NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

Index Returns volatility prediction using Machine Learning Techniques

A study on PSI-20 Index volatility

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

TITLE

Subtitle

Student full name

Dissertation / Project Work / Internship report presented as partial requirement for obtaining the Master's degree in Information Management

BOOK SPINE


2019
Title:
Subtitle:
Student
full name
MEGI


2019
Title:
Subtitle:
Student
full name
MGI

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

INDEX RETURNS VOLATILITY PREDICTION USING MACHINE LEARNING TECHNIQUES

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: ***Name***

Co Advisor*:* ***Name***

October 2022

DEDICATION (OPTIONAL)

Full page to this!

ACKNOWLEDGEMENTS (OPTIONAL)

FULL PAGE TO THIS

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 PSI-20 returns and therefore volatility, during a 5-year period, 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.

The main goal is to predict the return for each day, so, daily returns and also the daily volatility which it will be then transformed on a 5-day daily volatility so it can be better comprehended how prices range between a business week. This is a very sensitive topic, so, short term volatility, since some Financial Companies and Investment/Pensions Fund are rated under European Securities Market Authorities and their volatility levels should not be Higher than what is expected, because this could lead members of pension Schemes or either an investor on the Investment fund exposed at higher levels of risk, even if it for a short period of time and not on the long run.

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, my main focus on this project will not be on trying to obtain the most accurate model to predict a 5-day volatility, but to compare how different models predict and if their predictions fall very far from one another.

Due to the lack of scientific research regarding the main Portuguese stock index forecasting methods, this study will be relevant for all of those who may try to obtain a better understanding of the PSI-20 Index behavior and also how different methodologies could be applied in the forecasting process.

KEYWORDS

Stock Index Volatility Prediction; PSI-20 Index; Machine Learning on Index volatility Prediction; Short term volatility impacts; European Securities Market Authority ratings on volatility; Machine learning techniques accuracy on Financial Risk; Add another; Add Another

INDEX

LIST OF FIGURES

LIST OF TABLES

Não foi encontrada nenhuma entrada do índice de ilustrações.

LIST OF ABBREVIATIONS AND ACRONYMS

ML Machine Learning Techniques

ESMA European Securities Market Authority

To be completed an soon as possible

1. 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. 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. Has it is doable to see, in the previous few years a lot of small investors, specially 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 outperform big Investment funds and the S&P 500 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, 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 and subsequentially to the PSI-20 index, even if there is no direct exposure to financial Russian Assets, for example oil prices have increase meaning that Energy companies within this Index could be facing relatively bigger market movements than what was expected.

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 the PSI-20 Index volatility. To do so, we will train, validate, test a series of models and compare their outputs. Some of the models, such as Time Series forecasting models rely solely on the past values of the PSI-20 Index, whilst others, such as Machine Learning algorithms, allow for additional information to be considered when forecasting future values.

2. LITERATURE REVIEW

2.1. FINANCIAL SYSTEMS

As it is Described by Dr. Sharma, (Sharma, 2019)**,** is a link between the savers and the investors. It is made up of all those channels through which savings become available for Investment. We can see this type of action occurring through banks, in which they use the available savings to lend money to borrowers and turning this into an investment, meaning that they are expecting to obtain a higher rate of return then the rate at which they will pay back to the person that save in first place.

Once again, Dr. Sharma reflects that a Financial System means the structure that is available in an economy to mobile capital from various surplus sectors of the economy and allocate as well distribute the same surplus to various needy sectors on the same economy.

In more simple Terms, a financial system is a set of institutions and instruments which foster savings and channelize them to their most efficient use.

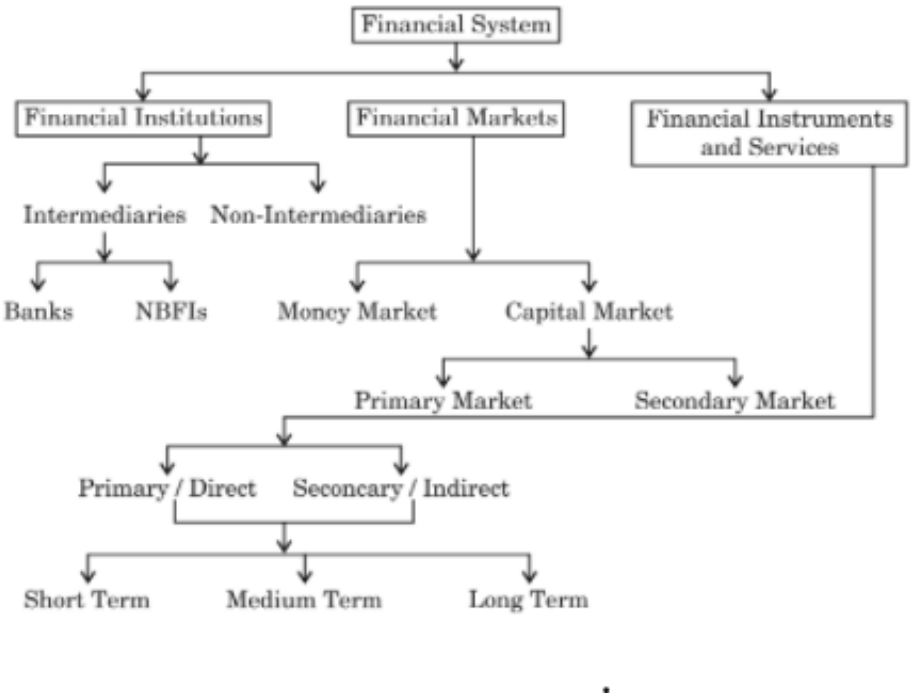

Figure 1- Structure of Financial System

Financial Markets are an important component of financial system, as they are mechanism for the exchange trading of financial Products**,** (Sharma, 2019).

John Hull, (Hull, 2018), confirm the idea above stating that financial institutions do a huge volume of Trading in a wide range of different financial instruments. There are a number of reasons for this, some trades are designed to satisfy the needs of their clients, some are to manage their own risks, some are to exploit arbitrage opportunities, and some are to reflect their own views on the direction in which market prices will move.

It's also stated in the book “Risk Management and Financial Institutions” by John Hull**, (Hull, 2018)**, that there two markets for trading financial instruments, being one the Over-the-counter, which will not be further discussed on this paper, and the Exchange-Traded Markets, that is the one the is relevant for this paper being that the PSI-20 index is traded on the latest.

The role of the exchange is to define the contracts that trade and organize trading so that market participants can be sure that the trades they agree to will be honored.

In the specific case underlying this report, the PSI-20 is quoted and therefore traded on the Euronext Lisbon, being that this entity will provide all type of honor that is required to the trading process.

2.2. STOCKS AND STOCKS INDEXES

As mentioned above, within the Securities Markets exists different assets that can be traded, and for each type of asset is expected a type of Return and also a type of Risk.

Stocks, as it is mentioned by the United States Securities and Exchange Commission, on their official website, **(U.S Securities and Exchange Commission, 2022)** are a type of securities that give stockholders a share of ownership in a company, meaning that they are also called “Equities”, since the account of a company is based on the fundamental equation that the fair value of assets must equal the value of liabilities summed with the value of equity.

$$Fair\ Value\ of\ Assets = Fair\ value\ of\ Liabilites + Equity$$

This means that the equity that a company has should be equal to the difference between the fair value of assets and the fair value of Liabilities.

This means that, when a company have different stockholders, there are a number of different people/companies that own different amounts of that same company (Shares).
Meaning that Equity, should be divided into shares, and each share should have the same value.

$$Share\ Value = \frac{Equity\ in\ €}{Number\ of\ Shares\ Issued}$$

This means, that at a given instant, the value of the share are sold at par, meaning that the market value they have should equal the accounting value.

Once again, the U.S Securities and Exchange Commission, elucidate the reason why people wish to buy shares, and those are mainly because:

- In first instance, they lead to Capital appreciation, which occurs when a Stock rises in price, meaning that now it is being sold above par, so above the accounting value, and that should be due to speculation on future revenue growth for the company. If revenue increases, and the expenses and interest rates due do not increase in the same proportion, this means that in the end of the period the company will have a positive net income, which means profit, and that should be accounted as an Asset, and therefore the Equity should increase in a given percentage.

- Other substantial reason why people buy stocks, is because these ones pay a Dividend, which means that a percentage of the net income (profit) is going to be distributed by all

stockholders, meaning that every share should have the same amount of dividend being paid. For example,

$$Dividend\ Payment = \frac{\%\ of\ Net\ income\ distributed\ to\ Stockholders}{Number\ of\ Shares}$$

- The ability to vote on the shareholder meetings and in this way decide and influence a company should also be a factor to considering when acquiring stocks.

Acknowledging that are diverse types of reasons to why acquire a stock, it is also understandable that stocks should have distinct categories at the eyes of the investors, and depending on which category a stock is included, also depends the inherent levels of risk and return that it is exposed to.

The U.S Securities and Exchange Commission, divides these categories between Growth, Income, Value and Blue-chip Stocks.
it could be assumed that stocks that target Growth, are the ones that should have higher values of risk, and also higher values of expected return, whilst Income and Blue-chip stocks, are expected to have a lower range of price movements and with that having less risk as well being more of a Long term investment, either because they pay a fair amount on dividends, or either because they can achieve an attractive rate of return along the years.

Whilst a single company could have their stocks categorized as one of the above, stock indexes are not that basic, as they combine multiple stocks from multiple companies within it's own portfolio, and due to that, depending on the actual allocation to each category it could be categorized accordingly.

An Index, in this case since the analysis is over PSI-20, a Stock index, accordingly to Evan Houpt and John Border, on their book "Stock Market for Beginners", **(Houpt & Border, 2014)**, could be considered as a survey of the stock Market, in which a company, in the case of PSI-20, Euronext, picks a limited number of stocks that it believes represent the performance of the entire market, and averages their performance to arrive at a number that investors can use to gauge the performance of the market. Every index has it's own stock chart with Opening and Closing prices just like individual stocks.

This difference between the closing price and the opening price on the next day of the index, will be the main focus of this paper since that represents the daily return and consequently the daily volatility.

2.3. EFFICIENT MARKET THEORY

Now that Stocks and Stocks index have been defined is crucial to understand how markets work in theory and how they actually work in practice.

The efficient market theory, as it could be found in the paper written by William Goetzmann "The Efficient Market Theory and Evidence: Implications for active Investment Management", **(Goeztmann, 2011),** asserts that, at all times, the price of a security reflects the available information about is fundamental value. This statement, under the investor's perspective, means that at all times the cost for speculation is high, and therefore should be a losing game.

He also states that, all investors at all times, are faced with the obligation of using a Passive Investment Strategy, an Active Investment Strategy, or a combination of both in their portfolio.

Goetzmann, **(Goeztmann, 2011)**, defines that a passive management strategy is one that uses an Index as a proxy, meaning that the assets within are invested accordingly a specific set of rules an seek to replicate the Index returns and risk metrics.

Whilst Active management could be ascertained as a strategy that is characterized as trading that seeks to exploit miss-priced assets relative to a risk-adjustment benchmark.

Since this paper is based on the actual levels of performance and volatility of the PSI-20, the Euronext Lisbon should be considered as a Passive Investment fund, as it seeks to replicate the actual returns and risk levels of the benchmark index, and not to overperform those. This being said, moving forward on the research, active management should not be studied in depth.

As described by Eugene Fama, the father of this theory in 1970, on its paper "Efficient Capital Markets: A review of Theory and Empirical Work", **(Fama, 1970)**, a market in which prices always "fully reflect" available information is called "efficient".

Eugene, breaks the market in three subsets of efficiency, being those the Weak, in which the information is solely based on historical prices, Semi-Strong, in which the concern is whether prices efficiently adjust to other information that is publicly available such as annual earnings, stock splits and so on, and finally the Strong form, in which only monopolistic groups of investors have access to some given information.

This theory, states that there are no "Gurus" in the stock market, and that the entire market always seeks to be the most efficient as possible, meaning that it is not possible to overperform the market.

In this paper, the focus will be on what is called "weak form" of market efficiency, because it will only be taken in consideration historical prices and all decisions will be solely based on statistics that generate from the analysis of the same prices. Therefore, there should not be expected the research on the impact over stock prices, that fundamental factors have such as news, announcements, crisis and so on. This will be reflected on the volatility during the underlying period, but it will not be based on a sentimental analysis.

2.4. Market Returns and Volatility

All fund managers know that there is a tradeoff between risk and return when money is invested, and when greater risks are taken, the higher the return that can be realized, **(Hull, 2018)**.

Hull also defend that the trade off between risk and return is not based on the actual return but instead on the expected return, that for statisticians the expected value of a variable should be it's average. Therefore, expected returns are a weighted average of possible returns, where the weight applied to particular return equals the probability of that return occurring. The possible returns and their given probabilities could be either estimated from historical data or assessed subjectively.
In this paper, it's going to be assessed based on historical prices, once again reinforcing the idea underlying the weak form of efficiency, and no other subjective variables will be considered.

In mathematical terms, the Expected return of a portfolio should be given by:

$$Expected\ Return = P_i * Return_i + P_j * Return_j + P_n * Return_n$$

Where, P equals the probability of that return to occur, and [i,j,n] represent different options of return.

Once again, using John Hull example, we may conclude that the expected return will always be a combination of possible returns, multiplied by the probability of them to occur.

Probability	Return
0.05	+50%
0.25	+30%
0.40	+10%
0.25	-10%
0.05	-30%

Figure 2- Returns Example

Meaning that in the above case, the Expected return should equal 10%, and this value is given by applying the equation above.

$$Exp.Return = 0.05 * 50\% + 0.25 * 30\% * 0.40 * 10\% + 0.25 * (-10\%) + 0.05 * (-30\%)$$

The returns are the base for all statistic calculations on the financial markets, since they are the best quantitative variable that describes how markets are growing or not. As the expected return is assumed to be the average value of all returns, this means that the return distribution is expected to be Normal, or Gaussian, since random variables with unknown distribution tend to be often assumed as Normal, **(Mathworld, 2022).**

When assessing the normal distribution, financial analyst tend to focus on the 4th moments of the distribution, being the first one the average, in which it is possible from all samples to calculate the Expected value, the second one it the standard deviation, that is the square root of variance, and variance in simple terms equals to the sum of the square distance between every single observation and the Expected value.

$$\sigma^2 = \frac{\sum(x - \bar{x})^2}{N}$$

Equation above, variance is usually denoted as σ^2, Sigma squared, the Expected value is given by $\bar{x}$, x is the individual observation and N is the number of observations.

With this being said, standard deviation, 2nd moment of the distribution, could be defined as the square root of variance, meaning:

$$\sigma = \sqrt{\sigma^2}$$

Since the normal distribution has a probability distribution function, this means that an observation that is y times the standard deviation far from the average value, should have a probability

assumption to it and as it is moving away from the expected value, the probability for that observation to occur reduces.

The third moment is the skewness, which indicates any asymmetric leaning to either the right of left, depending on if the mode is bigger or smaller than the Average, and the 4th moment is the Kurtosis, which indicates the degree of "tailedness", **(Westfall, 2014).**

Even acknowledging that the 3rd and 4th moment of the distribution represent a high value of information when analyzing the distribution of a given variable, this report will not cover in depth the use of the two metrics and will focus more on the 1st and 2nd moment, being the Expected Value and Standard Deviation.

As it is stated by Hull, **(Hull, 2018)**, the 2nd moment of the distribution, so the standard deviation should be the main quantitative risk metric. The standard deviation of the returns is also denominated as volatility of returns, and can be calculated on different time frames, being the most common ones the 1 year and other long-term volatility time basis.

For this paper, it will be assumed the daily returns and as consequence the daily volatility of the same, being after converted into a 5-day daily volatility, so it may be possible to predict with a week in advance what will be the expected risk for the upcoming week.

2.5. 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, **(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.

In the concrete case of this paper, it should not be covered in depth an analysis of each single stock represented on the PSI-20 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 these 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 o volatility as given by:

$$Sharpe\ Ratio = \frac{[E(R)] - Risk\ free\ rate}{\sigma}$$

Within the Capital Allocation Line, the point with Highest Sharpe ratio, will work the tangent point for the Capital Market Line.

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 PSI-20 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.

2.6. 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 a 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 use quantitative variables, since these ones are easier to model and also, easier to obtain, sometimes it is also a kay factor to use qualitative variables since this 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 since 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 PSI-20, 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 or 3 sets, namely the Training, the Validation, 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", **(Xu & Goodacre, 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.

The Validation set consists of a set of samples with known provenance, but these classifications are not known to the model, therefore, predictions on the validation set allow the operator to assess model accuracy. Based on the errors on the validation set, the optimal model parameter set is determined using the one with the lowest validation error. This procedure is called model selection, **(Xu & Goodacre, 2018).**

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.

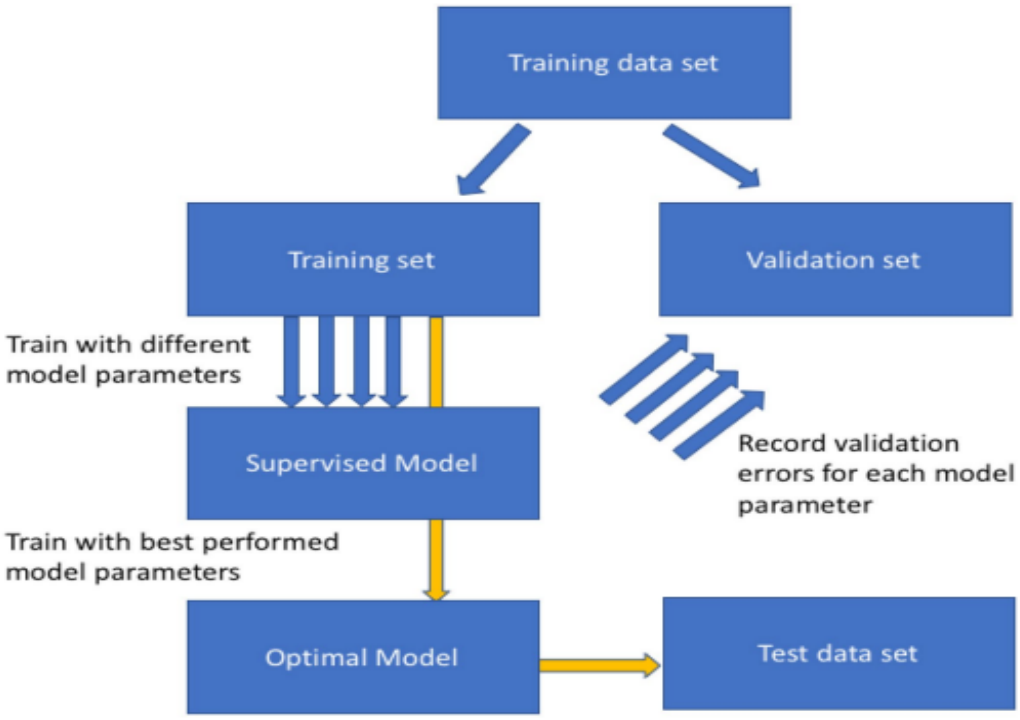

Figure 3- General flowchart used for model selection

By considering prices of the Psi-20 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 validation/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.

2.7. Value at Risk and is importance for Investment Managers

Value at risk is a key indicator when accessing the overall risk that a Portfolio Contains. This metrics is so fundamental that even Basel II demands that ,financial institutions that face market risk, are able to provide this value for a confidence interval of 99% at a 10-day basis, **(Settlements, 2009).**

Bearing this in mind, is therefore crucial that these same financial institutions keep on track the actual volatility they are facing in their portfolios. V@R can be calculated in numerous forms, but for the scope of this dissertation, the focus will be only on the parametric calculation, i.e., the assumption that returns follow a normal distribution and therefore, an Investment Manager, should be able to predict a confidence interval at a given percentage.

Even by understanding that any type of V@R calculation will be covered on this dissertation, it is one more empirical argument that fundaments the importance of the Distribution moments, i.e., mean and Standard Deviation of returns.

2.8. 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 should be believed, 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.

2.9. 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), the Exponential Weighted Moving Average (EWMA) and the Autoregressive Integrated Moving Average (ARIMA).

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, σ_n^2 is calculated based on a long-run average variance rate, V_L, as well from σ_{n-1} and μ_{n-1}. A given weight is attributed to each of these variables, which means that this is a weight model. The objective is to Maximize γ, which is the weight of V_L, by changing the allocations between α, weight given to μ_{n-1} and β, weight given to σ_{n-1}, **(Hull J. C., 2018).**

$$\sigma_n^2 = \gamma V_L + \alpha\mu_{n-1}^2 + \beta\sigma_{n-1}^2$$

The weight allocation function is provided by:

$$\gamma + \alpha + \beta = 1$$

The EWMA model, is a model similar to the above, yet since the weight given to older observations decreases exponentially as we move back in time, the Long-run variance has no impact on the model, i.e., $\gamma V_L = 0$.
Also, the parameters $\alpha \; and \; \beta$, are also replaced by λ, which is variable between 1 and 0, **(Hull, 2018).**
In this case, λ should be the weight provided to the previous day variance, i.e., σ_{n-1}^2, and $(1-\lambda)$ should be the weight given to the previous day squared mean return, i.e., μ_{n-1}^2.

$$\sigma_n^2 = (1-\lambda)\mu_{n-1}^2 + \lambda\sigma_{n-1}^2$$

Finally, the ARIMA model, also similar to the above, an ARIMA model converts a non-stationary data to stationary data. ARIMA(p,d,q) is where p denotes the autoregressive parts of the data set, d refers to integrated parts of data and q denote moving averages parts f the data set and all of them, i.e., pdq are non-negative integrals, **(Mondal, Shit, & Goswami, 2014).**

$$yt = \beta 0 + \beta 1 yt - 1 + \beta p yt - p + \varepsilon t$$

2.10. 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 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 in 2004, **(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.

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t$$

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, et al., 2012).**

By using a Monte Carlo Simulation, it is possible to generate a Stock Price for a given day, and from that price calculate the return and volatility.

The formula is breakdown in three steps:

$$Z = rnorm\ (x, \mu = 0, \sigma = 1) \qquad Wt = \sqrt{T} * Z$$

$$St = S_0 e^{[\ (\hat{\mu} - 0.5 * \sigma^2)T + \sigma * Wt\ ]}$$

Where:

- Z is given by a random normal distribution, with x number of simulations, and assuming that mean is 0 and standard deviation is 1.
- Wt is described as the Wiener process, and is given by multiplying the Square root of time by the Z variable.
- St (spot price at time t) is given by multiplying stock price at time 0 by the log normal distribution, i.e., drift ($\hat{\mu}$) – ½*variance, multiplied by time, plus standard deviation multiplied by the Wiener Process.

2.11. MACHINE LEARNING MODELS

By undertaking that the data series under analysis in this dissertation is a Quantitative Time series data set, the use of some models may be more accurate than others.

The models below should account for a multivariable data set, that for the concrete case of this dissertation, will be defined by adding other information that it appears to be relevant, such as Oil prices, since majority of the companies in the PSI-20 index are exposed, directly to this variable, IBEX-35, based on the assumption that should exists a negative correlation between the Spain index and Portuguese index, meaning a tradeoff between investing in Portugal or Spain, and finally, the EURIBOR 12 Months free interest rate, based on the assumption that there is also a tradeoff, in this case, between free interest rate financial assets and equities.

With this in mind, the following ones will be cover on the dissertation:

Support Vector Regression, that is similar to Support Vector Machine. SVM 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)**.

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 2^{n} different labels. This capacity is related to the amount of training data available, **(Awad & Khanna).**Based on the above, the VC dimension a h affects the generalization error, as it is bounded by $\|\omega\|$ where ω is the weight vector of separating hyperplane and the radius of the smallest sphere $\mathcal{R}$ that contains all the training points, according to :

$$h < \frac{R^2}{\|\omega\|^2}$$

The overall error of a machine learning model consists of:

$$\varepsilon = \varepsilon_{emp} + \varepsilon_g$$

Where ε_{emp} is the training error, and $\varepsilon_g$is the generalization error.

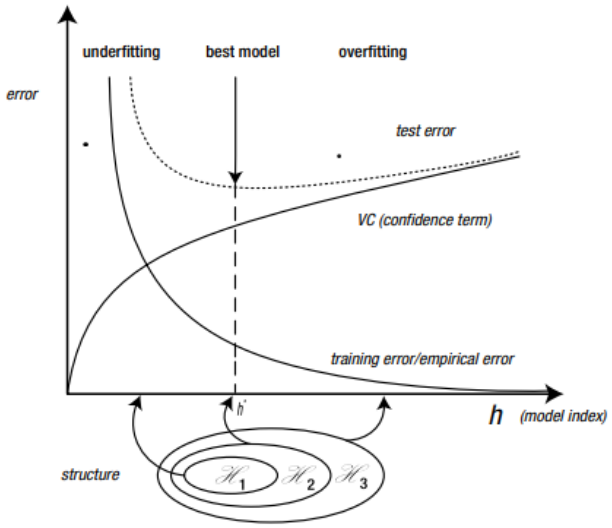

Figure 4- Relation between Error and Model Index

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).**

For a SVR the formula should be:

$$y = f(x) =< \omega, x > +b = \sum_{j=1}^{M} \omega_j x_j + b, y, b \in \mathbb{R}, x, \omega, \in \mathbb{R}^M$$

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

$$f(x) = \begin{bmatrix} \omega \\ b \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix} = \omega^T x + b \quad x, \omega \in \mathbb{R}^{M+1}$$

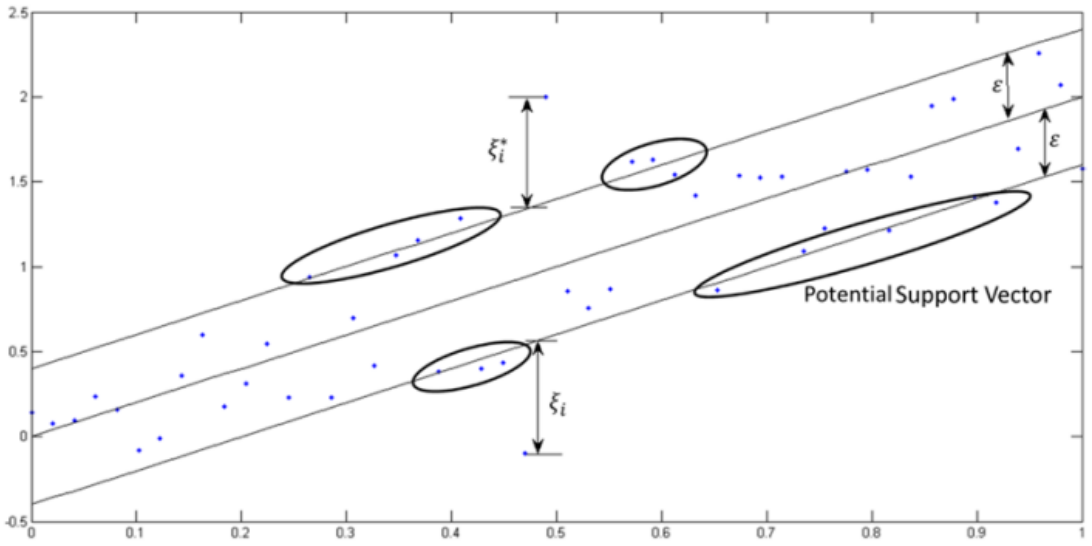

Figure 5- Support Vector Regression example

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 $x^{(t)}$ and the output of that LSTM unit $y^{(t-1)}$ in the last iteration.

$$Z^{(t)} = g(W_Z x^{(t)} + R_Z y^{(t-1)} + b_Z$$

Where, W_Z and R_Z are the weights associated with $x^{(t)}$and $y^{(t-1)}$ respectively, whilst b_Z represents the bias weight vector.

Input Gate- it combines the current input $x^{(t)}$, the output of that LSTM unit $y^{(t-1)}$ and the cell value, $c^{(t-1)}$ in the last iteration.

$$i^{(t)} = \sigma\left(W_i x^{(t)} + R_i y^{(t-1)} + p_i \blacksquare c^{(t-1)} + b_i\right)$$

Where ■ denotes the point-wise multiplication of two vectors, W_i, R_i, p_i are the weights provided to $x^{(t)}, y^{(t-1)}, c^{(t-1)}$ respectively, whilst $b_i$represent the bias vector of the component.

Forget Gate- The LSTM unit determines which information should be removed from its previous cell states $c^{(t-1)}$. Therefore, the activation values, $f^{(t)}$, of the forget gates at time step t, are calculated based on the current input $x^{(t)}$, the outputs $y^{(t-1)}$, and the state $c^{(t-1)}$ of the memory cells ate previous time step $(t-1)$, and b_f is the bias terms of the forget gates.

$$f^{(t)} = \sigma\left(W_f x^{(t)} + R_f y^{(t-1)} + p_f \blacksquare c^{(t-1)} + b_f\right)$$

Where ■ denotes the point-wise multiplication of two vectors, W_f, R_f, p_f are the weights provided to $x^{(t)}, y^{(t-1)}, c^{(t-1)}$ respectively.

Cell- this step computes the cell value, which combines the block input $Z^{(t)}$, the input gate $i^{(t)}$ and the forget gate $f^{(t)}$, with the previous cell value.

$$c^{(t)} = Z^{(t)} \blacksquare\ i^{(t)} + c^{(t-1)} \blacksquare f^{(t)}$$

Output Gate- is a combination of the current input $x^{(t)}$, the output of that LSTM unit $y^{(t-1)}$ and the cell value $c^{(t-1)}$ in the last iteration.

$$o^{(t)} = \sigma\left(W_o x^{(t)} + R_o y^{(t-1)} + p_o \blacksquare c^{(t-1)} + b_o\right)$$

Where ■ denotes the point-wise multiplication of two vectors, W_o, R_o, p_o are the weights provided to $x^{(t)}, y^{(t-1)}, c^{(t-1)}$ respectively, whilst $b_o$represent the bias of the weight vector.

Block Output- combines the current cell value $c^{(t)}$ with the current output gate.

$$y^{(t)} = g(c^{(t)}) \blacksquare\ o^{(t)}$$

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

$$\sigma(x) = \frac{1}{1 + e^{1-x}}$$

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

$$h(x) = g(x) = \tanh(x)$$

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

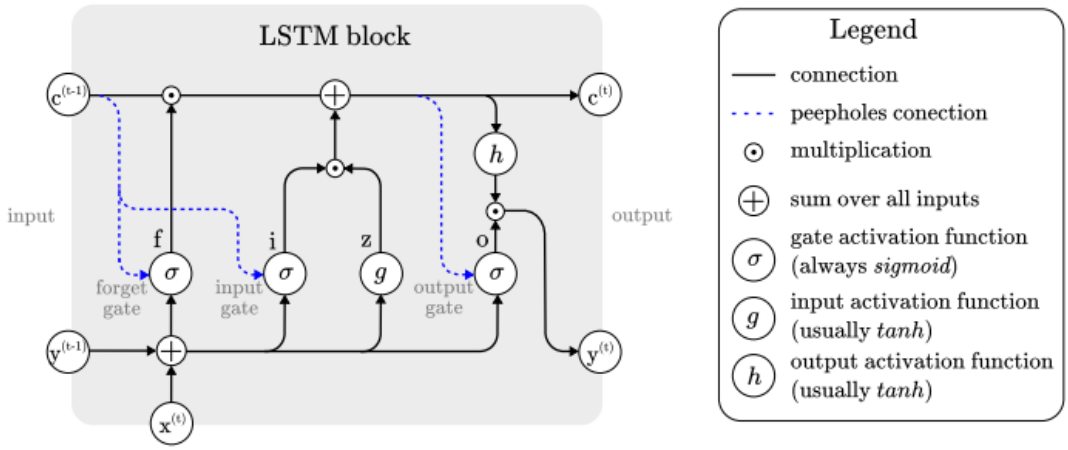

Figure 6- LSTM process

Other model often used as a predictive method for quantitative continuous time series data is the Classification and Regression Tree, CART, that is a form of decision tree.

Based on the publication paper " The CART decision tree for data mining data streams", **(Rutkowski, Jaworski, Pietruczuk, & Duda, 2014),** is possible to denote that the most important task in constructing decision trees for data sets is to determine the best attribute to make a split in the considered node. To solve this problem, it may be applied the Gaussian approximation.

Some of the CART advantages are that this one is Nonparametric, i.e., there are no probabilistic assumptions made over the distribution of the variable, it automatically performs variable selection, uses a combination of continuous or discrete variables, and establishes interactions among variables, **(Sharma & Kumar, 2015).**

CART uses Gini index to rank tests, and this tests in CART are always binary. Also, CART prunes trees with a cost-complexity model whose parameters are estimated by cross-validation.

Builds both classifications and regression trees. The classification tree construction by CART is based on binary splitting of the attributes. Gini index is used as splitting measure in selecting the splitting attribute. CART is different from other based algorithm because it is also use for regression analysis with the help of the regression trees. The regression analysis feature is used in forecasting a dependent variable given a set of predictor variables over a given period of time. CARTS supports continuous and nominal attribute data and have average speed of processing, **(Sharma & Kumar, 2015).**

The Gini index is defined by:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}}$$

where, x_i is a daily return, n is the number of observations, and $\bar{x}$ is the mean of variable x_i in this case, the mean of the daily returns.

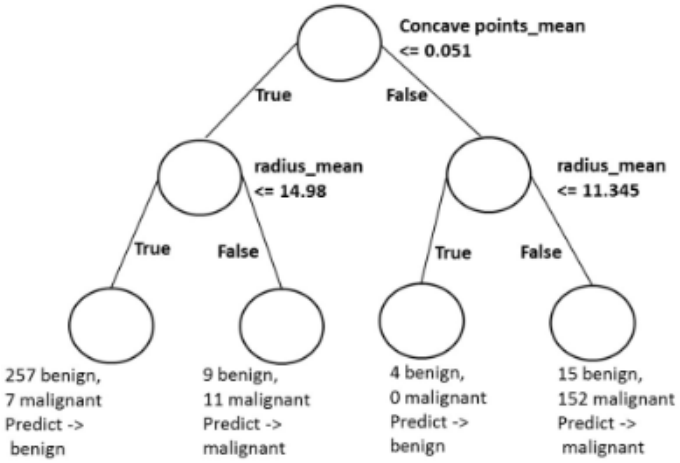

Figure 7- Example of CART Machine

Finally, the last model covered on this thesis will be the Linear Regression. This is one, if not, the most well-known machine learning algorithm and it could be applied in any field of study, so it is relevant to include it.

Assuming correlation as a numerical summary that describes the degree of which two continuous variables, X, Y, are linear related to each other. A simple linear regression of , X, Y, takes this one step further and formalizes a statistical model between the two variables. X is variously known as the covariate, or the predictor, explanatory, or independent variable. Correspondingly, Y is known as the outcome, or the predicted, response, or dependent variable. This is in contrast with a correlation, which does not make this distinction between which variable is explanatory and which is outcome. **(Ambrosius, 2007).**

Despite the assumption that this is a model that is used often, in order to use it, the target variable should comply with these five assumptions, **(Ambrosius, 2007):**

Linear Relationship: The relationship between the independent and dependent variables should be linear. This can be tested using scatter plots.

Multivariate Normal: All the variables together should be multivariate normal. For all the variables to be multivariate normal each variable separately has to be univariate normal means a bell-shaped curve. This can be tested by plotting a histogram.

No Multicollinearity: There is little or no multicollinearity in the data. Multicollinearity happens when the independent variables are highly correlated with each other. Multicollinearity can be tested with correlation matrix.

No Autocorrelation: There is little or no autocorrelation in the data. Autocorrelation means single column data values are related to each other. In other words, $\mathrm{f}(\mathrm{x}+1)$ is dependent on value of $\mathrm{f}(\mathrm{x})$. Autocorrelation can be tested with scatter plots.

Homoscedasticity: Homoscedasticity is there. This means “same variance” .In other words residuals are equal across regression line. Homoscedasticity can also be tested using scatter plot.

This linearity is formalized by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_n X_n + \varepsilon$$

Where Y is the target variable, $\beta_0$is the interception point, i.e., the value of Y when X is equal to zero, β_n is the slope of the distribution of X_n, i.e., the weight that the variable X_n will have on the final outcome. Finally, ε is the error, i.e., the part of the regression that our model is not able to explain and could be also undertake as the difference between the actual value and the predicted one.

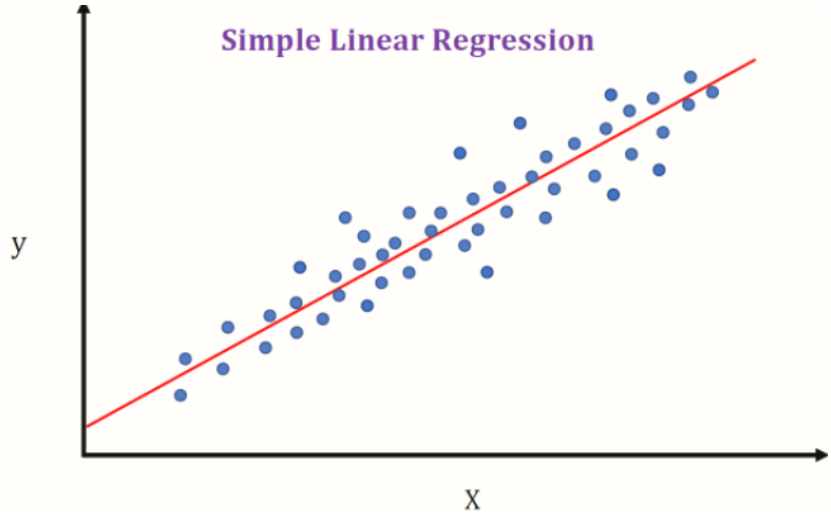

Figure 8- Linear Regression Example

2.12. 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 is 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 n, **(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.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

Where, n is the number of observations, x_i is the output generated from the model, x is the actual, observed value and $|x_i - x|$ 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.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|^2$$

Where, n is the number of observations, x_i is the output generated from the model, x is the actual, observed value and $|x_i - x|$ 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.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|x_i - x|^2}$$

Where, n is the number of observations, x_i is the output generated from the model, x is the actual, observed value and $|x_i - x|$ is the absolute error.

3. METHODOLOGY

Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology Text about methodology.

4. RESULTS AND DISCUSSION

Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion Text about results and discussion.

5. CONCLUSIONS

Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text Conclusion text.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works Text of limitations and recommendations for future works.

7. BIBLIOGRAFIA

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. (n.d.). *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*.

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

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

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.

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*.

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*.

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*.

8. APPENDIX (OPTIONAL)

9. ANNEXES (OPTIONAL)