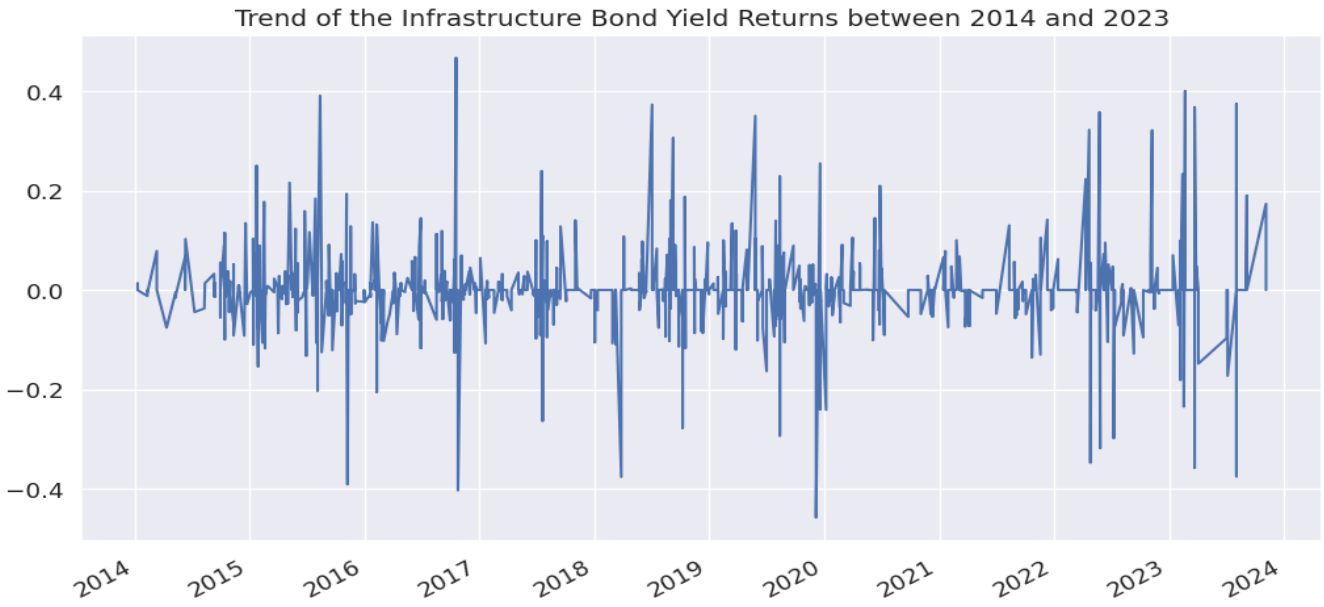

Figure 12: Trend of the Infrastructure Bond Returns between 2014 to 2023

In evaluating future predictions, the Multi-Layer Perceptron Regressor achieved the lowest error values this however, Random Forest and Regression achieved the best R squared scores.

Table 2:Prediction Evaluation of Various models in forecasting future returns of the Bonds market

	Error Metrics	IFB1	FXD
Random Forest	Mean Squared Error	0.0177	0.0093
	Mean Absolute Error	0.0794	0.05441
	Root Mean Squared Error	0.01766	0.0093
	R squared score	0.35592	0.8388
XGboost	Mean Squared Error	0.0616	0.03608
	Mean Absolute Error	0.23746	0.17126
	Root Mean Squared Error	0.0616	0.03608
	R squared score	-6.835	-1.43954
Regression	Mean Squared Error	0.0128	0.01025
	Mean Absolute Error	0.08642	0.07041
	Root Mean Squared Error	0.0128	0.01025
	R squared score	0.6616	0.80302
Multi-Layer perceptron Regressor	Mean Squared Error	0.00624	0.00274
	Mean Absolute Error	0.04709	0.02786
	Root Mean Squared Error	0.07901	0.05238
	R squared score	-11.887	-4.1425

4.4. 3. Derivatives Market

The derivatives market in the Kenyan was launched on 04th July 2019. Under this market twelve contracts. The two Index futures were found to be more volatile than the Single Stock futures recording daily volatilities of 3420% recorded for NSE 25 Index and 36550% for the NSE 25 mini-Index. In predicting future values the Regression model was found to be more robust than the other models. The Multi- Layer perceptron Regressor on the other hand achieved low errors.

Table 3:Prediction Evaluation of Various models in forecasting future returns of the Derivatives market

	Error Metrics	KCB	NCBA	NSE 25	25 Mini	SCOM	SCBK	Absa	BAT	COOP	EABL	Equity	IHMP
Random Forest	Mean Squared Error	0.014578	0.02238	0.00897	0.010496	0.03303	0.015575	0.02646	0.01047	0.01529	0.01653	0.0194	0.01024
	Mean Absolute Error	0.09282	0.11777	0.07199	0.075719	0.12546	0.09055	0.11144	0.080103	0.08745	0.094454	0.1012	0.08179
	Root Mean Squared Error	0.01458	0.022385	0.00897	0.010496	0.03303	0.015575	0.02646	0.01047	0.01529	0.01653	0.0194	0.01024
	R squared score	0.85123	0.91377	0.9139	0.92993	0.7326	0.93143	0.90502	0.92672	0.94258	0.898	0.8573	0.97393
XGboost	Mean Squared Error	0.02333	0.05958	0.02626	0.029832	0.05535	0.06899	0.04786	0.03729	0.06244	0.0499	0.0492	0.05937
	Mean Absolute Error	0.12703	0.21789	0.1429	0.15291	0.20832	0.25275	0.19639	0.17259	0.23431	0.205701	0.20314	0.23107
	Root Mean Squared Error	0.023332	0.05958	0.0263	0.029831	0.05535	0.06899	0.047859	0.03729	0.06244	0.0499	0.049185	0.059373
	R squared score	0.61888	0.38911	0.26201	0.434016	0.24916	-0.34566	0.68916	0.06973	0.04189	0.07013	0.07961	0.12347
Regression	Mean Squared Error	0.01174	0.02746	0.0125	0.01456	0.02837	0.0178	0.03373	0.00814	0.02033	0.01954	0.02018	0.018977
	Mean Absolute Error	0.08602	0.13757	0.08778	0.098777	0.12426	0.11497	0.14528	0.075982	0.11814	0.10909	0.10945	0.12956
	Root Mean Squared Error	0.01174	0.027465	0.0125	0.014557	0.02837	0.0178	0.033732	0.008144	0.02033	0.01954	0.020183	0.01898
	R squared score	0.90352	0.8702	0.83261	0.86524	0.80274	0.91043	0.84559	0.9556	0.8985	0.85742	0.84502	0.91046
Multi-Layer perceptron Regressor	Mean Squared Error	0.1099	0.00931	0.00318	0.00212	0.01128	0.00264	0.00403	0.00310	0.00648	0.002127	0.01292	0.00167
	Mean Absolute Error	0.07738	0.05530	0.03513	0.03571	0.06544	0.03284	0.03523	0.03365	0.041132	0.029102	0.03053	0.02230
	Root Mean Squared Error	0.33145	0.09648	0.05647	0.04606	0.10619	0.05141	0.06350	0.05571	0.08052	0.04612	0.11367	0.04078
	R squared score	-75.909	-0.6020	-2.4142	-0.3494	-1.7637	0.25272	0.45270	-1.0762	-0.5932	0.20562	-3.9159	0.58643

4.4.4. Exchange Traded Funds

The Absa New Gold Exchange Traded Fund is the only Exchange Traded Fund available in the Kenyan market and it was listed in 2017. The predictions prove that regression model achieved the lowest error metrics while also achieving high level of the R squared score.

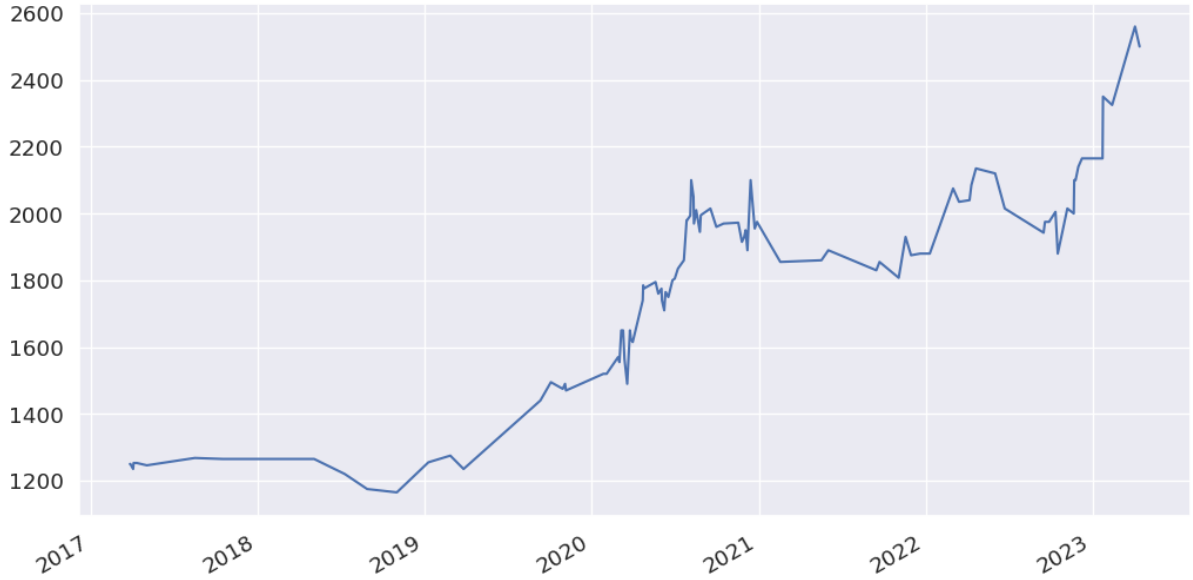

Figure 13: Trend of the Average Price of the Absa New Gold ETF

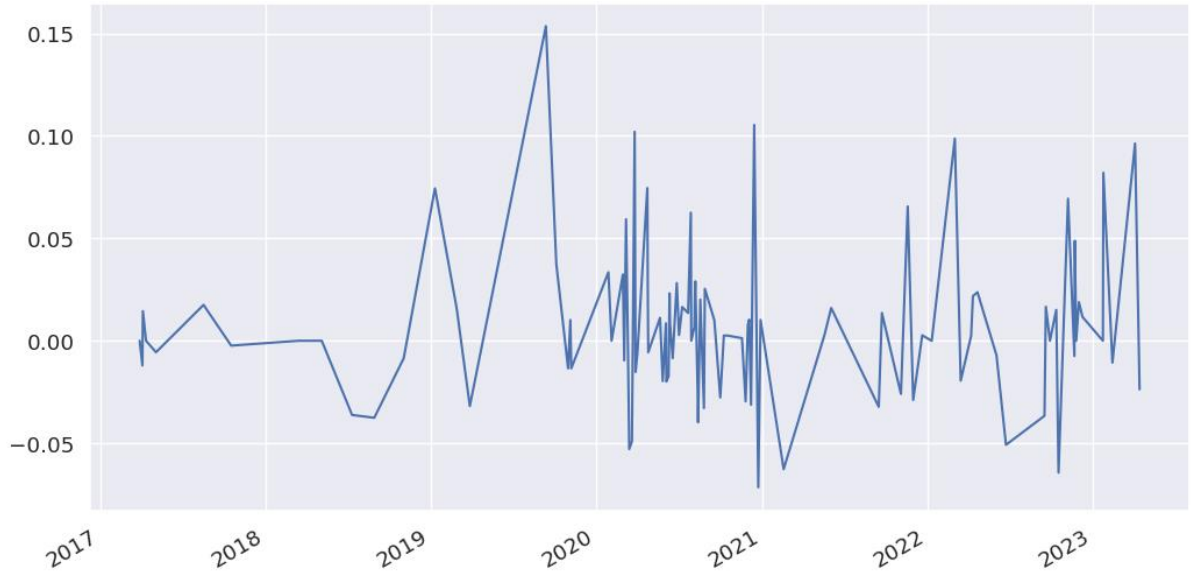

Figure 14: Trend of the Absa Gold ETF Returns between 2017 and 2023

Table 4:Prediction Evaluation of Various models in forecasting future returns of the Absa Gold ETF

Metric	**Random Forest**	**XGboost**	**Regression**	**Multi-Layer perceptron Regressor**
Mean Squared Error	0.01176	0.041491	0.009073	0.2486
Mean Absolute Error	0.10077	0.18445	0.08244	0.4280
Root Mean Squared Error	0.011764	0.041491	0.009073	0.4986
R Squared	0.91778	-0.022784	0.95110	-146.69

4.4.5. REITs (Real Estate Investment Trusts)

The Nairobi Securities Exchange has four listed Real Estate Investments Trust namely; The Acorn I-REIT, Acorn D-REIT, Ilam FAHARI I-REIT and the Laptrust Imara I-REIT. Out of the four two are listed in the Unquoted Securities Platform which takes the form of an Over-the-Counter market thus making price discovery difficult. Additionally, the Laptrust Imara I-REIT which was listed on 22nd March 2023 was listed as a restricted offer and was granted exemption from trading for the next three years. The Ilam Fahari I-REIT therefore remains to be the only active counter in this category. Price data for the REIT was obtained from 2015 when it was listed to March 2023. The price of the REIT has been deteriorating over the period of study due to lack of awareness and low uptake of the product.

Figure 15: Trend of the Average Price of the ILAM I-REIT

Figure 16: Trend of the ILAM I-REIT Returns

An evaluation of the predictive power of the various machine learning models fitted indicates that the Multi-Layer Perceptron was able to achieve the lowest error metrics. This however, regression model achieved high accuracy scores of the r squared score.

Table 5:Prediction Evaluation of Various models in forecasting future returns of the ILAM FAHARI I-REIT

Metric	**Random Forest**	**XGboost**	**Regression**	**Multi-Layer perceptron Regressor**
Mean Squared Error	0.01287	0.02795	0.00896	0.00287

Mean Absolute Error	0.092443	0.14581	0.077430	0.03511
Root Mean Squared Error	0.01287	0.027953	0.00896	0.05355
R Squared	0.77926	-0.041701	0.89298	-2.8224

4.5. Comparison of Volatility between the products

Daily volatility of the returns for the period between July 2019 to March 2023 were analyzed for each product yielding the results in Table 6. The NSE 20 Share Index recorded the lowest level of volatility a volatility of 0.689%. The ILAM I-REIT came in second with a volatility of 2.74% followed by the Absa New Gold ETF which recorded a volatility of 4%. The bonds market followed with and average volatility of 6.97% while the derivatives market exhibited the highest level of volatility which averaged at 6850%. This therefore indicates that the derivatives market is the most volatile market.

Table 6: Evaluation of Volatility by Product

Market	**Index**	**Daily Volatility**
Equities	NSE 20 Share Index	0.00689
Derivatives	Absa	1.3582
	BAT	57.043
	EABL	20.99
	KCB	5.50
	NCBA	5.65
	SCBK	15.01
	SCOM	6.401
	NSE 25	342.07
	25 Mini Index	365.50
	COOP	0.88
	Equity	0.813
	IHMP	0.819
	Average for Derivatives	68.50
REITS	ILAM FAHARI I-REIT	0.0274 or 2.74%

ETFs	Absa New Gold ETF	0.04 or 4%
Bonds	IFB	0.082 or 8.2%
	FXD	0.0574 or 5.74%
	Average Bonds Market	6.97%

5. Summary, Conclusion and Suggestion for future Research

5.1.Summary

This research analyzed the predictive power of various machine learning models in forecasting future stock market volatility. Data was used for the prices of the five products available in the Nairobi Securities Exchange namely; Equities, Bonds, Exchange Traded Funds, Real Estate Investment Trusts and Derivatives. The study further sought to establish the volatility of each product and infer on their volatilities especially a comparison between the volatility of traditional and alternative products. Diagnostic tests were conducted on the price indices used in the analysis which the price indices used for the above products were skewed and not following a normal distribution. This however it was noted that their returns were stationary. In making projections of the future volatility six technical indicators were utilized namely; Simple Moving Average, Exponential Moving Average, Relative Strength Index, Rate of price Change, Stochastic Oscillator and Moving Average Convergence Divergence. The predictive power was evaluated using the error metrics namely; Mean Squared Error, Mean Absolute Error and Root Mean Squared Error. Additionally, out of sample analysis was done through the R- squared score. The fitted models included; the random forest model, XGBoost model, Linear Regression and Multi-Layer Perceptron Regressor. The Multi-Layer Perceptron Regressor achieved a good fit achieving low error values, this however it recorded negative values with regards to the R squared score thus not a good model in making predictions. The Linear regression model proved to be the most robust model achieving a combination of both low level of errors as well as high R squared score. In the evaluation of volatilities of the different products, daily volatility of the returns for the period between July 2019 to March 2023 were analyzed for each product. The NSE 20 Share Index recorded the lowest level of volatility a volatility of 0.689%. The ILAM I-REIT came in second with a volatility of 2.74% followed by the Absa New Gold ETF which recorded a volatility of 4%. The bonds market followed with and average volatility of 6.97% while the derivatives market exhibited the highest level of volatility which averaged at 6850%. This therefore indicates that the derivatives market is the most volatile market.

5.2.Conclusion and Areas of further Studies

The study had to main objectives to study and evaluate the predictive power of various machine learning models as well as study and evaluate the volatility of the listed products at the Nairobi Securities Exchange. The stationarity and normality tests indicated presence of leverage effect, leptokurtosis, volatility clustering and persistence. Secondly, the Multi-Layer Perceptron Regressor was found to be a good fit model achieving low level of errors. This however the root mean squared error for this model was negatives thus failing to predict future values accurately. The Linear Regression model was therefore derived as the most robust model as it was able to achieve low errors as well as well as high R squared score.

This research utilized technical indicators in the prediction of future values. Future studies should look into combining both technical and fundamental analysis in their analysis for better results. Secondly, the field of Artificial Intelligence is a fast-evolving area, recent developments have seen the developments of hybrid models which combine different models in a bid to achieve better results. Future researchers focusing on this area should therefore incorporate such models in the analysis.