Forecasting Low Frequency Volatility using a Combination of Econometrics Models and Random Forests

Katlego Aubrey Lentswe (katlego@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Dr. Michael Kateregga
Impact Radius, Forensic South Africa

24 October 2019

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

In this project, we implement three volatility models ARCH(1),GARCH(1,1),EGARC(1,1) and Random Forest machine learning technique to forecast volatility. These models were trained using a stock dataset composed of daily of S&P 500 stock index. The results showed that the GARCH(1,1) outperformed other econometric models. The Random Forest outperformed all econometric models in terms of forecasting volatility.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Katlego Aubrey Lentswe, 24 October 2019

Contents

1. Introduction

1.1 Background

The stock market has a huge influence on the Gross Domestic Product (GDP) of a country, it can use financial parameters and consumer confidence to influence the economy of the country (Adjasi and Biekpe, 2006). Investors and companies use the stock market for different reasons, to make a profit to settle debts, to introduce their products, and to expand their business interests (Berger and Udell, 1998). The stock market is considered by many to be a highly volatile environment, ranging from annual, quarterly and daily swings (Drakopoulou, 2016).

Volatility can be described as the variance overtime or an unequal variance across the time series. Volatility is one of the key features in the financial market, it provides input in portfolio management for investors and companies (Pine and Shmurun, 2007). Financial institutions and other organizations must be informed of current volatility to better manage their assets and estimate their future returns (Allen and Carletti, 2008). Over the years, the volatility of stock returns has been a great concern to an organization, the reason is that they use volatility to measure financial risks (Schwert, 1990). When dealing with derivatives and securities, thorough research must be done on volatility and critical information must be extracted to forecast volatility accurately. A proper investigation on the nature of volatility and an accurate forecasting method informs the investors on how to better manage their portfolios. This has prompted analysts and researchers to further their investigation in different techniques and methods on how to predict volatility.

Abu-Mostafa and Atiya (Abu-Mostafa and Atiya, 1996) reported that various techniques for predicting financial markets by extracting meaningful information have been adopted and implemented over the years. These techniques are fundamental analysis and technical analysis. Fundamental analysis is responsible for in-depth analysis of the stock market whereas technical analysis deals with historical data such as past prices and trade volumes to predict future prices (Poon and Granger, 2003; Pike et al., 1993). When predicting financial indicators, technical analysis is the technique to consider with complex mathematical models (Danıelsson, 2002; Kayacan et al., 2010). However, such complex mathematical models are not easy to use when predicting volatility. Forecasting models are a useful tool for financial institutions to estimate future market trends. There has been a lot of models proposed over the years by researchers, Sjhölm et al. (Sjöholm, 2015) proposed the ARCH model which was one of the first time series models used to model the change in volatility, and it was developed by (Engle, 1982). ARCH can be described as a statistics model used for time series data to describe the variance of the present residual result as a component of the real sizes of the past periods residual result. One of the disadvantages of the ARCH model is that, it can not record the lag variance, an extension of the ARCH model named Generalized ARCH model was proposed and developed by (Bollerslev, 1986). The GARCH model incorporates the ARCH model within it, the variance of the residual term follows the process of the autoregressive moving average. In any case, leverage effects and asymmetric effect of the news cannot be detected by the GARCH model. When the stock price reacts more strongly when influenced by bad news or the good news from a related firm, this can be described as an asymmetric effect. Hence researchers have proposed the extension of a GARCH model with certain specifications, e.g Exponential GARCH (Nelson, 1991) and Glosten-Jagannathan-Runkle GARCH(expressed in a quadratic form) (Glosten et al., 1993). The leverage effect is defined as a measurement to quantify how much business risk a company is currently experiencing. Since GARCH-type models have been frequently used as tools for forecasting volatility, they have become very important in modelling risk analysis and uncertainty in

financial time series.

The econometric models above cannot adapt to changes in stock fluctuation, this makes it difficult to capture patterns in the data. Classification and regression tree models and other machine learning methods have been adopted to capture these patterns. A Random Forest(RF) algorithm is considered to be one of the most powerful machine learning technique that can predict patterns. The RF is used for classification and regression problems. On the regression part, it takes the output of other trees, sums them up, and take the average. In this study, we use the RF algorithm to forecast the volatility of a given stock price.

The essay is structured as follows, chapter 2 is the literature review of previous work done and discussion. Chapter 3 is composed of an understanding of Econometric models is presented, in chapter 4 a brief description of the Random forest technique is presented. In Chapter 5 results of the econometric models and the random forest algorithms are discussed..

1.2 Objective of the study

This study aims to compare three econometric models namely; ARCH, GARCH and EGARCH and a machine learning method known as random forests.

1.3 Motivation of the study

In the financial world to minimize risk and manage portfolios the best technique to forecast volatility must be deployed. Financial institutions and researchers are looking for the best, reliable, cost-effective and easy method to forecast volatility. In this study, we will be evaluating the parameters of the three econometric model and the Random Forest.

2. Literature review

Forecasting volatility accurately is crucial for financial institutions that make capital out of options trading and portfolio management. Models of volatility have been used to forecast the direction of the stock market and assess its nature (Poon and Granger, 2003; Fleming et al., 1995).

Franses and Van Dijk (Franses and Van Dijk, 1996) conducted a study of forecasting volatility of the stock market using non linear GARCH models. An evaluation of the skewness of the stock market indices was described by two GARCH models named Quadratic GARCH (Q-GARCH) and Glosten-Jagannathan-Runkle GARCH (GJR-GARCH). The performance of the two proposed models for forecasting stock market was studied and compared. It was concluded that the Q-GARCH model had shown better results of volatility when compared to EGARCH. Gokcan et al.,(Gokcan, 2000) used non-linear and linear GARCH models to predicted the volatility of upcoming stock market. A comparison study of linear (GARCH(1,1) and non-linear (EGARCH) was performed using month to month financial exchange returns of seven developing nations. Results indicated that the GARCH(1,1) model outperformed EGARCH model when compared.

Regime switching GARCH models study was done to forecast the stock market volatility (Marcucci, 2005). A comparison study of various ordinary GARCH models and Markov Regime-Switching GARCH (MRS-GARCH) was done. The performances of the models were assessed using statistics and risk management analysis. The outcome indicated that the Mark Regime Switch-GARCH model using normal distribution outperformed the ordinary GARCH models and similar Markov regime switching GARCH in forecasting volatility at horizons under statistics and risk management losses. (Luo et al., 2017) proposed a study on forecasting of the stock market volatility and MCS Test. An investigation of both return and variance of the SSE380 index was done to evaluate GARCH, EGARCH and TGARCH models with normal distribution and student's t distribution. The results indicated that using different loss functions of the GARCH model and using Student's t distribution model showed to be the most suitable model for forecasting SSE380 volatility.

Zhang and Pan (Zhang and Pan, 2006) did a study on forecasting the volatility of the Chinese Stock Market. Several other models were used to forecast the stock market volatility of the Shanghai and Shenzhen indices. The following models were tested, moving average model, mean model, random walk model, GARCH model, GJR-GARCH model, EGARCH model and APARCH model under different distributions. Volatility forecasts were evaluated using statistical error measures that included traditional error measures and asymmetry error statistics. Results showed that the GJR-GARCH and EGARCH model outperformed similar GARCH-type models. However, the ARCH and GARCH models are focused on volatility clustering and leptokurtosis, both of cannot record the economics shocks of the stock market namely, the leverage effect, the EGARCH was introduced to address such problems. In recent years Machine Learning technique has been introduced to finance. One of the most important reason why ML and AI were introduced to finance is that finance involves pattern recognition using data, where multifarious inputs are modelled to predict outputs.

Supervised learning is formed by regression and classification which are important components of statistics and machine learning for analysis of data sets or solving complex problems (Zhu and Goldberg, 2009). Recently Artificial Neural Network(ANN) has been used to resolve and identify difficulties in finance. Malliaris and Salchenberger (Malliaris and Salchenberger, 1996) used ANN to forecast the S&P 100 implied volatility. An ANN model was used to forecast volatility using historical volatility and

different options market factors. The results showed that the performance of the ANN demonstrated was impressive.

Hajizadeh et al.,(Hajizadeh et al., 2012) proposed a hybrid model to forecasting volatility of S&P 500 index returns. The performance of the hybrid GARCH models for forecasting the volatility was studied. Two-hybrid models (EGARCH) intertwined with ANN was proposed to forecast the volatility of S&P 500 index. Results obtained showed that the best model was EGARCH (3, 3) enhanced with ANN. (Tseng et al., 2008) proposed an ANN model of a hybrid EGARCH to forecast the volatility of the Taiwan stock index option prices. A combined asymmetric volatility of Artificial Neural Network for option pricing model was proposed. The results showed that Grey-EGARCH volatility outperformed the EGARCH in terms of forecasting volatility. A hybrid model for forecasting volatility was integrated with LSTM and multiple GARCH-type models was proposed by (Kim and Won, 2018). A novel hybrid LSTM model which combined different standard GARCH model to forecast the volatility of the stock price was examined. The results indicated that the EGARCH model indicated to be better at forecasting volatility among the GARCH models.

There are currently few studies on applying Random Forest(RF) in finance. A study of forecasting the direction of stock market prices using random forest was done by (Khaidem et al., 2016). A new method was proposed to limit the risk of investment in the stock market by forecasting the returns of a stock using machine learning algorithms called ensemble learning. The results indicated that by using the non-linear nature of the problem and linear discriminant type machine learning algorithms the model performed better.

The purpose of this project is to compare ARCH, GARCH and EGARCH and Random Forest methods in forecasting volatility.

In this section, we discussed the literature review of different GARCH models such as ARCH, QGARCH, RS GARCH. There has been limited work done on comparison of ARCH, GARCH, EGARCH model and Random Forest.

3. Econometric Models

3.1 ARCH(p)

Auto Regressive Conditional Heteroscadasticity is a statistical model proposed by (Engle, 1982). ARCH model is used to describe the volatility of the variance in a time series. ARCH model uses previous returns to forecast volatility, the returns was described by JingLi (Li, 2016) in the lecture nodes as the log return series. The advantage is that, the series is closely related to Taylor series expansion which can describe the change of the price overtime. Let P_t be the current asset price, P_{t-i} be the historical price and r_t the return can be defined as:-

$$r_t = \frac{P_t - P_{t-i}}{P_{t-i}} \approx log\left(\frac{P_t}{P_{t-i}}\right). \tag{3.1.1}$$

The mean of returns of the financial assets is very close to 0. ARCH(1) is the simplest model for modelling volatility, the model for returns given by:-

$$r_t = \sigma_t \epsilon_t \tag{3.1.2}$$

ϵ_t is noise with zero mean and variance σ_t^2, ϵ_t follows the normal distribution and it is defined as:-

$$\epsilon_t \sim \textbf{Noise}(0, \sigma_t^2) \tag{3.1.3}$$

σ_t is the standard deviation and it is defined as:-

$$\sigma_t = \sqrt{\omega + \alpha_i r_{t-i}^2} \tag{3.1.4}$$

and ω is the constant coefficient. The proof of ARCH(1) follows the properties of normal distribution which is $N(0, 1)$. The conditional mean of r_t is given as:-

$$\begin{aligned} E(r_t|r_{t-1}, r_{t-2}, ...) &= E(\sigma_t \epsilon_t|r_{t-1}, r_{t-2}, ...) && (3.1.5)\\ &= \sigma_t E(\epsilon_t|r_{t-1}, r_{t-2}, ...) && (3.1.6)\\ &= \sigma_t * 0 && (3.1.7)\\ &= 0 && (3.1.8) \end{aligned}$$

The conditional variance of r_t given as:-

$$\begin{aligned} var(r_t|r_{t-1}, r_{t-2}, ...) &= E(r_t^2|r_{t-1}, r_{t-2}, ...) && (3.1.9)\\ &= E(\sigma_t^2 \epsilon_t^2|r_{t-1}, r_{t-2}, ...) && (3.1.10)\\ &= \sigma_t^2 E(\epsilon_t^2|r_{t-1}, r_{t-2}, ...) && (3.1.11)\\ &= \sigma_t^2 t * 1 && (3.1.12)\\ &= \sigma_t^2 && (3.1.13) \end{aligned}$$

The unconditional variance is given as:-

$$var(r_t) = E(r_t^2) - [E(r_t)]^2 = E(r_t^2) \tag{3.1.14}$$
$$= E[E(r_t^2|r_{t-1}, r_{t-2}, ..)] \tag{3.1.15}$$
$$= E[\omega + \alpha_i r_{t-i}^2] \tag{3.1.16}$$
$$= \omega + \alpha_i E[r_{t-i}^2] \tag{3.1.17}$$
$$= E(r_t^2) = \frac{\omega}{1-\alpha_i} \tag{3.1.18}$$

if the following condition holds $0 < \alpha_i < 1$. Hence the general formula ARCH(p) is:-

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 \tag{3.1.19}$$

The ARCH(p) model like any other model, it has its disadvantages, it was found that the positive and negative shocks were different in terms of effects on volatility. Another disadvantage is that the ARCH(p) model could over predict the volatility when the model response is slow on the large isolated shock of return series (Matei, 2009).

3.2 GARCH(p,q)

(Bollerslev, 1986) proposed the GARCH (Generalized Auto Regressive Conditional Heteroscedasticity) model which is an extension of the ARCH model, the difference is that it considers both the lag residuals(p) and lag variance(q). The GARCH can be defined as:-

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{3.2.1}$$

whereby σ_{t-i}^2 is the historical variance at particular time series and β_j is a constant and must be greater than zero. The unconditional variance of the GARCH model is given by:-

$$var(r_t) = \frac{\omega}{1-\alpha_i+\beta_i} \tag{3.2.2}$$

if the following condition holds, $0 < \alpha_i + \beta_i < 1$.

3.3 EGARCH(p,q)

The GARCH model is better in terms of estimating the returns and volatility, but if you want to estimate the leverage effect the EGARCH is the effective model to use. Exponential Generalized AutoRegressive

Conditional (EGARCH) Heteroscadasticity was proposed by (Nelson, 1991). EGARCH can be defined as:-

$$\log(\sigma_t^2) = \omega + \sum_{i=1}^{p} \alpha_i g(Z_{t-i}) + \sum_{j=1}^{q} \beta_j \log(\sigma_{t-j}^2) \tag{3.3.1}$$

$$Z_t = \frac{\epsilon_t}{\sigma_t} \tag{3.3.2}$$

whereby $g(Z_t) = \theta Z_t + \lambda[|Z_t| - E|Z_t|]$, θ and λ are constant coefficients, Z_t is the standard normal variable. If $\theta > 0$, this implies that the negative shocks from the past have a strong influence on the present conditional volatility as compared to positive shocks from the past. If θ and $|Z_t| - E|Z_t| > 0$ there is an increase of conditional variance, and when $|Z_t| - E|Z_t| < 0$ there will be a decrease of conditional variance.

3.4 Implementing econometrics model algorithm in python

Below is step by step process of building the algorithm, the code used to implement these econometric models was done by (Sheppard, 2019).

Algorithm 1 Econometrics model algorithm

Input: Log returns

Output: Estimated volatility

1: Import all the necessary libraries of python.
2: Calculate the log-returns using adjusted close price.
3: Determine or allocate the parameters(p,q distribution, mean, dependent variable and volatility model) of the econometric models.
4: Fit your model
5: Display the estimated volatility and a summary of your parameters.
6: Forecast for 5 horizons.

4. Decision Trees and Random Forests (RF)

4.1 Decision Trees

A Decision Tree (DT) is a flow chart diagram that is used to determine an outcome or show a statistical probability. Every single branch of the decision tree represents a possible decision, outcome, or reaction. A DT is displayed as a sequence of steps that produce an effective and easier method to visualize and understand the potential outcomes of a decision and its range of possible outcomes. When DT is constructed it involves the dividing a set of training data, which is split into similar subsets based on conditions of the feature values.

4.1.1 Procedure of decision tree. Algorithm steps to build a decision tree is given as follows, (1) Assign the data sample to the root node, (2) divide the data samples into subsets and choose the features and threshold value, (3) when a feature has the best information gain it is set as a criterion node, (4) divide again, choose the node which has the best value or condition and (5) if the node cannot be further divided assign it as a leaf node. The definition of Root node, inner node and leaf node can be found at the appendices section were found in Wikipedia. Below is a figure of a decision tree.

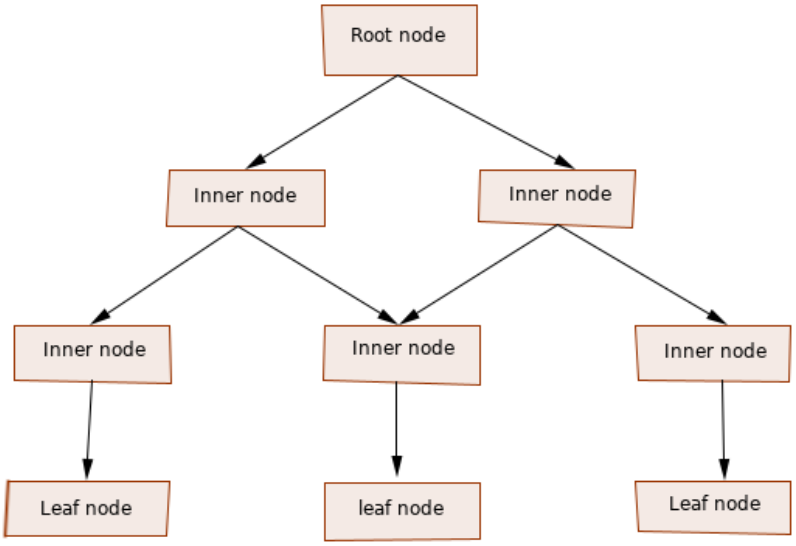


Figure 4.1

An example that can be solved using a decision tree, suppose you want to buy a stock currently and sell it in the future to gain capital, the feature that you will use to make your decision is the volatility. You have two options, (1)Sell the stock or (2) Hold the stock.

- If the volatility is greater or equal to the specified condition, the leaf node should display (1) Sell the stock.
- If the volatility is less than or equal to the specified condition the leaf node should display (2) Hold the stock.

4.1.2 Advantages. It is easier to use Decision trees when manipulating data for processing, another advantage is that normalizing or standardizing data is not required in decision trees. Missing data points does not influence the process of constructing a decision tree. (Dietterich, 2000).

4.1.3 Disdvantages. A minor change in the data set can cause instability in a decision tree and calculation can become extremely complicated when compared to other algorithms. Decision trees requires more time to be trained.

4.2 Random Forests

A random forest can be described as a group or a collection of decision trees that uses the "majority vote" concept to produce an outcome or action. Random forest uses bootstrap aggregating (Bagging) techniques or ensemble learning to reduce the variance of the prediction function. The figure 4.2 below represents the random forest and was done by Kumar (Kumar, 2018)

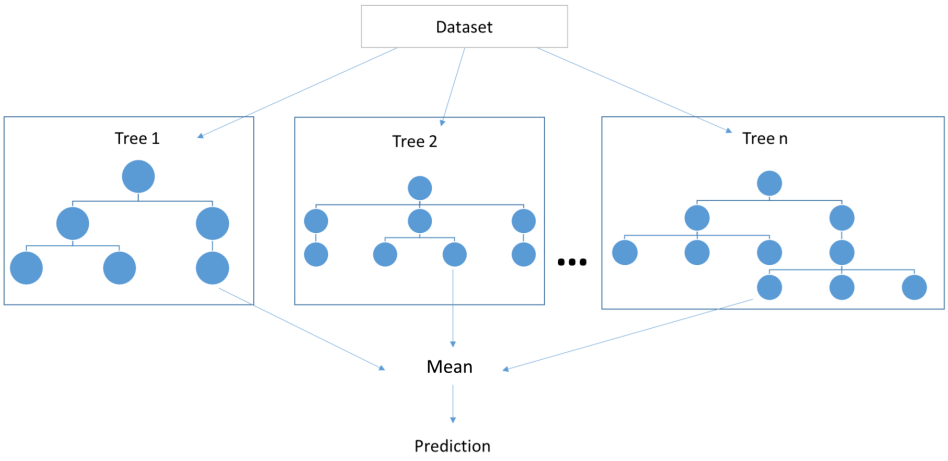


Figure 4.2

4.2.1 Random forest pseudocode. A pseudocode for random forest is arranged in this manner:-

- Select the number of features in a dataset randomly using bootstrap mechanism.
- For every tree x in the forest, a bootstrap sample data from the dataset is selected.
- A subset of features is selected randomly for possible feature split.
- If a feature cannot split, that will be the end results of that tree.

4.2.2 Implementing Random Forests algorithm in python 2. below is step by step process of building the algorithm, the code for the Random Forest regressor was done by (Breiman, 2001a)

Algorithm 2 Random Forest algorithm

Input: Historical Volume and adjusted close

Output: Estimated volatility

1: Import all the necessary libraries of python
2: Preprocessing of stock dataset
3: The next step is to prepare the data for training by dividing the dataset into features and labels. The data will also be divided into training and testing
4: Scale the data into a proper format for training and testing.
5: Fit your training and testing dataset into the random forest model
6: Lastly evaluate your algorithm by using metrics such as mean square error and root mean squared error

5. Performance measurement

Measurement and evaluation of the performance of a model is very important when forecasting. Zhang et al. (Zhang et al., 2014) recommended that when measuring the forecast performance of a model, the Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE) are the most popular methods to use. MSE is mostly used for bias, precision and accuracy problems. In this study MSE is used to monitor the precision and accuracy of the Random Forest algorithm. The equation of the MSE is defined as:-

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y_t})^2 \tag{5.0.1}$$

whereby n is the number of observations, y_t is the actual value and $\bar{y_t}$ is the predicted value at time t. A small value in the outcome or results of the MSE indicates a model has a good forecast model, meaning that, the smaller value the better the forecast model.

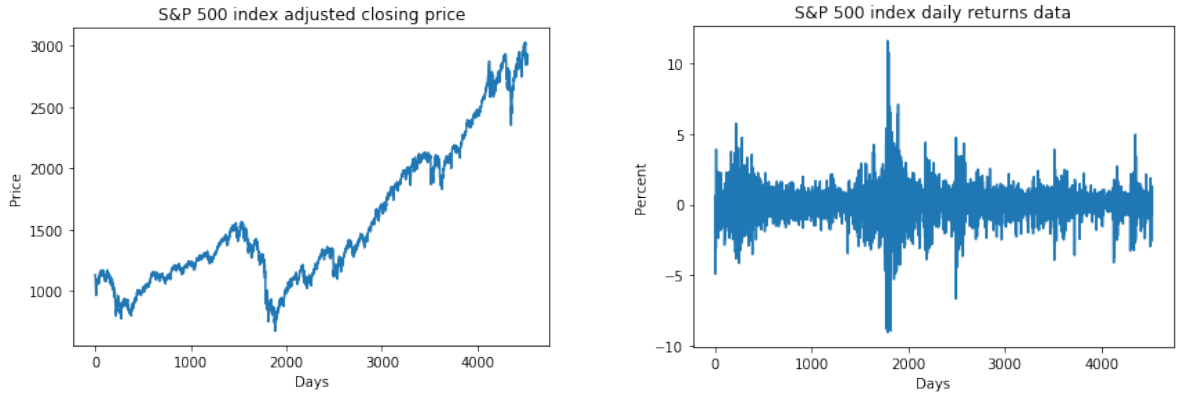
Another method prediction performance is evaluated by statistical metrics, Root Mean Square Error (RMSE) and mean absolute error (MAE). The mathematical expressions of these criteria methods are described by the following equations below

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \bar{y_t}| \tag{5.0.2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y_t})^2} \tag{5.0.3}$$

6. Data description

In this section a brief description of S&P 500 index is explained. The S&P(Standard and Poor) 500 index can be described as a measurement method used to measure the stock performance of 500 well-established companies in the United States of America (Denis et al., 2003). Figures 6.1 below represent plots of the closing price and the return using 3.1.1. When observing the closing price graph between days 1000 and 2000, the price of the stock was increasing at a constant rate, then suddenly it dropped massively unexpectedly. This was caused by a terrorist attack, whereby an American Airlines Flight crashed on to the world trading centre in New York. After the terrorist attack, the price of S&P 500 index increased at a steady pace rapidly. When observing the log data returns 6.1b, S&P 500 index was highly volatile and shows that it risky to invest in it or on the other hand it is an opportunity to gain high returns, furthermore there is the presence of volatility clustering. The negative shocks of the returns has more influence on the volatility than the positive shocks. Figure 6.1 has the effect as figure 6.1b but it has been expressed with a stock price graph and you can fully observe the market drift.



(a) Price

(b) Return

Figure 6.1: S&P 500 index

The dataset of S&P 500 index is composed with 6 columns (Opening, Close, Low, High, Adjusted closing and the Volume) and 4526 rows. The historical dataset that was used was acquired in from Yahoo finance (finance), the time period that was used was from 2001-09-04 to 2019-09-03.The historical dataset was divided into two parts 75% for training and 25% was used for the evaluation of Random Forest. This is an important step when you splitting data for training and testing to determine your features and label. This section aims to select features that will help to predict volatility for Random Forest. The heat map (figure 6.2) below indicates how variables correlate to each other. Open, High, Low and Close are stock prices, the squared returns are used as a proxy of volatility, and the proxy is represented as Var in the heat map. Open, High, Low and Close stock prices do not affect the proxy volatility of the market.

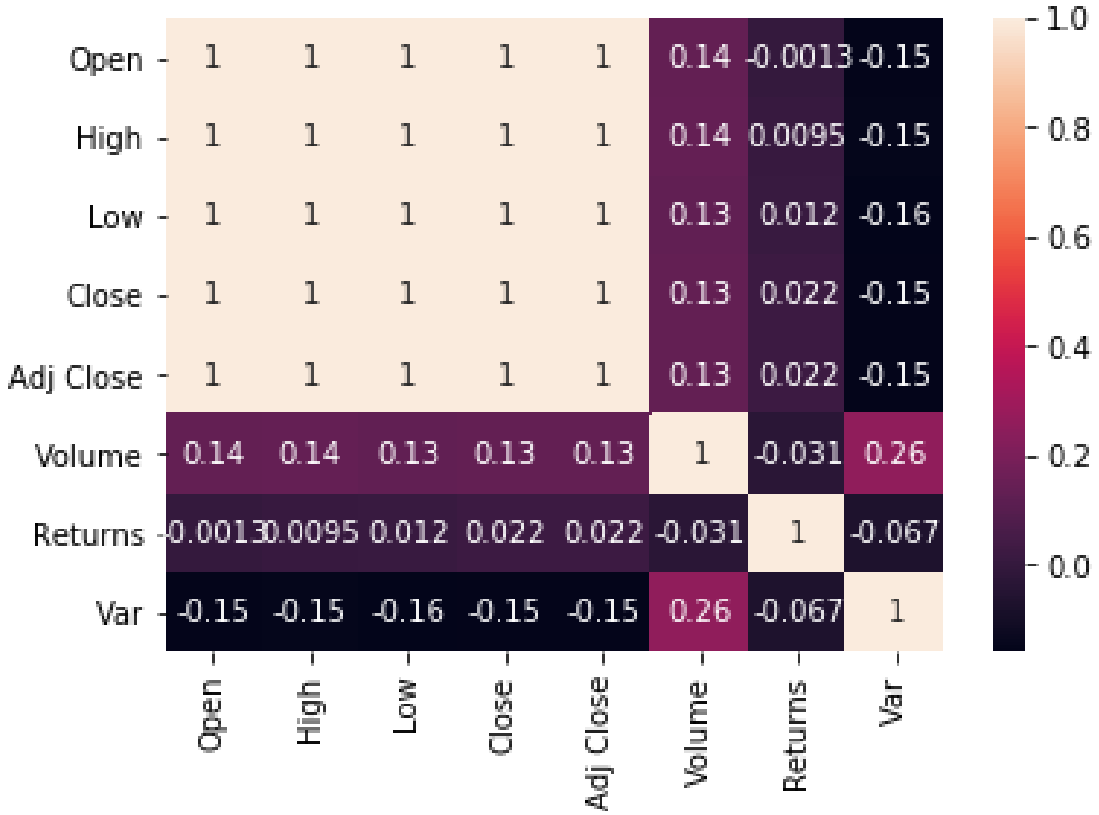


Figure 6.2: Heat map

6.1 Data manipulation

The RF, ARCH, GARCH and EGARCH algorithms were implemented using a programming language called python 2. Packages that were used for data extraction and data visualization were pandas, seaborn and arch model. The following criterion and variables were used to implement the models, for ARCH(p) model, p was chosen as one, the distribution that was chosen as the normal distribution, the dependent variable was the returns 3.1.1. For the GARCH(p,q) and EGARCH(p,q) model the same parameters that were chosen for ARCH(p) model were exploited but with an addition of q which was 1. p=1 and q=1 were chosen because they are the most convenient and easier methods to work with. The RF parameters were chosen differently from the econometrics models, the RF uses features and label to estimate or forecast. The Volume was chosen for features because they have a relationship with the proxy volatility and the proxy volatility(variance) was chosen as the label. 75 % of the dataset was chosen for training and 25 % for testing.

7. Results and Discussion

In this section, we analyse ARCH(1,1), GARCH(1,1), EGARCH(1,1) and RF model, using the performance measurement we identify which model performed better in estimate volatility .

7.1 Econometric models

7.1.1 ARCH(1,1), GARCH(1,1) and EGARCH(1,1) . The table below shows the results of ARCH(1,1) model, the method of estimation that was chosen was Maximum likelihood(MLE), because MLE usually has lower variance compared to other methods and on large variety of estimation problems it can be modified. Nagilo (Ngailo, 2011) concluded that when the coefficients of the volatility are less than one, it indicates that the overall volatility is low and when the coefficients are greater than one it indicates high volatility. The standard errors are of the models are close to zero meaning that we can assume that the coefficients are precise according to (Brooks and Gelman, 1998). Since our confidence interval is 95 % we can assume that the significant level is 0.05 which is 5%, we can reject the null hypothesis of non-stationary returns and deduce that the returns data is stationary. From table 7.1,7.2 and 7.4, since the p-value is less than the significant level for the ARCH(1,1), GARCH(1,1) and EGARCH(1,1), we can conclude that coefficients are significant. All the p-values in the table are significant, we can conclude that the parameters are significant and best to fit for the model.

	ARCH(1)				
	Coefficients	**St err**	**t**	**p>|t|**	**95.0% Conf. Int.**
ω	0.9262	6.106e-02	15.168	5.725e-52	[0.807, 1.046]
α_1	0.3738	6.399e-02	5.842	5.167e-09	[0.248, 0.499]
μ	0.0387	1.892e-02	2.048	4.060e-02	[1.657e-03,7.581e-02]

Table 7.1: ARCH table

	GARCH(1,1)				
	Coefficients	**St err**	**t**	**p>|t|**	**95.0% Conf. Int.**
ω	0.0211	4.901e-03	4.303	1.687e-05	[1.148e-02,3.069e-02]
α_1	0.1104	1.331e-02	8.294	1.091e-16	[8.433e-02, 0.137]
β_1	0.8707	1.396e-02	62.361	0.000	[0.843, 0.898]
μ	0.0594	1.149e-02	5.176	2.269e-07	[3.694e-02,8.196e-02]

Table 7.2: GARCH

	EGARCH(1,1)				
	Coefficients	**St err**	**t**	**p>|t|**	**95.0% Conf. Int.**
ω	7.7392e-03	3.345e-03	2.314	2.067e-02	[1.184e-03,1.429e-02]
α_1	0.2259	2.432e-02	9.286	1.595e-20	[0.178, 0.274]
β_1	0.9761	5.274e-03	185.102	0.000	[0.966, 0.986]
μ	0.0702	1.168e-02	6.008	1.877e-09	[4.729e-02,9.307e-02]

Table 7.3: EGARCH

The table 5.4 below shows the results of measures used to forecast the accuracy of the model, the measures that were used are, the MSE, Mean Absolute Error(MAE) and RMSE. (Bowerman et al., 2005) concluded that when evaluating and analysing models, the smaller the error the better accuracy. In terms of Mean Square error, the GARCH(1,1) outperformed all the econometric model that were evaluated in this study followed by ARCH(1) and EGARCH(1,1). For the Mean Square Error, the EGARCH(1,1) performed better other models, on the Root Mean Square Error GARCH(1,1) came out top compared to other models. The overall analysis and evaluation showed that the GARCH(1,1) model is the best model to forest volatility. According to (Bollerslev et al., 1992) GARCH(1,1) is considered to be adequate to model volatility of long time periods.

	ARCH(1)	**GARCH(1,1)**	**EGARCH(1,1)**
MSE	21.8473888	20.3866581	20.66736
MAE	1.5698184	1.4021370	1.410133
RMSE	0.011621	0.0115776	0.011168

Table 7.4: Performance to volatility forecasting

Horizon	1	2	3	4	6
Mean	0.038733	0.038733	0.038733	0.038733	0.038733
Volatility	1.126024	1.347192	1.429874	1.460783	1.472338

Table 7.5: 5-steps ahead volatility forecast using ARCH(1)

The conditional volatility and the conditional mean of S&P 500 index were computed using a built-in function in python 2 called **forest**. Table 5.5, 5.6 and 5.7 presents 5 days ahead forest of the conditional mean of the three models, the conditional mean 0.000458 for ARCH(1), 0.000612 for GARCH(1,1) and 0.00738 for EGARCH(1,1). It can be observed that the conditional mean of all the models is closer to 0, which implies that for financial assets, the expected value of log returns is closer to zero. The EGARCH(1,1) model could not forecast anything greater than a day, because it can not capture volatility properly.

Horizon	1	2	3	4	5
Mean	0.059448	0.059448	0.059448	0.059448	0.059448
Volatitliy	1.279075	1.276019	1.273022	1.27008	1.267194

Table 7.6: 5-steps ahead volatility forecast using GARCH(1,1)

Horizon	1
Mean	0.07018
Volatility	1.33816

Table 7.7: 5-steps ahead volatility forecast using EGARCH(1,1)

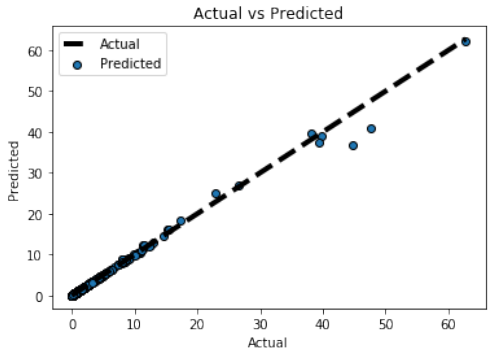
	Model Selection		
	ARCH(1,1)	**GARCH(1,1)**	**EGARCH(1,1)**
AIC	13771.9	12045.7	12091.7
BIC	13791.1	12071.3	12117.4

Table 7.8: Model selection table

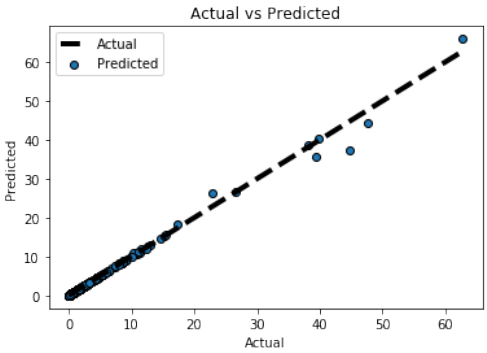
Looking at the table 7.8 above, AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) is for model selection to forecast or estimate volatility (Acquah, 2010). AIC is used to find a model that can best approximate unseen data, the BIC similar to AIC but it uses the Bayesian approach. A smaller value of AIC and BIC indicates that the model supports the data, and it is considered to be the best model compared to others. The results showed that the GARCH(1,1) showed to be the best model when compared with ARCH(1) and EGARCH(1,1) model.

7.2 Random Forests

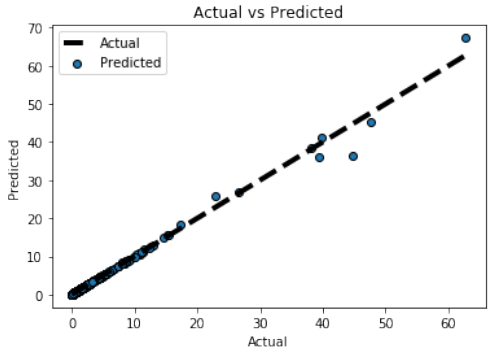
In figure 7.1 below we compare the actuals and the predicted values using different numbers of trees. When observing the graphs, the random forest is performing very well with minimum errors. Looking at the graphs it is quite clear that analysing these graphs via observation is not adequate to come to a solid conclusion without considering other methods of analysis. The table 7.9 below shows a performance analysis of the Random Forest using performance measurements techniques such as MSE, MAE and RMSE. Generally the increase of the number of trees suggests that the will be fewer errors and the Random Forest will perform better, but in essence this approach will lead to overfitting, the best option to exploit is optimization.



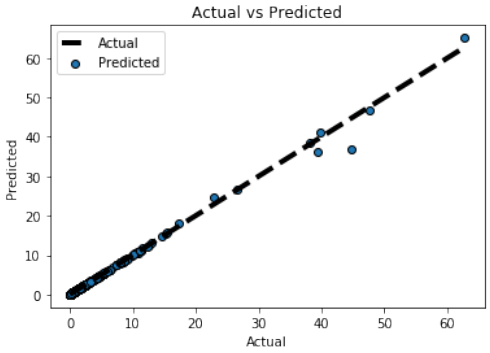
(a) 10 trees



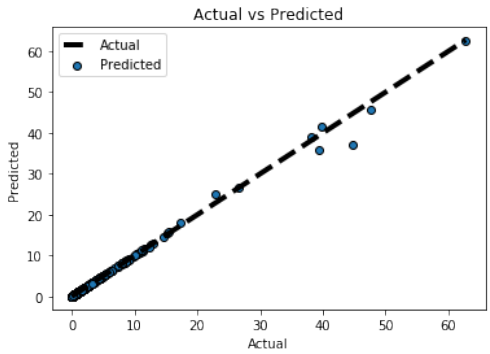
(b) 50 trees



(c) 100 trees



(d) 200 trees



(e) 500 trees

Figure 7.1: Comparison of number of trees

Number of trees	10	50	100	200	500
MSE	0.112774	0.105444	0.09212	0.07541	0.0734330
MAE	0.0312256	0.0278845	0.0282915	0.023363	0.022877
RMSE	0.33581	0.32472	0.30350	0.27460	0.27099

Table 7.9: Performance analysis

7.3 Comparison

Since the GARCH(1,1) model outperformed the other econometric models, in this section we will be comparing the performance analysis of the GARCH(1,1) model and the Random Forest. The table 7.10 below shows the comparison of the GARCH(1,1) and the Random Forest. Random Forest indicates that it has smaller values of MSE, MAE and RMSE, meaning it has minor errors compared to the GARCH(1,1). It is evident that Random Forest performed better estimating at estimating the volatility than the GARCH(1,1).

	MSE	MAE	RMSE
GARCH(1,1)	20.386658	1.4021370	4.51515869
RF	0.09184	0.02673	0.30192

Table 7.10: Comparison of Models

8. Conclusion and Future Work

In this thesis, we compare three econometric models and the random forest to determine which technique is more accurate at forecasting volatility of a stock market. The returns were calculated and we examined the distribution of the returns. After fitting the return data to the econometric models, the results indicated that these models were exceptional at estimating volatility. The MSE, MAE and RMSE were used to select the best model among the three, the results indicated the GARCH model had fewer errors compared to other models. Another method that was used is the information criteria (AIC and BIC), the results showed that GARCH(1,1) had low AIC and BIC meaning that it was the best model among the other models. For the Random forest results, features and label were selected by using the heat map used for training and testing, it was discovered that selection of features affects the performance of the model. In overall the Random Forest was more accurate than the GARCH(1,1) model. For future work, we can study econometric models with longer lags and also other GARCH models. We should also study other machine learning models in future like the neural network to forecast volatility.

Acknowledgements

First and far most I would like to thank God for giving me strength to finish this thesis. I appreciate and I'm very thankful to AIMS funders, without your contributions this project would not be possible. I am grateful to my supervisor, Dr. Michael Kateregga for the guidance, discussions and introducing me to this topic, I would also like to thank Prof. Ronnie Becker for his inputs and suggestions. I would like to thank the AIMS leadership and staff particularly the academic director Dr. Simukai Utete for her support and encouragement. I would like to expand my gratitude to Prof. Jeff Sanders for his 'talks' that he gave throughout the year, they were very helpful. I would also like to show my gratitude to the tutors Emilia, Reem and Rock who guided me on completing this thesis. Furthermore, I would like also thank my AIMS colleagues who contributed directly and indirectly with their fruitful discussions on this project. Lastly I would like to thank my family especially my parents Mr MA Lentswe and Mrs DM Lentswe (late) for laying out a solid educational foundation in my life.

Appendices

AppendixA. Definitions

Concept	Description:
Heteroskedasticity	An unequal or change of variance in a time series (Hayes and Cai, 2007) .
Leverage effect	It is a strong correlation between the stock and the returns (Ait-Sahalia et al., 2013).
MCS(Model Confidence Set) test	It is a group of models when given a level of confidence it produces the best model (Luo et al., 2017).
Student-t	Is a family member of probabilistic distribution which is used widely for statistical analyses (Zhu and Galbraith, 2010).
Implied volatility	It is the market volatility (Fleming, 1998).
LSTM	Long Short Term Memory is a Recurrent Neural Network that solves tasks that cannot be solved by previous learning algorithms (Gers et al., 1999).
S&P 500 index	Is a list of top five hundred companies in the USA influence the stock market of the USA Denis et al. (2003).
Node	is a point in the network were point interact or branch (wikipedia, 2001).
Root node	Is the topmost node in the tree that has links to one or more nodes or decision tree (wikipedia, 2001).
Inner node	Is a node that can be split into subsets node (wikipedia, 2001).
Leaf node	is a node that cannot be split further or an outcome of a decision tree (wikipedia, 2001).
Splitting	Is a node that can be classified into two or more sub nodes (wikipedia, 2001).
Over-fitting	When a model depends on the historical data to a point that it is unable to make good predictions on its own or when the model follows the noise pattern or error (Domingos, 2012).
Feature	it is an attribute which its characteristics can be observed (Hall, 1999).
Bootstrap	It is a statistical method that uses random sampling with replacement (Efron and Tibshirani, 1997).
Bagging	Is an ensemble machine algorithm that is used to improve the accuracy and stability of a learning algorithm (Dietterich, 2000).

References

Abu-Mostafa, Y. S. and Atiya, A. F. Introduction to financial forecasting. *Applied Intelligence*, 6(3): 205–213, 1996.

Acquah, H. D.-G. Comparison of akaike information criterion (aic) and bayesian information criterion (bic) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics*, 2(1):001–006, 2010.

Adjasi, C. K. and Biekpe, N. B. Stock market development and economic growth: The case of selected african countries. *African Development Review*, 18(1):144–161, 2006.

Ait-Sahalia, Y., Fan, J., and Li, Y. The leverage effect puzzle: Disentangling sources of bias at high frequency. *Journal of Financial Economics*, 109(1):224–249, 2013.

Alberg, D., Shalit, H., and Yosef, R. Estimating stock market volatility using asymmetric garch models. *Applied Financial Economics*, 18(15):1201–1208, 2008.

Allen, F. and Carletti, E. The role of liquidity in financial crises. *Available at SSRN 1268367*, 2008.

Bauwens, L., Preminger, A., and Rombouts, J. V. Regime switching garch models. *Available at SSRN 914144*, 2006.

Berger, A. N. and Udell, G. F. The economics of small business finance: The roles of private equity and debt markets in the financial growth cycle. *Journal of banking & finance*, 22(6-8):613–673, 1998.

Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3): 307–327, 1986.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. Arch modeling in finance: A review of the theory and empirical evidence. *Journal of econometrics*, 52(1-2):5–59, 1992.

Bowerman, B. L., O'Connell, R. T., and Koehler, A. B. *Forecasting, time series, and regression: An applied approach*. Thomson Brooks/Cole, 2005.

Breiman, L. RandomForestRegressor random forest code description, (accessed: 2019-9-21), 2001a. URL https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/modules/generated/sklearn.ensemble.RandomForestRegressor.

Breiman, L. Random forests. *Machine learning*, 45(1):5–32, 2001b.

Brooks, S. P. and Gelman, A. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

Cheng, W.-H. and Hung, J.-C. Skewness and leptokurtosis in garch-typed var estimation of petroleum and metal asset returns. *Journal of Empirical Finance*, 18(1):160–173, 2011.

Danıelsson, J. The emperor has no clothes: Limits to risk modelling. *Journal of Banking & Finance*, 26(7):1273–1296, 2002.

Denis, D. K., McConnell, J. J., Ovtchinnikov, A. V., and Yu, Y. S&p 500 index additions and earnings expectations. *The Journal of Finance*, 58(5):1821–1840, 2003.

Dietterich, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000.

Domingos, P. M. A few useful things to know about machine learning. *Commun. acm*, 55(10):78–87, 2012.

Drakopoulou, V. A review of fundamental and technical stock analysis techniques. 2016.

Dwipa, N. M. S. Glosten jagannathan runkle-generalized autoregressive conditional heteroscedastics (gjr-garch) methode for value at risk (var) forecasting. *Proceeding of ICMSE*, 3(1):M–63, 2016.

Efron, B. and Tibshirani, R. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

Engle, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

finance, Y. SP 500 financial data. URL https://finance.yahoo.com/quote/%5EGSPC?p=^GSPC.

Fleming, J. The quality of market volatility forecasts implied by s&p 100 index option prices. *Journal of empirical finance*, 5(4):317–345, 1998.

Fleming, J., Ostdiek, B., and Whaley, R. E. Predicting stock market volatility: A new measure. *Journal of Futures Markets*, 15(3):265–302, 1995.

Franses, P. H. and Van Dijk, D. Forecasting stock market volatility using (non-linear) garch models. *Journal of Forecasting*, 15(3):229–235, 1996.

Gers, F. A., Schmidhuber, J., and Cummins, F. Learning to forget: Continual prediction with lstm. 1999.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance*, 48(5):1779–1801, 1993.

Gokcan, S. Forecasting volatility of emerging stock markets: linear versus non-linear garch models. *Journal of forecasting*, 19(6):499–504, 2000.

Hajizadeh, E., Seifi, A., Zarandi, M. F., and Turksen, I. A hybrid modeling approach for forecasting the volatility of s&p 500 index return. *Expert Systems with Applications*, 39(1):431–436, 2012.

Hall, M. A. Correlation-based feature selection for machine learning. 1999.

Hamid, S. A. and Iqbal, Z. Using neural networks for forecasting volatility of s&p 500 index futures prices. *Journal of Business Research*, 57(10):1116–1125, 2004.

Hayes, A. F. and Cai, L. Using heteroskedasticity-consistent standard error estimators in ols regression: An introduction and software implementation. *Behavior research methods*, 39(4):709–722, 2007.

Hazen, T. L. Volatility and market inefficiency: A commentary on the effects of options, futures, and risk arbitrage on the stock market. *Wash. & Lee L. Rev.*, 44:789, 1987.

Hibbert, A. M., Daigler, R. T., and Dupoyet, B. A behavioral explanation for the negative asymmetric return–volatility relation. *Journal of Banking & Finance*, 32(10):2254–2266, 2008.

Im, J. and Jensen, J. R. A change detection model based on neighborhood correlation image analysis and decision tree classification. *Remote Sensing of Environment*, 99(3):326–340, 2005.

Jorion, P. Predicting volatility in the foreign exchange market. *The Journal of Finance*, 50(2):507–528, 1995.

Kayacan, E., Ulutas, B., and Kaynak, O. Grey system theory-based models in time series prediction. *Expert systems with applications*, 37(2):1784–1789, 2010.

Khaidem, L., Saha, S., and Dey, S. R. Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*, 2016.

Kim, H. Y. and Won, C. H. Forecasting the volatility of stock price index: A hybrid model integrating lstm with multiple garch-type models. *Expert Systems with Applications*, 103:25–37, 2018.

Kim, K.-j. and Han, I. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert systems with Applications*, 19(2):125–132, 2000.

Kumar, V. Random forests and decision trees from scratch in python random forest description, (accessed: 2019-9-15), 2018. URL https://towardsdatascience.com/random-forests-and-decision-trees-from-scratch-in-python-3e4fa5ae4249.

Laurent, S. Analytical derivates of the aparch model. *Computational Economics*, 24(1):51–57, 2004.

Li, J. Lecture 5: (g)arch models slide, January 2016.

Luo, L., Pairote, S., and Chatpatanasiri, R. Garch-type forecasting models for volatility of stock market and mcs test. *Communications in Statistics-Simulation and Computation*, 46(7):5303–5312, 2017.

Malliaris, M. and Salchenberger, L. Using neural networks to forecast the s&p 100 implied volatility. *Neurocomputing*, 10(2):183–195, 1996.

Marcucci, J. Forecasting stock market volatility with regime-switching garch models. *Studies in Nonlinear Dynamics & Econometrics*, 9(4), 2005.

Matei, M. Assessing volatility forecasting models: why garch models take the lead. *Romanian Journal of Economic Forecasting*, 12(4):42–65, 2009.

Nelson, D. B. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, pages 347–370, 1991.

Ngailo, E. Modelling and forecasting using time series garch models: An application of tanzania inflation rate data. *Unpublished Master's Thesis). University of Dar es Salaam*, 2011.

Pike, R., Meerjanssen, J., and Chadwick, L. The appraisal of ordinary shares by investment analysts in the uk and germany. *Accounting and Business Research*, 23(92):489–499, 1993.

Pine, A. and Shmurun, A. L. System and method for analyzing financial market data, May 1 2007. US Patent 7,212,997.

Poon, S.-H. and Granger, C. W. Forecasting volatility in financial markets: A review. *Journal of economic literature*, 41(2):478–539, 2003.

Schwert, G. W. Stock market volatility. *Financial analysts journal*, 46(3):23–34, 1990.

Sheppard, K. arch documentation arch code description, (accessed: 2019-08-30), 2019. URL https://readthedocs.org/projects/arch/downloads/pdf/latest/.

Sjöholm, S. Heteroscedasticity models and their forecasting performance, 2015.

Tseng, C.-H., Cheng, S.-T., Wang, Y.-H., and Peng, J.-T. Artificial neural network model of the hybrid egarch volatility of the taiwan stock index option prices. *Physica A: Statistical Mechanics and its Applications*, 387(13):3192–3200, 2008.

wikipedia. Node (computer science) computer science definitions, (accessed: 2019-9-21), 2001. URL https://en.wikipedia.org/wiki/Node_(computer_science).

Wu, J. Threshold garch model: Theory and application. *The University of Western Ontario*, 2010.

Zhang, G. P. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.

Zhang, X., Zhang, T., Young, A. A., and Li, X. Applications and comparisons of four time series models in epidemiological surveillance data. *PLoS One*, 9(2):e88075, 2014.

Zhang, Z. and Pan, H. Forecasting financial volatility: Evidence from chinese stock market. 2006.

Zhu, D. and Galbraith, J. W. A generalized asymmetric student-t distribution with application to financial econometrics. *Journal of Econometrics*, 157(2):297–305, 2010.

Zhu, X. and Goldberg, A. B. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.