

**Forecasting Indexes Volatilities by using Machine Learning Techniques, Econometric and Random Models**

A study on the forecasting capacity prediction under each model on the first days of the Ukraine’s Conflict

Francisco Gonçalves Cruces Matos Bettencourt

Dissertation report presented as partial requirement for obtaining the Master’s degree in Statistics and Information Management

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**  
Universidade Nova de Lisboa

**Forecasting Indexes Volatilities by using Machine Learning techniques, econometric and random models**A study on the forecasting capacity prediction under each model on the first days of the Ukraine’s Conflict

by

Francisco Gonçalves Cruces Matos Bettencourt

Dissertation report presented as partial requirement for obtaining the Master’s degree in Statistics and Information Management , with a specialization in Risk Analysis

**Advisor / Co Advisor:**

**Co Advisor***:*

October 2022

Dedication (optional)

NOT YET REQUIRED TO BE COMPLETED ONCE THE DISSERTATION IS FULLY COMPLETED

Acknowledgements (optional)

NOT YET REQUIRED TO BE COMPLETED ONCE THE DISSERTATION IS FULLY COMPLETED

Abstract

Predicting the volatility on returns for a stock index is an attractive and defying task in the field of Machine Learning (ML). The comparison of Machine Learning models, and their resulting predictions, with several Time Series algorithms and Monte Carlo simulations, could provide valuable insight regarding the advantage of using more recent Machine Learning methods to predict stock index volatility. In this paper, it is presented a study on the ability of various models to predict for four worldwide Indexes, the returns and therefore, their volatilities, on the beginning of the Ukraine’s conflict, by applying and comparing them in order to prove if recent machine learning models, could bring a better capacity to predict or not, than old models or just basic fundamental and accountant analysis.

By acknowledging the recent macroeconomic variables and factors, that inflation is rising, probably powered by the Ukraine’s conflict, there is a direct relationship between the increase of Gas and food prices due to the lack of production and difficulties on the supply chain, it is crucial to understand how markets worldwide are behaving regarding these factors. Therefore, this study will be based on four indexes, namely the Euronext 100 (Europe), the National Stock Exchange India (Asia), the São Paulo Stock Exchange (South America) and the NASDAQ (North America).

The main goal is to compare the predicted volatility under each model, for each region/Index, with the actual return, and observe which model performed better at forecasting the actual value. The study and forecasting of volatility are of high value, since Pension/Investment funds, as well as other stakeholders in Financial Markets, recognize that the risk should be minimized to the maximum level, and be within the standards that Pension/Fund members agreed upon.

Being that Data driven solutions are in the core of any business today, it is important to understand how we can obtain, prepare, analyze, and get different outcomes based on the same data being analyzed. With this being said, the main focus on this project will not be on trying to obtain the most accurate model to predict the daily volatility, but to compare how different models predict and if their predictions fall very far from one another. As well, by using multivariable models, allows for other type of information beside the return and return distribution to be considered, and this could equate in higher accuracy capacity. In this study will be use not only High frequency variables, based on the same timeline as the returns, i.e., daily, but will be also used low frequency variables that could be monthly, quarterly or even annual.

Keywords

Stock Index Volatility Prediction; Ukraine’s/Russia stock volatility forecasting; Machine Learning Volatility forecasting; Monte Carlo Stock; GARCH; GARCH-MIDAS;LSTM Stock; Transformer Stock; Worldwide volatility comparison.

INDEX

[1. Introduction 1](#_Toc114072317)

[2. Literature review 2](#_Toc114072318)

[2.1. Risk and Return Metrics 2](#_Toc114072319)

[2.2. Data Analysis in Financial Terms 3](#_Toc114072320)

[2.3. Models Accuracy and Prediction Capacity on timeseries Financial Data 5](#_Toc114072321)

[2.4. Econometric Models 5](#_Toc114072322)

[2.5. Monte Carlo Simulation 7](#_Toc114072323)

[2.6. Machine learning Models 8](#_Toc114072324)

[2.7. Accuracy Measurement Models 14](#_Toc114072325)

[2.8. Diebold-Mariano Test 15](#_Toc114072326)

[2.9. Similar Papers and results 15](#_Toc114072327)

[3. Methodology 17](#_Toc114072328)

[3.1. Research Approach and Design Strategy 17](#_Toc114072329)

[3.2. Data Collection 17](#_Toc114072330)

[3.3. Data preparation 18](#_Toc114072331)

[3.4. Results Obtention 19](#_Toc114072332)

[4. Results and discussion 22](#_Toc114072333)

[5. Conclusions 23](#_Toc114072334)

[6. Limitations and recommendations for future works 24](#_Toc114072335)

[7. Bibliography 25](#_Toc114072336)

[8. Appendix (optional) 27](#_Toc114072337)

[9. Annexes (optional) 28](#_Toc114072338)

List of Figures

[*Figure 1- Structure of Financial System* (Sharma F. C., 2019) 2](#_Toc107680518)

[Figure 2- Returns Example (Hull J. C., 2018) 6](#_Toc107680519)

[Figure 3- General flowchart used for model selection (Xu & Goodacre, 2018) 9](#_Toc107680520)

[Figure 4- Relation between Error and Model Index (Awad & Khanna, 2015). 14](#_Toc107680521)

[Figure 5- Support Vector Regression example (Awad & Khanna, 2015) 14](#_Toc107680522)

[Figure 6- LSTM process (Houdt, Mosquera, & Nápoles, 2020) 16](#_Toc107680523)

[Figure 7- Example of CART Machine (Sharma & Kumar, 2015) 17](#_Toc107680524)

[Figure 8- Linear Regression Example (Ambrosius, 2007) 19](#_Toc107680525)

[Figure 9- The Overall Research Process 22](#_Toc107680526)

[Figure 10- Data used by model 23](#_Toc107680527)

[Figure 11- Step Breakdown by Model Type 24](#_Toc107680528)

[Figure 12- Training Set Accuracy Error Example 24](#_Toc107680529)

[Figure 13- Validation Set Accuracy Error Example 25](#_Toc107680530)

[Figure 14- Test Set Accuracy Error Example 25](#_Toc107680531)

[Figure 15- Example of Best and worst model by Index and by Data splitting 25](#_Toc107680532)

LIST OF Equations

[*Equation 1- Fair Value of Assets* 3](#_Toc107680485)

[*Equation 2- Share value of individual Stock* 3](#_Toc107680486)

[*Equation 3- Dividend Payment* 4](#_Toc107680487)

[*Equation 4- Expected Return* 5](#_Toc107680488)

[*Equation 5- Variance* 6](#_Toc107680489)

[*Equation 6-Standard Deviation* 6](#_Toc107680490)

[*Equation 7- Sharpe ratio* 7](#_Toc107680491)

[*Equation 8- GARCH* 11](#_Toc107680492)

[*Equation 9- Long Run Variance* 11](#_Toc107680493)

[*Equation 10- EWMA* 11](#_Toc107680494)

[*Equation 11- ARIMA* 11](#_Toc107680495)

[*Equation 12- Monte Carlo Simulation* 12](#_Toc107680496)

[*Equation 13- Random Normal Distribution* 12](#_Toc107680497)

[*Equation 14- Wiener Process* 12](#_Toc107680498)

[*Equation 15-Spot Price at time t* 12](#_Toc107680499)

[*Equation 16- Support Vector Regression Hypothesis* 13](#_Toc107680500)

[*Equation 17- SVR Error* 13](#_Toc107680501)

[*Equation 18- Support Vector Regression* 14](#_Toc107680502)

[*Equation 19- Support Vector Regression Augmented* 14](#_Toc107680503)

[*Equation 20-LSTM Block Input* 15](#_Toc107680504)

[*Equation 21-LSTM Input Gate* 15](#_Toc107680505)

[*Equation 22-LSTM Forget Gate* 15](#_Toc107680506)

[*Equation 23-LSTM Cell* 16](#_Toc107680507)

[*Equation 24-LSTM Output Gate* 16](#_Toc107680508)

[*Equation 25- LSTM Block Output* 16](#_Toc107680509)

[*Equation 26- LSTM Logistic Sigmoid* 16](#_Toc107680510)

[*Equation 27- LSTM Hyperbolic Tangent* 16](#_Toc107680511)

[*Equation 28- CART Gini Index* 17](#_Toc107680512)

[*Equation 29- Linear Regression* 18](#_Toc107680513)

[*Equation 30- Mean Absolute Error* 19](#_Toc107680514)

[*Equation 31- Mean Squared Error* 20](#_Toc107680515)

[*Equation 32- Root Mean Squared Error* 20](#_Toc107680516)

[*Equation 33- Logarithmic Return* 23](#_Toc107680517)

List of Abbreviations and Acronyms

NOT YET REQUIRED TO BE COMPLETED ONCE THE DISSERTATION IS FULLY COMPLETED

# Introduction

Forecasting of financial assets has always been a vital topic in finance, given that the ability to overperform the market, and therefore, break the market efficiency theory, could generate huge profits to those who would be able to do it, as it is corroborated by Poon and Granger (2001). From the beginning of the history of the stock market and trading, the evolution of technology has narrowed the gap to a reliable future value prediction. Nowadays, with the widespread use of Machine Learning algorithms and Auto Regressive models, and due to the recent computational power increase and ease of access, big funds and banks are trying to get to the perfect prediction, in a way that would help them have larger profits and also a better understanding of the risk they are facing in the market.

At each day it is possible to realize that accessing real live data from the markets is getting easier and it that, also the number of models, statistics, risk metrics is increasing. This means, that even the small investment funds or the single investor that likes to go on the markets by himself, is being able to have reliable and worth trusty information on a daily basis, that consequentially allows him to have a better understanding of the risk and profit opportunities that the same is exposed to, such as it is described by Ma, Xiong and Feng (2021). Has it’s doable to see, in the previous few years a lot of small investors, especially those that have a background in Finance and Engineering have started to trade on their own, using and creating machine learning algorithms that allow them to sometimes, even outperformed big Investment funds and the NASDAQ in terms of returns for example.

Nevertheless, and even acknowledging that returns are one of the most important factors that weights on the investors investment decision, as it is explained by Chaudhuri and Koo (2001), is also important to understand that different investors have a different risk profile, and even if some are willing to undertake a significative risk on the longer and shorter time, others are not, and due to this, is really important that all the stakeholders on the process, have a clear view in which are the levels of risk they are exposed to, and if this level is the one-to-go level for the Investor.

Nowadays, is also remarkably important to acknowledge the weight and influence that some Externalities and Macro environment factors have on the investment decision. Social and environmental awareness are increasing on a really fast pace, which lead sometimes to big market, not expected, movements, such as for example, the Ukraine/Russia Crisis, which is leading to an unprecedent disinvestment on Russian Companies and assets. Not either the best risk metrics can predict what could be the impact of such conflict for the world economy, as well to the impact on the European Central Bank/American Federal Bank, even if there is no direct exposure to financial Russian Assets, for example oil prices have increase meaning that Energy companies within these Indexes could be facing relatively bigger market movements than what was expected, (Engelhardt, Ekkenga, & Posch, 2021).

With all that being said and recognizing that a fundamental basis is always really important to be able to understand and calculate the expected volatility for a given period of time, in this paper, the main goal is to obtain a prediction as accurate as possible of these Indexes volatilities. To do so, data will be trained and tested for a series of models, and compared their outputs. Some of the models, such as Time Series forecasting models rely solely on the past values of the Indexes prices, whilst others, such as Machine Learning algorithms, allow for additional information to be considered when forecasting future values.

# Literature review

## Risk and Return Metrics

By recognizing that markets seek efficiency and that different investors at different stages of their life’s seek different risk profiles, there were studies performed on this field, with the goal to achieve the most efficient allocation of capital, and in this way achieve a better relation between risk and return. If an analysis is solely made on a single stock, there should not exist the need for efficiency since the risk and return will be given by the intrinsic distribution moments of that same stock. Since a portfolio combines a big number of different stocks, with each having different means and standard deviations, a combination between the weight allocated to each stock will lead to more or less efficient portfolios.

Has it is stated by Myles Mangram(2013), when he referrers to the theory that Harry Markowitz presented on his doctoral dissertation, the most important factor for the risk of a portfolio is given by the individual risk that each security represents and the way the Securities within the same portfolio correlate with each other, meaning, if the volatility in one increases how will the other securities be affected. This works as the base for the principal of diversification on a Portfolio, meaning that a portfolio manager, or an Index for a more concrete and related example, should not only assess the risk in each single stock that constitutes their portfolio, but also, in the correlation between themselves.

The returns for a given stock/index should be given by,

Equation 1- Daily return

In which P represents the price, and i represents the day of the Price.

The mean return is given by dividing the daily returns by the number of days N, i.e,

Equation 2- Mean Return

Volatility is the same as the standard deviation of the returns, given by:

Equation 3- Volatility

In the concrete case of this paper, it should not be covered in depth an analysis of each single stock represented on the selected indexes and the way that these same stocks are correlated, since the focus will be more on predict the volatility of the actual portfolio, and not trying to obtain the most efficient one. By dividing the Expected Return by the Volatility, an Investment Manager is able to understand how many units of return he is obtaining for unit of risk taken. The biggest this value, the more efficient the portfolio, since this will equate into higher returns with lower levels of volatility.

The Sharpe ratio, as it was presented by William Sharpe, allows an investment manager to compare how many units of return he is obtaining over the Risk-Free rate for a given level of volatility as given by:

*Equation 4- Sharpe ratio*

Within the Capital Allocation Line, the point with Highest Sharpe ratio, will work the tangent point for the Capital Market Line and it is also notable to understand the difference between Systematic and Unsystematic risk.

In the book “Capital market Theory: An Overview on Corporate Finance”,(Ross, Westerfield, & Jaffe, 2002), systematic risk is described as being a macro-level form of risk that affects a large number of assets to one degree or another, such as inflation and interest rates, that virtually affect all securities, and cannot be eliminated. Unsystematic risk, on the other hand, is a micro-level form of risk that specifically affects a single asset, or a narrow group of assets (Volatility), (Ross, Westerfield, & Jaffe, 2002)**.**

The topics above will not be covered in detail during this dissertation paper, but they are substantial empirical knowledge that may help the reader to better understand why investment managers and stock indexes use different combinations of allocations to different stocks, and how that can affect the overall performance and key risk metrics of an Index/Portfolio. Without going further on the methodology applied to this paper, is crucial to understand that any type of index even being a combination of stocks, will be assessed as it was one stock, so the correlation between stocks within the portfolio, will not, once again, be a subject of further detail.

## Data Analysis in Financial Terms

Since the beginning of the 21st century, data driven companies and data driven business models have been one of the most profitable. As such, and defining data as an individual set of facts, statistics, and information, that is fitter for a deep analysis and allows to achieve conclusions from it, sometimes, and by using predictive methods, it allows data managers and data scientists to achieve a high level of accuracy when predicting future outcomes. It this being said, is expected that some type of information that exists on the Financial Markets, with the help of this same predictive methods, could be used by investment managers in order to take decisions.

There are usually two types of data, qualitative and quantitative, being that for the majority of the predictive models in Finance they use quantitative variables, since these ones are easier to model and also, easier to obtain, sometimes it is also a key factor to use qualitative variables since these ones, represent the mindset of the global market, and could have a really big impact on future prices, volatility, trends and so on (Wong, Chin, & Tan, 2016). Despite the fact that qualitative variables may impact the future price of assets, and as a direct consequence the return and volatility of the same assets, the study on this qualitative variables and the actual impact they have is still vague, in the sense that there are not yet many models that have performed within the expected level of accuracy when trying to predict the actual impact (Guo, Shi, & Tu, 2017).

Due to the complexity of this models and the lack of scientific evidence to corroborate their actual impact on the target variable, the use of sentimental analysis models will not be covered in this dissertation, and the main focus will be on the quantitative variables that in fact may or may not, depending on concrete cases, impact the target variable. In the concrete case of this paper, the data to be used across all models are the actual prices of the selected indexes, since they are the base for return calculation and consequently for volatility as well.

When using a predictive method for forecasting it is always necessary to split the data set into 2 sets, namely the Training and the Test set, as it is stated by Yun Xu and Royston Goodacre in their paper “On Splitting Training and Validation Set: A Comparative Study of Cross‑Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning” (2018). The Training set consists of building the model with multiple model parameter settings and then each trained model is challenged with the validation set (not to be used)**.** The Test set is the last set of data , that should be a set with new data that was never considered when drafting the model, and the actual accuracy of the model on this set, will determine the actual prediction capacity of the same.

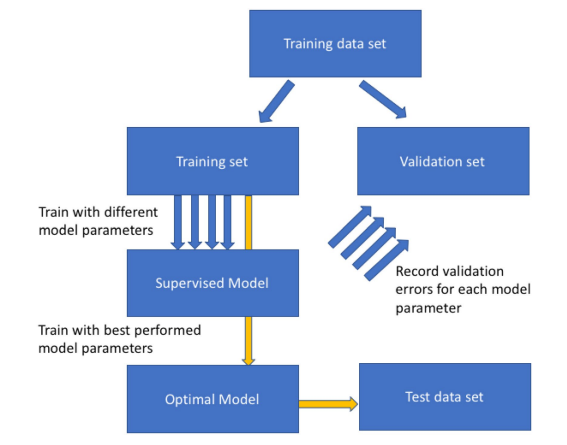


Figure 1- General flowchart used for model selection (Xu & Goodacre, 2018)

By considering prices of the Indexes, as the main source of data for the model, it is also important to denote that these ones are sequential, meaning that the order in which they are presented on the data set affects the outcome of the model. In this type o situations, and basing that the data should be ordered by day, i.e., the first observation should be the day of the first Price used, is expected that the training set should be the older prices, and the test set the most recent ones, being that the main objective of the dissertation is to corroborate which of the models used, if any, could help to predict volatility in the future, and in this way help Investment Managers and small investors to be more aware of the risk they are facing.

## Models Accuracy and Prediction Capacity on timeseries Financial Data

There are different types of models, that have different types of assumptions that can be used to predict target variables. Some models, require less information, i.e., they only need to be supplied with stock data, or the main Key Statistics that are based on the stock data, whilst some other models, may require a bit more information, in order to also produce, what could equate into a more accurate result. Since this type of data is a Time Series data, what can also be understand as a collection of values obtained from sequential measurements over time (Esling & Agon, 2012),in this dissertation, was decided to split the models that can predict a time series in three.

First, the econometric models, that are models that are able to describe the application of statistical methods to the quantification and critical assessment of hypothetical relationships using data (Dougherty, 2016).Second, randomized models, i.e., models that based on given assumptions of the overall distribution, will randomly provide values for the target variable. One of the most famous is the Monte Carlo Simulation based on a Geometric Brownian Motion. At last, machine learning algorithms will be used, these ones can be based on the actual price of stocks, i.e., they will account every single observation in the model, they can be based solely on the distribution moments of all observations combined, or they can use multiple variables and their key statistics in order to predict the target variable. Since the models described above are predictive models, could be therefore assumed, that these ones are able to predict values that could be accurate or highly inaccurate. Therefore, accuracy models can be used to fairly compare the accuracy capacity between models. The models that are better explaining relationships between variables/assumptions used, should be the ones with a higher accuracy rate.

## Econometric Models

Based on the book “Handbook of Financial Time Series” (Andersen, Davis, Kreiss, & Mikosh, 2009)**,** it is possible to denote that the most well-known Econometric models to be used are the Generalized Autoregressive Conditional Heteroskedasticity (GARCH). In this work it will be presented two type of GARCH Models, namely the GARCH, which is univariable model, that works directly with the returns and their distribution parameters, and also the GARCH-MIDAS, in which MIDAS stands for Mixed Data Sampling, is based on the same ‘basis’ as the other one, but have the peculiarity of allowing low frequency variables to provide thoughtful insights regarding the parameters (Conrad, Custovic, & Ghysels, 2018).

The GARCH model, is a model for the variance of a time series. Despite their capacity to predict long run volatility, they actually tend to perform a more accurate prediction result, when accessing short term volatility. In GARCH, is calculated based on a long-run average variance rate, , as well from and . A given weight is attributed to each of these variables, which means that this is a weighted model. The objective is to Maximize , which is the weight of , by changing the allocations between, weight given to and , weight given to (Hull J. C., 2018).

*Equation 5- GARCH*

The weight allocation function is provided by:

*Equation 6- Long Run Variance*

Regarding the technical details of GARCH-MIDAS models, in which the conditional variance is multiplicatively decomposed into a short-term (High frequency) and a long term (low-frequency) component. The long-term component is the returns of each Index, and the low-frequency are the quarterly/annual explanatory variables (Engle, Ghysels, & Sohn, 2013).

For the return calculation, denotes the quarterly frequency, and , denotes the number of days in the quarter (for simplification terms, with will be 252/4, i.e., 63 business days).

Equation 7- Return for GARCH-MIDAS

The conditional mean of returns is constant, i.e.,

Equation 8- Mean Return on GARCH-MIDAS

The innovation is assumed to be independent and identical distributed with mean 0 and variance one. and denote the short- and long-term component of the conditional variance, respectively.

The short-term component, varies at the daily frequency and follows a unit-variance GARCH(1,1) process, i.e.,

Equation 9- Short-term Component in GARCH-MIDAS

Where >0, 0, and 1. The Long-term Component varies at the quarterly frequency and is given by,

Equation 10-Long-term Component in GARCH-MIDAS

Where denotes the explanatory variable and a certain weighting scheme.

For this case the weighting scheme to be used will be the Beta weighting Scheme, which is given by,

Equation 11- Beta Weighting scheme

Engle et al., define that GARCH-MIDAS is estimated by quasi-maximum likelihood and construct heteroscedasticity and autocorrelation consistent (HAC) standard errors.

All the methodology above is benchmarked on the Engle et al. article and provide thoughtful insights regarding this approach to mixed data sampling variables, in order to forecast volatility.

## Monte Carlo Simulation

By acknowledging that the return of prices follow a given distribution, in this case, a normal distribution, it may be assumed that generating random variables, for the target variable, should not be totally random. A Geometric Brownian motion is often used to explain the movement of time series variables and, when adapted to corporate finance, explains the movement of asset Prices (Reddy & Clinton, 2016), in this concrete case, a Stock Market Index. Since volatility of an asset is measured by its returns, which are based on the logarithmic difference between the price of an asset in a day and the day immediately before that, it may be assumed that the returns distribution for the long term follows an uncertain distribution (random walk) , that will probably be approximately normal within a width range of samples.

(Sengupta, 2004) states that for the Geometric Brownian assumption to be effective regarding modeling stock price, or Index price, in a time series, the following conditions must be verified:

* The underlying asset must be continuous into time and value.
* A stock must follow a Markov process, meaning that only the current stock price is relevant for predicting future prices.
* The proportional return of a stock is Log-Normally distributed
* The continuously compounded return for a stock is normally distributed.

*Equation 12- Monte Carlo Simulation*

Regarding the formula, it is made up of two parts, the first one being a certain component and the second one an uncertain or variable component. The first part is called the drift of the stock and it is assumed as the return that a stock will earn over a short period of time. The uncertain component represents a stochastic process that includes the annual volatility of returns on an Index, and also a Wiener Process which is the Stochastic component (Reddy & Clinton, 2016). For each random number generated from a normal distribution, and this distribution is used due to the fact that returns are normally distributed, the Wiener process consists of the multiplication of this random number by the square root of time, which in turn creates the stochastic process.

When it comes to a Monte Carlo simulation, it is a process that consists in simulating values, for a given variable, n times, in order to predict the most probabilistic outcome, i.e., the one that appears the most times within the simulation. When applying the Monte Carlo simulation to the Geometric Brownian Motion, it should be applied the drift value and the annual volatility, being this one the daily volatility times the square root of 252 business days (Brewer, Feng, & Kwan, 2012)**.** By using a Monte Carlo Simulation, it is possible to generate a Price for a given day, and from that price calculate the return and volatility.

The formula is breakdown in three steps:

*Equation 13- Random Normal Distribution*

*Equation 14- Wiener Process*

*Equation 15-Spot Price at time t*

Where:

* is given by a random normal distribution, with number of simulations, and assuming that mean is 0 and standard deviation is 1.
* is described as the Wiener process and is given by multiplying the square root of time by the variable.
* (spot price at time t) is given by multiplying stock price at time 0 by the base of natural logarithm () raised to the power of the log normal distribution, i.e., drift () – ½\*variance, multiplied by time, plus standard deviation multiplied by the Wiener Process.

## Machine learning Models

**Support Vector Regression**, that is similar to Support Vector Machine, offers a principled approach to machine learning problems because of its mathematical foundation in statistical learning theory. SVM constructs its solution in terms of a subset of the training input and has been extensively used for classification, regression, novelty detection tasks, and feature reduction (Awad & Khanna, 2015). Vapnik-Chervonenkis (VC) theory proves that a VC bound on the risk exists. VC is a measure of the complexity of the hypothesis space. The VC dimension of a hypothesis H relates to the maximum number of points that can be shattered by H. H shatters n points, if H correctly separates all the positive instances from the negative ones. In other words, the VC capacity is equal to the number of training points n that the model can separate into 2n different labels. This capacity is related to the amount of training data available (Awad & Khanna, 2015).Based on the above, the VC dimension a h affects the generalization error, as it is bounded by where is the weight vector of separating hyperplane and the radius of the smallest sphere that contains all the training points, according to:

*Equation 16- Support Vector Regression Hypothesis*

The overall error of a machine learning model consists of:

*Equation 17- SVR Error*

Where is the training error, and is the generalization error.

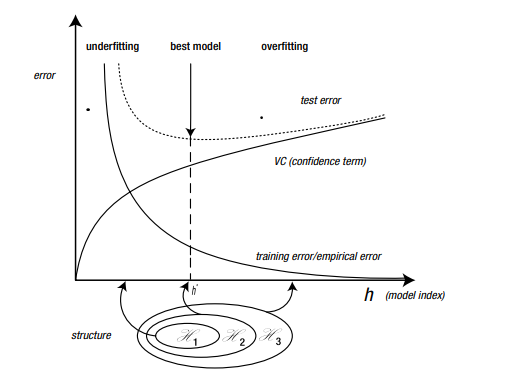


Figure 2- Relation between Error and Model Index (Awad & Khanna, 2015).

Bearing this in mind, the actual difference between SVM and SVR, is that the regression problem is a generalization of the classification problem, in which the model returns a continuous-valued output, as opposed to an output from a finite set (Awad & Khanna, 2015).

For a SVR the formula should be:

*Equation 18- Support Vector Regression*

Or by augmenting by one and include in the vector, it is possible to obtain:

*Equation 19- Support Vector Regression Augmented*

Uma imagem com texto, mapa, céu

Descrição gerada automaticamente

Figure 3- Support Vector Regression example (Awad & Khanna, 2015)

Other model often used in timeseries data is the **Long-Short Term Memory, LSTM**, which is a recurrent neural network. Recurrent or very deep neural networks are difficult to train, as they often suffer from the exploding/vanishing gradient problem (Houdt, Mosquera, & Nápoles, 2020). Overall, this can be prevented by using a “Constant Error Carousel” (CEC), which maintains the error signal within each unit’s cell. The input gate and output gate, form the memory cell. The self-recurrent connections indicate the feedback with a lag of one-time step. A plain vanilla LSTM unit is composed of a cell, an input gate, an output gate and a forget gate, that allows the network to reset is state. In short, the architecture of a LSTM model, is based in a set of recurrently connected sub-networks, also known as, memory blocks. The main function of this blocks is to maintain its state over time and regulate the information flow through non-linear gating units (Houdt, Mosquera, & Nápoles, 2020).

**Block input**- this step is devoted to updating the block input component, which combines the current inputs and the output of that LSTM unit in the last iteration.

*Equation 20-LSTM Block Input*

Where, and are the weights associated with and respectively, whilst represents the bias weight vector.

**Input Gate**- it combines the current input , the output of that LSTM unit and the cell value, in the last iteration.

*Equation 21-LSTM Input Gate*

Where denotes the point-wise multiplication of two vectors, are the weights provided to respectively, whilst represent the bias vector of the component.

**Forget Gate**- The LSTM unit determines which information should be removed from its previous cell states . Therefore, the activation values, , of the forget gates at time step , are calculated based on the current input , the outputs , and the state of the memory cells ate previous time step , and is the bias terms of the forget gates.

*Equation 22-LSTM Forget Gate*

Where denotes the point-wise multiplication of two vectors, are the weights provided to respectively.

**Cell**- this step computes the cell value, which combines the block input , the input gate and the forget gate , with the previous cell value.

*Equation 23-LSTM Cell*

**Output Gate**- is a combination of the current input , the output of that LSTM unit and the cell value in the last iteration.

*Equation 24-LSTM Output Gate*

Where denotes the point-wise multiplication of two vectors, are the weights provided to respectively, whilst represent the bias of the weight vector.

**Block Output**- combines the current cell value with the current output gate.

*Equation 25- LSTM Block Output*

Where in the steps above, denote point-wise non-linear activation functions.   
The logistic Sigmoid is used as a gate activation funct1ion,

*Equation 26- LSTM Logistic Sigmoid*

While the hyperbolic tangent is often used as the block input and output activation function.

*Equation 27- LSTM Hyperbolic Tangent*

All the process above described, as well as all formulas were based solely on the research performed under the publication article “A review on the long short-term memory model” (Houdt, Mosquera, & Nápoles, 2020).

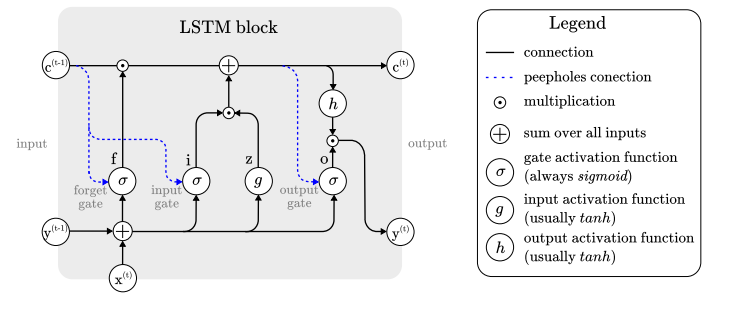


Figure 4- LSTM process (Houdt, Mosquera, & Nápoles, 2020)

Finally, the last model to be used is going to be the **Transformer**, being this one also a recurrent neural network, and it’s proven to be successfully applied in the field of neural language processing. In contrast to the LSTM model, Transformer models do not incorporate recurrence in their structure, they in fact are based on a multi-head attention mechanism and positional embeddings in order to forecast time series. In this paper, a transformer model will be combined with the GARCH model, because it is believed predict with higher degree of accuracy the level of volatility. By using a rolling window methodology, the model uses a fixed sample length for fitting the model and, then the following step is forecasted. In this paper, the window size is **XXXXXX** and the forecast horizon is 1, meaning that it would be used the XXXX previous days, to forecast the next day volatility (Ramos-Peréz, Alonso-González, & Núñez-Velázquez, 2021).



Equation 28- Rolling Window Methodology (Ramos-Peréz, Alonso-González, & Núñez-Velázquez, 2021)

The Transformer will take in consideration not exogenous or endogenous variables but will be based on an **FIGARCH Model and on a GARCH algorithms**. GARCH model was explained above, so FIGARCH algorithm is given by,

Equation 29-FIGARCH

Where L is the lag-operator and d the fractional differencing parameter.

As previously stated, Transformer layers were developed for Neural language processing purposes thus, some modifications are needed in order to apply this layer for volatility forecasting purposes.



Figure 5-Transformer and Multi-head attention Mechanism (Ramos-Peréz, Alonso-González, & Núñez-Velázquez, 2021)

The two main components used by these layers in order to deal with time series are the following:

**Positional Encoder**- Since Transformer Layers have no recurrence structure the information about the relative position of the observations within the time series needs to be included in the model. The positional encoder that will be used on this paper changes based on the Lag, but remains the same across different explanatory variables, as provided by,

Equation 30 - Positional Encoder

Where is the position of the observation within the time series and is the maximum lag.

**Multi-Head attention**- this one is composed of several scaled dot-product attention units running in parallel. Scaled dot-product attention is given by,

Equation 31- multi-head attention

Where Q,K,V are input matrices and the number of input variables taken into consideration.

## Accuracy Measurement Models

Since all the models above described, are used to make a prediction, i.e., based on a multitude of assumptions, these models will predict a value for the Target Variable, it could be acknowledged that they will sometimes be right, and sometimes wrong. If a model has around 95% accuracy on the training data, this could mean that this same model is overfitted to the training data set and will not perform so well on the test data set. Bearing this mind, the main focus of the models below described, is to explain what the actual difference between the real output and the output generated by each model and present it as an average of the model capacity to predict. With this being said, the three accuracy models to be used on this dissertation are the following:

**Mean absolute error-** it involves summing the magnitudes (absolute values) of the errors in order to obtain the total error, and the dividing it by (Willmott & Matsuura, 2005).This measures the absolute average difference between the real data and the predicted data, but it usually tends to fail to punish large errors in prediction.

*Equation 32- Mean Absolute Error*

Where, is the number of observations, is the output generated from the model, is the actual, observed value and is the absolute error.

**Mean Squared error-** This one, is really similar to the one above, but since with will square absolute error, the geometric difference between both observations will be emphasized.

*Equation 33- Mean Squared Error*

Where, is the number of observations, is the output generated from the model, is the actual, observed value and is the absolute error.

**Root Mean Squared error**- also very similar to the one above, this one is able to explain the second moment of the error distribution, i.e., the standard deviation of the error.

*Equation 34- Root Mean Squared Error*

Where, is the number of observations, is the output generated from the model, is the actual, observed value and is the absolute error.

## Diebold-Mariano Test

The Diebold-Mariano test can discriminate the significant differences of forecasting accuracy between different models based on the scheme of quantitative analysis. It works on a Hypothesis basis and allows to check if the results are statistically significant or not for a forecasted series (Constantini & Knust, 2011). By assuming S as a sample of T, and where denotes the prediction errors for the two rival forecasts g, is given by,

Equation 35- Diebold-Mariano Test

And the predictor error between the two forecasts is some type of function, for example, for the MSE,

Equation 36-Diebold-Mariano for MSE

## Similar Papers and results

This dissertation tends to compare the volatility predictive capacity, across four Worldwide indexes, by using a panoply of different models. Previous papers have also followed a similar approach, sometimes only by focusing in one type of model, such as Machine learning for example, or by using a multitude of models. With this subsection of the literature review, it is intended to show the main results/outcomes, that those papers had, and with this have enough empirical knowledge to compare them with the results from this paper. It is also needed to understand that the comparisons should not be straight forward, as there are a multitude of factors that impact the outcomes, either by the Macro environment factors, by the country, date that was performed or even by the model used. With this being said, the following papers should work as a comparison basis for the expected outcomes:

* “Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models” (Kim & Won, 2018), is a paper based on the KOSPI 200 stock index (South Korea), and considers a data period ranging from January 1, 2001 to September 30, 2011, and the main objective was to predict the daily volatility until January 2, 2017. It was also used other variables beside the Stock data, such as the Korean Treasure bonds and the 3-year AA-grade corporate bond. In addition, other variables, commodities, such as Oil and Gold, were also used as a variable that would help to determine the target variable. Models such as the MSE, MAE and others, were used to compare the prediction capacity. The GARCH model performed within the expected values that the researchers expected, nevertheless, the LSTM model outperformed the GARCH model.

My personal assumption for this one, is that GARCH models accrue a value to the long run variance, that might have a significant impact in a such long time period, and also the LSTM by using other variables, beside the prices and statistics of the same, might have a better capacity to explain the price movements and therefore the volatility. Statistics such as correlation between variables, could have a big impact in the predicted outcome, i.e., by assuming that variables are heavily correlated, a change in variable A could equate in an immediate change on the Target variable, by a given, believed, amount.

* The paper “A machine learning approach to predicting stock returns” (Silva, 2021), is based on the monthly returns of the NASDAQ and uses a couple of machine learning techniques that help to understand the behavior of volatility. It also, considers other variables such as technical indicators, being these ones, based on price movement and statistics. In the concrete case of this dissertation, the only two models that will be undertaken are the Linear regression and the CART, in which the Linear regression had the worst perform of both, showing a MSE almost three times bigger than the CART model, despite that in the Training have performed better.
* On the paper “Multi-Transformer: A New Neural Network-Based Architecture for forecasting S&P Volatility (2021)” it is possible to understand how different neural networks, such as Transformer, Multi-Transformer, LSTM and ANN can, by using a wide range of GARCH models, such as EGARCH, FIGARCH and so on, improve these ones, by providing thoughtful inputs to the models, that in the end will equate in a much higher prediction capacity. I consider this paper to be of crucial relevance for any further work, because it not only provides an in-depth study about a recent Neural Network-Architecture, as well as combine a couple of them in a Multi-Transformer model, and can still, based on “older” methods such as the GARCH, improve significatively the Forecasting capacity.
* Summing up, this three papers create the possibility to understand that, models that have a better R2, i.e., use a combination of variables, and weight attributed to the same variables, have a larger capacity prediction, in the sense that they generate outcomes that are based not solely on the previous behavior of the Target variable, and that models that not consider a long run variance, will be able to predict better in the shorter period. It is my fundamental, a priori, believe that the Transformer model will surpass all the other models for all sets of data under analysis.

# Methodology

*“A research design is the strategy for a study and the pan by which the strategy is to be carried out. It specifies the details of how the project should be conducted in order to fulfil the research objective*” (Falinouss, 2007). In this part of the dissertation, it is intended to provide a full description on how the results will be achieved, and with that, the answer to the research problem.

## Research Approach and Design Strategy

Explanatory research aims to develop an initial hunch or insight, and to provide direction for any further research needed. The primary purpose of explanatory research is to shed light on the nature of the situation and to identify any objectives or data needs to be addresses through additional research, working as a sort of benchmark (Falinouss, 2007).This dissertation aims to explain three fundamental questions, that could help clarify on future research, or even provide further empirical proven knowledge that might help when making investment decisions. By acknowledging that the time period under research will be from 01/10/2020 until 04/03/2022, it is expected, due to large macrosocial events that occurred during the same period, that volatility levels turn out to be directly impacted by those, i.e., the market was exposed to large amounts of external factors beside the normal trading that occurs under the expected levels. The main research question that this paper tries to prove, is if there were any type of model, i.e., from the ones above described, that could have helped predicting the first 10 days of volatility since the start of Ukraine’s conflict (for reference it will be assumed the 20th of February (Sunday) as the first day of conflict). Furthermore, in order to provide more globalized research, will be used four different Indexes from 5 different Continents on Earth, namely the Euronext 100 (Europe), the National Stock Exchange India (Asia), the São Paulo Stock Exchange (South America), the SP500 (North America) and the SP/AXS 200 (Oceania), as mentioned in the Abstract. Once again, furthermore, due to the specificity of the models used, some other exogenous variables will be used in order to achieve a lower level of error in predicting.

## Data Collection

It is widely believed that the success of any data solution, is based on the quality of the data that the models use. With this being assumed, is totally crucial to achieve the maximum amount of information during the time period under analysis. For this paper, there will be used two types of variables, the endogenous and the exogenous ones.

* For the Endogenous variables, it will be considered the actual closing daily prices of each index during the period under analysis. Based on these prices will be possible to determine multiple metrics that will be accounted in every single model, as explained below, and this should be considered the most valuable variable of each and every single model. This information is publicly available, either through “Yahoo Finance”, “Bloomberg” or any other information provider.
* Regarding Exogenous variables, these ones will be split between two sets, in order to allow for different frequencies. Low frequency variables, in each it will be used the “Inflation” as a percentage change from quarter to quarter and “House Pricing” as a percentage change from quarter to quarter. All this information is publicly available through the [OCED Public Website](https://data.oecd.org/price/housing-prices.htm#indicator-chart). Regarding the High frequency variables, it will be used the Commodities prices, namely Corn and Brent, since changes on both of them are heavily correlated with the war on Ukraine, being both Ukraine and Russia large producers and exporters, and with their roll there is a shortage on supply. Finally, the volume, i.e., number of shares traded in one day, will also be accounted as a variable to the model, but in this concrete case, every country Index will account their volume only, so for any Indian model it will not be used the Brazil volume. All of this information is widely and publicly available through “Yahoo Finance” or any other free provider.

The Figure 6, provide a direct web link from each it is possible to obtain the same data that it was used on the models (the full path will be available on the Appendix as well).

|  |  |  |
| --- | --- | --- |
| Variable | Data Link Shortcut | Variable Type |
| Euronext 100 | **[^N100](https://finance.yahoo.com/quote/%5EN100/history?period1=1602288000&period2=1646352000&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true)** | **Endogenous** |
| São Paulo Stock Exchange (Ibovespa) | [**^BVSP**](https://finance.yahoo.com/quote/%5EBVSP/history?period1=1601510400&period2=1646352000&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true) | **Endogenous** |
| National Stock Exchange of India (NIFTY 50) | [**^NSEI**](https://finance.yahoo.com/quote/%5ENSEI/history?period1=1601510400&period2=1646352000&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true) | **Endogenous** |
| NASDAQ | [**^IXIC**](https://finance.yahoo.com/quote/%5EIXIC/) | **Endogenous** |
| Inflation | [**Inflation-OCED**](https://data.oecd.org/chart/6Ol5) | **Exogenous Low Frequency** |
| House Pricing | [**House Pricing-OCED**](https://data.oecd.org/chart/6Ol6) | **Exogenous Low Frequency** |
| Brent | [**Brent**](https://datahub.io/core/oil-prices/r/0.html) | **Exogenous High Frequency** |
| Corn | [**CORN**](https://www.macrotrends.net/2532/corn-prices-historical-chart-data) | **Exogenous High Frequency** |
| Volume | **Volume for each Endogenous variable on the links above** | **Exogenous High Frequency** |

Figure 6- Information Source

## Data preparation

The data that will be collected is in the pure form of prices, i.e., the values there are still brute, and in order to get them in the perfect shape for the model, is still needed to do a data analysis work.

For all missing values, i.e., prices that are not available for a single business day, this day will be omitted from the project, and the return for the day immediately after the missing value, will be given by the logarithmic difference between the day missing value plus one and the day missing value minus one.

Firstly, since this is a time series data set, the split between training and test data set should be performed on a basis of temporal continuity, i.e., not randomly, or in a percentage of the total set. Since the main goal in here is to forecast for the first 10 days of war in Ukraine, and assuming that this one as started on 20th February 2022, the training set will account for all observations since 10/01/2020 until 18/02/2022. Therefore, the test set, in which the trained models will perform, will be from 21/02/2022 until 04/02/2022.

Secondly, starting with the endogenous variables, these ones are provided in a Price format. When analyzing stock data, it is highly recommended to convert the same to returns, especially daily, in order to be possible to have a relevant sample size, that will be approximately Normal Distributed, and with this, easier to work with. Otherwise, prices tend to be Log-Normal distributed, and the key statistical parameters will be harder to calculate. Bearing this information in mind, the first step is to calculate the daily return of the Stock index, based on the “Closing Value”, for the entire range of observations. Then, it can be calculated the mean return, as well as the standard deviation for the period.

Thirdly, when it comes to the exogenous variables, for the high frequency ones, since they are in the same time basis that the endogenous ones, i.e., they are also daily and the number of observations matches between both variables, for corn and brent will be used the daily difference, so same as return for a normal stock, and for volume with will be used the face value. For low frequency variables, since these ones are on a different time basis that the target variable, they will be under a process noted as MIDAS, described above on the Literature review topic, and provide quarterly information, in this concrete case, the percentual change from each quarter, that will impact the daily returns by a given percentage under the model, and in the end achieving different parameter values.

Fourthly, the data above described will not be considered under all of the models. In order to better understand this, the Figure 6, in which is defined the type of variables that each model will use.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Daily Returns | Statistic Parameters of Returns | High Frequency Indicators | Low Frequency indicators |
| Monte Carlo Simulation | Yes | Yes | No | No |
| GARCH | Yes | Yes | No | No |
| GARCH-MIDAS | Yes | Yes | No | Yes |
| SVR | Yes | ***Yes/No*** | Yes | ***Yes/No*** |
| LSTM | Yes | ***Yes/No*** | ***Yes/No*** | ***Yes/No*** |
| Transformer | Yes | ***Yes/No*** | No | No |

Figure 7- Variable Inputs by Model

## Results Obtention

Now that is possible to understand each type of model uses each type of date, is time to define the process of achieving the results.

Information about each model as been previously described in the Literature Review segment, but in order to obtain a more graphical example, with will be covered here as well, the main characteristics of the process.

**Monte Carlo Simulation-** At this moment in time, should be assumed that the returns for a given index are already calculated as well as the distribution of the same, and with that the underlying moments , i.e., first and second moment, equating in the mean and standard deviation. Based on this, the annualized mean return will be calculated, by multiplying the mean return per two hundred fifty-two days, and the annualized volatility, will be given by multiplying the daily standard deviation per square root of two hundred and fifty-two. Once these values are obtained, the Drift is defined, and with that the model can start producing random outcomes. Whenever the model produces a random outcome for price at time S, by using *Equation 15*, it is obtained only one price. In order to predict for ten days, it is needed to repeat the process on *Equation 15* ten times. By doing so, the prices obtained will not follow a specific time frame, i.e., they will not be in sequential order. To overcome this difficulty, it was determined that the simulation number one, will equate in the price for the 21st of February 2022, and so on, until the tenth simulation that will equate in the price at 4th of March 2022. By doing so, the output data is now under a continuous, sequential basis, and the calculus of the returns is possible to be obtained. Since the main goal of a Monte Carlo Simulation is to simulate a huge number of times, the process above will be simulated one thousand times, meaning, that in the end will be obtained an array with ten rows, each one representing the daily return of the given day, and will have one thousand columns for each ten-day simulation. After so, the mean return for each row will be calculated, and that should be the return used to compare with the actual return. Then the absolute error and squared error will be calculated in order to obtain the performance metrics.

**GARCH (?,?)-** Once the daily returns are calculated, it is now possible to obtain the estimations for the GARCH model. GARCH model, as demonstrated in *Equation 5* works on a basis of weight attributed to the given variables, as shown in *Equation 6*. The Long run variance is given by deducting one per Alpha and Beta, and these parameters should be calculated using a solver system, in which the objective is to attribute values to either , being that the long run variance weight allocation is always higher than zero, and both range between one and zero. The target variable on the solver system is the likelihood, that should be maximized by changing the weights between the above-mentioned variables. By using RStudio software, it is possible through the “rugarch” library, and by using the “ugarchspec” to specify the conditions in which the model is supposed to work on. The only step that is still missing is to define the GARCH model parameters, so how many moving averages and auto regressions should be accounted in the simulation. The best and proper way to check this is by using the Autocorrelation function, with lag equal to one, and check for each point has a major breakdown. Using that point as reference, providing it to the ugarchspec, still it is needed to check the probability of P being higher than the p-value. If this value is higher than, let’s assume, 0.05, then this variable is not significative for the model, and it should be changed the number of Moving Averages and Auto regressions used. Finally, once those parameters are calculated, it may be used the “ugarchforecast” function to predict for n days ahead, that in this case is only ten. Once these values are calculated is just needed to compare them with the real values by applying the same approach as before, i.e., the absolute error and squared error.

**GARCH-MIDAS**- as explained above in the Literature Review topic, this one model is pretty similar to the normal GARCH but allows for low frequency variables that will be accounted on each day’s return. By using these new returns, all other metrics will be also changing due to the weight each variable have on the target variable. Since the information that the exogenous variables will provide to the model are in return terms, for the first value it will be used the average return of the entire period, and therefore, the normal return. Once these values are calculated is just needed to follow the same process of forecasting and then compare the results. This model will use the same amount of Moving averages and Auto regressions as it’s counterparty.

SVR- DESCRIBE LATER ONCE CODE IS READY

LSTM- Describe LATER WHEN CODE IS READY

TRANSFORMER- Describe LATER WHEN CODE IS READY

# Results and discussion

NOT YET REQUIRED, NEXT STEP

# Conclusions

NOT YET REQUIRED, NEXT STEP

# Limitations and recommendations for future works

NOT YET REQUIRED, NEXT STEP

# Bibliography

Ambrosius, W. T. (2007). *Topics in Biostatistics.* New Jersey: Humana Press.

Andersen, T., Davis, R., Kreiss, J.-P., & Mikosh, T. (2009). *Handbook of Financial Time Series.* Berlim: Springer.

Awad, M., & Khanna, R. (2015). *Efficient Learning Machines.* Apess Open.

Brewer, K., Feng, Y., & Kwan, C. (2012). Geometric Brownian motion, option pricing, and simulation: some spreadsheet-based exercisies in financial modelling. *Spreadsheets in Education*.

Chaudhuri, K., & Koo, K. (2001). Volatilty of Stock Returns: Importance of Economic Fundamentals. *Economic&PoliticalWeekly*, 3852-3856.

Dougherty, C. (2016). *Introduction to Econometrics* (Vol. 5). United Kingdom: Oxford University Press.

Engelhardt, N., Ekkenga, J., & Posch, P. (2021, June 25). ESG Ratings and Stock Performance during the Covid-19 Crisis.

Esling, P., & Agon, C. (2012). Time-Series Data Mining.

Falinouss, P. (2007, June). Stock Trend Prediction Using News Articles.

Fama, E. (1970). *Efficient Capital Markets: A Review of Theory and Empirical Work.* New York, United States of America: Wiley.

Goeztmann, W. (2011). *The Efficient Market Theory and Evidence: Implications for Active Investment Management.* United States of America, Boston: Academia.

Guo, L., Shi, F., & Tu, J. (2017, March 11). Textual analysis and machine Learning: Crack unstructured data in Finance and Accounting. *Textual analysis and machine Learning: Crack unstructured data in Finance and Accounting*.

Houdt, G. V., Mosquera, C., & Nápoles, G. (2020, May 13). A review on the Long-Short Term Memory Model. p. Springer Nature.

Houpt, E. J., & Border, j. (2014). *Stock Market for Beginners.* John Border.

Hull, J. (2018). *Risk Management and Financial Institutions.* Wiley.

Hull, J. C. (2018). *Risk Management and Financial Institutions.* Wiley.

Kim, H. Y., & Won, C. H. (2018, March 06). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models .

Ma, J., Xiong, X., & Feng, X. (2021). News Release and the Role of different types of investors. *International Review of Financial Analysis*.

Mangram, M. (2013). A simplified prespective of the Markowitz Portfolio Theory.

Mathworld, W. (2022). *Wolfram Mathworld*. Retrieved from Wolfram Mathworld: https://mathworld.wolfram.com/NormalDistribution.html

Mondal, P., Shit, L., & Goswami, S. (2014). STUDY OF EFFECTIVENESS OF TIME SERIES MODELING (ARIMA) IN FORECASTING STOCK PRICES. *International Journal of Computer Science*.

Poon, S.-H., & Granger, C. (2001, June 11). Forecasting Financial Market Volatility.

Reddy, K., & Clinton, V. (2016). Simulating Stock Prices Using Geometric Brownian Motion: Evidence from Australian Companies. *Australasian Accounting, Business and Finance Jornal*, p. 27.

Ross, S., Westerfield, R., & Jaffe, J. (2002). *Capital Market Theory: An Overview on Corporate Finance.* New York, United States of America: McGraw-Hill.

Rutkowski, L., Jaworski, m., Pietruczuk, L., & Duda, P. (2014, January 11). The CART decision tree for mining data streams.

Sengupta, C. (2004). *Financial Modeling Using Excel and VBA.* Sidney: Wiley Finance.

Settlements, B. o. (2009, March 13). Revisions to the Basel II. *Basel Committee*.

Sharma, F. C. (2019). *Financial Markets, Institutions and Services.* India: SBPD Publications.

Sharma, H., & Kumar, S. (2015). A Survey on Decision Tree alghoritms of classification in data mining. *International Journal of Science and Research*.

Silva, F. O. (2021, May 21). A MACHINE LEARNING APPROACH TO PREDICTING STOCK RETURNS.

U.S Securities and Exchange Commission, .. (2022). *U.S Securities and Exchange Commission*. Retrieved from U.S Securities and Exchange Commission: https://www.investor.gov/introduction-investing/investing-basics/investment-products/stocks

Westfall, P. H. (2014, August 11). Kurtosis as Peakdness.

Willmott, C. J., & Matsuura, K. (2005, December 19). Advantages of the mean absolute error over the root mean square error in assesing average model performance. pp. 79-82.

Wong, Z. Y., Chin, W. C., & Tan, S. H. (2016, December 19). Daily value-at-risk modeling and forecast evaluation: The Realized volatility approach.

Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*.

# Appendix (optional)

# Annexes (optional)