INTERNATIONAL
HELLENIC
UNIVERSITY

Data Mining Forecasting Stock Fluctuations

Dimitrios Kalliantasis

SID: 3305160004

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in e-Business and Digital Marketing

JULY 2019

THESSALONIKI – GREECE

INTERNATIONAL
HELLENIC
UNIVERSITY

Data Mining Forecasting Stock Fluctuations

Dimitrios Kalliantasis

SID: 3305160004

Supervisor:	Prof. Christos Tjortjis
Supervising Committee Members:	Assoc. Prof. Name Surname Assist. Prof. Name Surname

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in e-Business and Digital Marketing

JULY 2019

THESSALONIKI – GREECE

Abstract

This dissertation was written as a part of the MSc in e-Business and Digital Marketing at the International Hellenic University. The goal of this dissertation was to develop, amplify and test the accuracy of fast non-linear regression algorithms in an attempt to exploit mining knowledge either by utilizing existing algorithms or developing new methods in order to forecast the dynamic non-linear fluctuations of stocks and their price, bearing in mind and monitoring the gravity external or internal factors convey. External factors can be political, economic, social and technological among others, while internal factors are fundamental, technical and market sentiment factors as well.

Acknowledgements

I would like to express my appreciation and gratitude to my supervisor Professor Christos Tjortjis for his patience and his valuable guidance and support without which I wouldn’t have managed to integrate this work.

Dimitrios Kalliantasis

Date 31/07/2019

Contents

1 Introduction

In an era of ever-increasing technological achievements and the requirement of developing new sources of income the world experiences a wide variety of emerging businesses. This startup community contains diverse types of ventures addressing to multiple and separate market spaces. Such ventures can be technological, social media, agricultural, energy, transportation, and accommodation oriented among others.

Since many economies across the world are bouncing back from a recession or still experience financial crisis while other countries are in a developing state of economy, new or existing entrepreneurs are trying to exploit novelty in order to gain the most out of the market offering several products or services. As a result, a few of these companies accomplish to enter the Stock Exchange disturbing the balance of the respective market.

This unresolved matter attracts the interest of many researchers and scholars for quite a long time. As it happens the Stock Market drew attention since the very beginning that the first bourse was founded in Bruges, Belgium in the late medieval period. Since then, merchants and businessmen all over the world pursue to acquire the Holy Grail of economy, which is no other than money and uncountable earnings. This problem agonizes brokers and other stakeholders daily as a potential downfall of stocks prices will sustain severe loss not only to specific businesses or industries but also larger economic scales such as countries or unions. This is the so-called domino effect. In other words, there are two faces of the same coin, one that can lead to glory and another that can cause chaos, and chaos is exactly where the truth of stocks lies.

In order to reduce this devastating effect and determine the stock chaos in effective and profitable manner, a vast amount of surveys have taken place as well as algorithms have been developed throughout the decades aiming at forecasting the fluctuations of stocks. Many attempts to predict the Stock Markets worldwide have proven ineffectual while others scored quite good results yet, no adequate enough to be considered panacea to the cause they are delivering.

The aforementioned non-linear regression algorithm that we are about to develop aims at weighing the gravity conveyed by external and internal factors that affect stocks.

External factors can be political, economic, social and technological among others, while internal factors are fundamental, technical and market sentiment factors.

● With regards to the fundamental factors, which may slightly vary depending on each market domain, there are some metrics that should be mentioned. These are the earnings per share or alternatively free cash flow per share and the valuation multiple such as the P/E ratio that is determined with respect to each market segment, the actual stock price, the growth rate and the stock discount rate which is inflation based, e.g. high inflation translates to high discount rate or else lower stock price.

Fundamental analysis in simple words deals with the intrinsic value of stocks. Investors seem more eager to invest on a stock when the intrinsic value of the stock is higher than its current value. The rationale behind this is to purchase a valuable stock at the lowest price possible.

● Regarding technical factors there are also many elements that may influence the performance of a stock. These include the following:

i) inflation based, as low inflation drives high multiples and vice versa,

ii) deflation based, which is bad for stock as it entails loss in the pricing power,

iii) economic strength of the market and peers, which seem to be correlated at 90% of the cases,

iv) potential substitutes on a global stage,

v) incidental transactions motivated by other than the intrinsic value of the stock, which may impact supply and/or demand,

vi) demographics, meaning that the higher the number of middle-aged investors the higher the valuation multiples,

vii) trends that may affect companies that gather momentum conveyed in the stock price or have suffered from reversion induced by the trend,

viii) liquidity, i.e. large cap stocks that show high liquidity are well followed and transacted contrary to small cap stocks (liquidity discount)

Technical analysis takes into account the above parameters attempting to predict the way stocks fluctuate by tracing patterns which stem from the study of relevant to the stock market charts. Such charts describe historical data based on the previously mentioned factors.

● Market sentiment incorporates behavioral finance and social science while psychology affects the market participants either individually or collectively. Likewise, there are some characteristics here as well:

i) data over emphasizing,

ii) investors demonstrating greater pain to losses than to gains,

iii) investors persist to mistakes

● Eventually, political, economic, social and technological factors can be possible elections, turmoil, financial crisis, recession, unstable environment as well as environmental changes or disasters, technological improvements and innovations, renewable and non-renewable energy etc. Factors such the above are witnessed to have made a severe impact on the stock market in multiple occasions in the past and present meaning it is a certainty that they will have high effect in the short and long-term future as well.

1.1 Chapter Overview

Following the goals and objectives that were previously described, we present a brief summary of the content of each chapter.

Chapter 2 includes all the relevant literature that deals with stock prediction algorithms in the past. A research of several studies takes place to define whether there has been any progress made concerning this fields or not. In chapter 2 we attempt to find out what data mining has delivered as an outcome when predicting stock fluctuations, as well as the level of success data mining methods have provided in order to eventually present our findings. The final part involves the citation of the predictive technique which we will exploit to overcome possible issues concluded from the literature review of both subjects and the clarification of its benefits.

In Chapter 3, we attempt to present an analysis of the requirements each subject entail. Hence, it is necessary to mention and elaborate on investment theories, data effect and relevance, market predictability and predictive variables for the stock market.

Chapter 4 includes in detail all the elements that are related to the data which are going to be used for our task. The objective is to cover topics such as data description, quality and origin, thus, to configure the appropriate datasets in terms of compatibility with the data mining technique used and finally, proceed to a randomness test.

In Chapter 5 we present the whole process of developing and testing linear and non-linear regression algorithms along with the methods that will be applied. This Chapter is a detailed description of the functions of *Auto-Regressive models*, *Traditional Time Series* and *Neural Networks* as well as the parameters and the variables that we will be deployed.

Chapter 6 presents several tests that we will be applied to assess the algorithms regarding speed, accuracy and scalability. Furthermore, we attempt to benchmark the models under consideration in comparison with the other existing data mining algorithms and eventually, evaluate the work that was put through the project.

In Chapter 7 we present a summary of the findings this thesis has concluded. Finally, suggestions on future developments that could apply on the stock market valuation and forecast are cited.

2 Literature Review

This chapter contains the most notable and relevant literature that was published by accredited scholars and researchers with regards to the Stock Market.

2.1 Stock Background

The stock market background includes all the previous works, surveys, achievements and developments regarding the stock exchange that took place throughout the decades.

Stock exchange is widely known since the late medieval ages when the first stock market was founded in Bruges, Belgium. Since then it has dominated the interest of academic researchers, economists, bankers, traders, enterprises and financial behemoths. That wide interest drew such attention that many surveys and efforts have been made thus, to manage to master the fluctuation patterns of stocks. Yet, such attempts were proven a struggle in terms of accuracy and efficacy due to the diversity of the factors that can determine the movement of stocks.

Generally there are two different beliefs preserved with regards to stocks, the one claims that, as trends undergo certain patterns then the stock market can be predicted while the accuracy of the prediction depends on the method itself and the configuration of the right set of parameters and the other one which suggests that the stock market is purely based on randomness. On both occasions however, there has not been any method to provide acceptable and accurate results when predicting the next stock movements.

The most commonly used approaches of stock market trading are the fundamental analysis and the technical analysis. Fundamental analysis deals with external factors that may affect the market such as domestic and foreign events and political and economic current affairs. On the other hand, the underlying philosophy of the technical analysis is that stock prices comprise a solid performance indicator of the market hence, it evaluates solely on the prices of the stocks rather than external factors. In other words, it relies on Occam’s razor hypothesis where, out of two methods, the one with the least hypothesis is the most acceptable.

Since the predictability of the market was always attracting a lot of researchers and academics, there was a hypothesis formula developed known as the *Efficient Market Hypothesis* (EMH), which states that prices include all the information relevant to a market and every time new information arise the market itself corrects and absorbs it, in order to turn the market efficient. More specifically there are three formulas of EMH [1]:

- *Weak*: States that the prediction of future stock prices cannot be achieved on the basis of past stock prices.
- *Semi-Strong*: States that future prices cannot be predicted from utilizing published information.
- *Strong*: Claims that no matter what information there are available the market is unpredictable.

The above suggest that the market fluctuations utilize the *‘Random Walk’* model which is equivalent to:

$$y(t)=y(t-1) + rs$$

where y(t) is the value of the market on time t and rs is an *Independent and Identically Distributed* variable. This model implies that the best prediction about a future value is a present value.

The latest endeavors that have taken place towards the stock forecasting goal are inextricably bound to data mining and machine learning techniques. Data mining as a rapidly growing technology regarding the information processing industry has been applied to several fields including engineering, business and finance among others [9, 10]. Stock prediction is an extremely challenging task because of its complexity as well as its non-linearity as a dynamic system.

The most popular methods of soft computing exploited in market trends prediction are Traditional Time Series Prediction, Genetic Algorithms, Support Vector Machines and Artificial Neural Networks [ref].

Traditional Time Series Prediction models are used for modeling linear relationships among the factors that can potentially influence the market of stocks. Therefore, they are a widely used prediction method in econometrics dealing with univariate and multivariate regression. The most common univariate model is the Box-Jenkins model that exploits only one variable within the recurrent equation of the autoregressive moving average model:

$$X_t = \varphi_1 X_{t-1} + \ldots + \varphi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}$$

Where the φ's are the autoregressive parameters, the θ's are the moving average parameters, the X's are the original series, and the a's are a series of unknown random probability distribution errors (or residuals).

The Box-Jenkins model provides good results on the short-term, however it requires a large amount of data. Setting aside the univariate models, the multivariate are more specific ones in terms of the causality of factors affecting data behavior. In other words, more than one variable is taken into account.

The Support Vector Machine technique, which was proposed back in 1970s by Vapnik [ref], is considered to be the pinnacle of classifiers while, additionally, it can be applied in solving regression problems, especially in financial time series prediction. The model is based on the principle of the structural risk minimization as well as reducing the generalization error in order to achieve high performance. The technique is also better to solve linear programming problems contrary to the function of Neural Networks.

Neural Networks on the other hand, tend to discover non-linear relationships that exist within the training set input and output pairs which, makes them the best choice when it comes to model non-linear dynamic systems like stocks. They are quite tolerant to noise created by incomplete data representation. Neural Networks seem to be very beneficial in forecasting complex problems such as stock prices, due to its non-parametric and non-linear learning properties. However, they present a difficulty in determining the significance and the weight of each used variable and they also face overtraining problem, which occurs when the network fits data too well ending up to decreasing the ability of generalization. The latter happens because of the number of nodes and the long training time periods. Yet, overtraining can be avoided by testing or cross validation.

Genetic Algorithms are adaptive heuristic algorithms that repeatedly modify a population of individual solutions. The algorithm randomly selects individuals out of the current population as parents in order to produce the next generation, a procedure that is repeated towards an optimal outcome. Their main difference from classic algorithms is that at each iteration they generate a population and not a single point where, the best point is considered the optimal one as well as the selection of the next population is random and not deterministic. Genetic Algorithms are suitable for solving several optimization problems and functions that are non-linear, discontinuous or non-differentiable.

In addition to the above, there have been made attempts on hybridizing the previously mentioned techniques thus, to deploy and combine the advantages of each one them into one predicting method.

The Self-Organizing Feature Map (SOFM)-Support Vector Regression (SVR) comprises a hybrid forecasting attempt that uses a filter-based feature selection. The Self-Organizing Feature Map was proposed in 1989 by Kohonen [ref] and is an unsupervised learning algorithm that spatially converts the input samples into relationships among two-dimensional grids that respond to the inputs. The Support Vector Regression deals with real valued functions minimizing the error generalization. The combination of the SOFM with SVR tries to improve the algorithm's accuracy while reducing its training time. Moreover, by combining the above with filter-based feature selection, the ability of selecting the better feature is achieved.

In order to train a model such as this one, there is need in performing a two-stage process. The first step is to cluster the training data with the SOFM algorithm and then as second step is to construct an individual SVR model for each cluster while at the same time the kernel and loss functions have to be selected. Additionally, the kernel parameter gamma (γ), the loss function parameter ε and the soft margin constant C have to be optimized. The process is established as following:

At first, at the preprocessing stage all input variables need to be scaled thus, to avoid calculating difficulties and increase accuracy. Each feature is scaled to the [0,1] range according to the equation below:

$$x' = \frac{x - \min_a}{\max_a - \min_a}$$

Where the value x' is the scaled value, x is the original one, $\max_\alpha$ and $\min_\alpha$ are the maximum and the minimum value of feature α respectively.

Secondly, at the actual first stage the clustering of the training dataset takes place separated in two clusters one including similar objects and another with dissimilar ones so that high-dimensional data can be better visualized under a configured two-dimensional grid.

The second stage is to train individual SVR models for the two clusters by using the loss function parameter ε along with the parameters C and γ where the pair with best cross-validation accuracy regarding the Mean Square Error is selected.

2.1.1 Stock Background Assessment

The assessment part has to do with the evaluation of prediction methods in terms of strengths and weaknesses such as efficacy, accuracy and speed among others while monitoring and identifying existing or remaining issues.

- **Prediction Methods**

The market predictability is undoubtedly an extremely interesting task. According to the literature there are various methods used, ranging from chart studies to linear or non-linear regressions such as the following:

- Technical Analysis Methods,
- Fundamental Analysis Methods,
- Traditional Time Series Prediction Methods
- Machine Learning / Data mining Methods.

2.1.2 Stock Market Features Summary

This is a summary of what other methods that pertain to stocks offer, what are their main features and what focal point they are trying to address.

- **Technical Analysis**

"Technical analysis is the method of predicting the appropriate time to buy or sell a stock used by those believing in the castles-in-the-air view of stock pricing" [1]. The idea behind technical analysis is that share prices fluctuate dictated by the constantly changing attributes due to their dynamic nature. Technical data such as price, volume, highest and lowest prices per trading period are exploited in charts thus to predict future stock movements. Price charts are used to detect trends, which are originated from supply and demand issues. Technical analysts utilize these charts to extract trading rules so that they can use them within the stock market. The study of these charts suggests that the present actions of the investors will shed light on potential crowd actions, which makes it a very popular prediction approach. However, it has been heavily criticized due to the subjective nature of the charts resulting in different conclusions extracted by different analysts.

- **Fundamental Analysis**

'Fundamental analysis is the technique of applying the tenets of the firm foundation theory to the selection of individual stocks" [1]. Here, analysts use fundamental data in order to unveil all the aspects of the industry or market they will select to invest in. Their aim is to compute the 'real' value of the asset they will make their investment by studying variables such as the growth, the dividend payout, the interest rates, the risk of investment, the sales level, the tax rates etc. Their objective is to calculate the intrinsic value of an asset (e.g. of a stock). Since they accomplish this task, they apply a simple trading rule as follows: *If the intrinsic value of the asset is higher than the value it holds in the market, invest in it. If not, consider it a bad investment and avoid it.* The fundamental analysts are of firm belief that the market is 90 percent defined by logical and 10 percent by physiological factors.

- **Traditional Time Series Prediction**

The Traditional Time Series Prediction analyzes historical data attempting to exploit future values of a time series as a linear combination of these historical data. Time series prediction is classified in two different types, the *univariate* (simple regression) and the *multivariate* (multivariate regression) [2] which are the most common in econometrics. The application of this method starts with the formation of specific factors which are *the explanatory variables xi* of the prediction model. Then a mapping between their values xit and the values of the time series yt *(y is the to-be explained variable)* takes place in order to form specific pairs {xit , yt}. These pairs are used to define the importance of each *explanatory variable* in the formulation of the *to-be explained variable*. *Univariate* models are based on one *explanatory variable* (I=1) while *multivariate* models use more than one variable (I>1).

- **Data Mining Methods**

Data mining methods use a set of samples to generate an approximation of the underling function that generated the data. Knowledge is extracted from these samples when unseen data are presented to a model with the possibility to infer the *to-be explained variable* from these data. The Neural Networks methods and their derivatives have been widely applied to predict the market.

Neural Networks is a data processing technique that maps an input stream of information to an output stream of data [3]. Neural Networks (NNs) can be used to perform *classification* and *regression* tasks. Neural Networks are consisted of *neurons* (or *nodes*) distributed across *layers*. The *structure of the network* is defined by the distribution of these neurons and the way they are linked. Each neuron node is characterized by a *weight* value. A *neuron* is a processing unit that takes a number of inputs and gives a distinct output. Another characteristic of the neurons is the *transfer function f*. There are four most commonly used transfer functions, the *hardlimit*, the *pure linear*, the *sigmoid* and the *tansigmoid*. There are also three types of layers the *input layer*, the *hidden layers*, and the *output layer*. Each network uses one input and one output layer. The number of hidden layers can vary from 0 to any number. The only layer that does not contain transfer functions is the input layer. A Neural Networks example of two hidden layers is cited in the figure 1 below [4].

This network is briefly described by the string: 'R-S1-S2-S3', which suggests that the input layer consists of R different inputs, two hidden layers with S1 and S2 neurons respectively and the output layer with S3 neurons. In order to define a network's weights, the transfer function of each neuron should be defined first. The weights that are generated need further adjustment which is achieved with the *training* of the Neural Networks data creating a set of samples known as the *training set*. As such, the generalization on unseen samples is enabled.

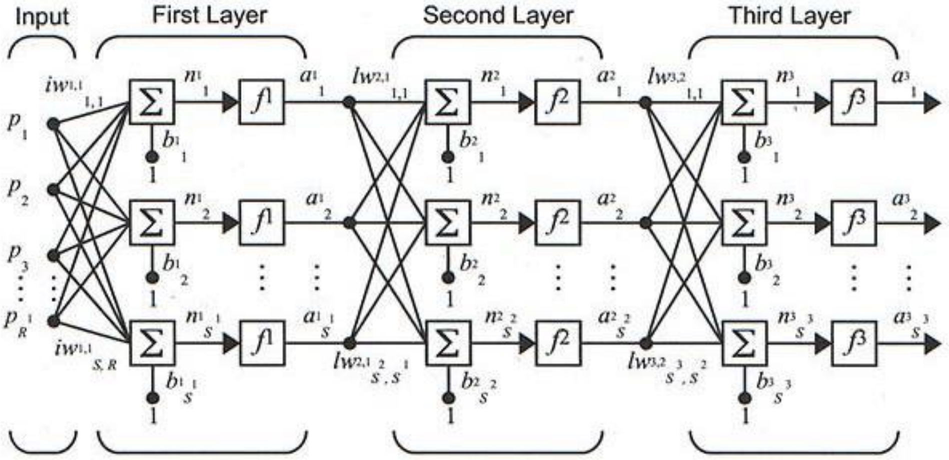


Figure 1: Two hidden layer Neural Network.

2.1.3 Algorithm Benefits

Once again, we provide the benefits the non-linear regression algorithm will offer. The goal is to transcend the efficiency of the existing data mining methods and lead to the next scale presenting a worthwhile and reliable technique in this demanding field.

According to the literature review a daily basis prediction can be achieved with models like the *Neural Networks* and the *Traditional Time Series*. However, it is required to list their pros and cons thus to determine the most suitable as cited below:

Neural Networks:

- Test both linear and non-linear patterns.
- High computational cost.
- Demand many parameter settings to increase their performance.

Traditional Time Series Models:

- Low computational cost.
- Difficulty in testing non-linear patterns.
- Demand less parameter settings to perform.

Since our objective is to test non-linear patterns it seems that the *Traditional Time Series prediction models'* efficiency is not recommended for this specific task. Hence, the use of *Neural Networks models* is a better option in attempting to test these patterns due to their higher efficacy in testing non-linear patterns.

3 Requirements/Analysis

This chapter aims at providing a short summary of the various elements and data linked to startups and stocks respectively which need to be further analyzed. It is an overview of investment theories, data related to startups and stocks in terms of functional or non functional, the predictability of these markets and already applied predictive techniques as well as the variables that influence startup growth and the fluctuations of stocks.

3.1 Investment Theories

An investment theory is a concept that takes into account various factors associated with the process of investments. Ideally, it comprises a thorough examination of a wide range of parameters in order to determine the right investment choice given a specific goal. Moreover, the theory attempts to assess investments based on certain criteria such as the risk they entail and at what extent it is acceptable as well as the potential amount of return. Simply put, an investment theory is about making informed decisions when it comes to investing. Investment theories are intertwined with the stock market however, they offer substantial knowledge to startups especially, if they are going to become huge success.

There is several investment theories regarding economics however, the major are two: the *Firm Foundation* and the *Castles in the Air* which was presented by John M. Keynes, a British economist. The two theories have contradictory connotation regarding the stock market. The *Firm Foundation* uses past data to determine stock prices and suggests that all assets have intrinsic value while on the other hand *Castles in the Air* is behavioral oriented meaning that it takes into consideration the actions of other investors in order to gain profit.

Both theories are bound to the *Efficient Market Hypothesis Theory* which suggests that prices vary and that all the relevant information need to be available to the investor thus, to make the market truly efficient. As the theory suggests if we can determine the true value of a stock then the *Firm Foundation Theory* holds otherwise, if we cannot predict the intrinsic value then the *Castles in the Air Theory* is the one that holds. The latter one seems to be the most prevailing. According to Keynes investments are made until "there is no longer any class of capital assets of which the marginal efficiency exceeds the current rate of interest". In simple words this means that investments take place until the net present value equals to zero.

Setting aside the *Keynesian Theory* several contemporary investment theories have emerged. Such theories are *Jorgenson's Neoclassical Theory of Investment*, *Clark's Accelerator Principle* and *Tobin's Q-Theory of Investment* while all incorporate diverse elements of past theories. All the above mentioned theories make assumptions on optimizing the investor's behavior.

The *Neoclassical Theory of Investment* assumes the maximizing of profit or present value of a firm in order to deliver a capital stock that is optimal. *Clark's Accelerator Theory* reduces the price variables to constants meaning that the output is delivered as a proportional outcome of the optimum capital stock. Finally, *Tobin's Q-Theory* serves as a solution to constraints that the former two theories induce providing a liaison of stock fluctuations with a company's decision making on investing. In other words, according to the Q-Theory, the share prices which are issued by a company actually, reflect its investment decisions.

3.1.1 Stock Market Predictability

The Stock Market predictability is an issue that agonize scholars and businessmen for decades with hundreds of researches made to prove it. As investment theories suggest predictions are depended on the *Efficient Market Hypothesis* which states that information relevant to a specific market should be included in the price and be accessible in order to present the market as efficient as possible. If the market proves to be efficient enough, then there is no need of predictions.

The *Efficient Market Hypothesis Theory* is stated in three forms. The first one is strong form efficiency and implies that the market is efficient as the prices reflect all relevant information, meaning that all risks would be eliminated. The second one is the semi-strong form efficiency that reflects only public information within the market while the last one, the weak form efficiency does not reflect any information in prices and investment strategies cannot be successfully deployed. The latter two efficiency forms tend to be more common regarding the stock market as even if more and more information become available and prices constantly fluctuate it is extremely difficult to reflect the true value of a stock in the current market price.

In order to determine the common stock value, the expected growth rate, the dividend payout, the risk level and the interest rate have to be determined first. The growth rate is the rate at which shareholder dividends are expected to increase and more generally suggests that current prices of shares comprise the present value all future dividends enclose.

Several studies have taken place to prove the predictability of the market yet there is no clear evidence to justify that. However, there are some points to dwell on and exploit in order to enhance this direction with the development of a non-linear regression algorithm.

3.1.2 Data Related to Stocks

As aforementioned, information that pertains to a specific market in order to acquire essential information derives from the study of relevant to the market data. The aim is to group the stock related data into categories and test them in terms of applicability. According to the literature there are four main categories that classify stock market data. These are fundamental data, technical data, market sentiment data and PESTEL analysis data.

Fundamental data are indicative of the intrinsic value of the stock. These data provide knowledge regarding the fundamental aspect of stocks. Each stock has its own value which scales up or down depending on various factors. If a stock is considered to be a high asset for its shareholders, then its intrinsic value increases and vice versa affecting dividends and stature. Along with that, there are some metrics with a huge impact on shares value and price.

Earnings per share or alternatively free cash flow per share is a fundamental metric to determine the value of a stock. The first has a more general concept that is earnings oriented taking into account the amount of money that is actually produced by each share, while the latter one describes a more specific metric in terms of money that increase a company's cash flow capabilities.

Another significant metric is the P/E ratio i.e. Price per Earnings ratio which is a type of valuation multiples that vary depending on the addressable market, the true value of a stock, its growth and stock discount rate. The last two rates are affected by inflation which is more of a technical factor in a way that when inflation is high, then the discount rate is also high while in parallel, the stock price lowers.

Technical data are all those exclusively referring to stocks including the closing price of a stock, its lowest and highest peak throughout a trading day as well as the volume of the stocks that was traded during this particular day.

Technical data are affected by several factors such as inflation and deflation. Low inflation drives interestingly high valuation multiples while high inflation rates drive low multiples. Deflation on the other hand, conveys a loss in the power of pricing which is undoubtedly a negative impact. Technical data can be derived from the understanding of the economic scales within the market as there is a 90% correlation among peers and the market itself. Additionally, potential product or service substitutes on a global stage could provide significant information.

Another source of technical data is transactions that occur incidentally not being motivated by the intrinsic value of stocks. Such transactions may impact a company's supply chain and its product or service demand, which impact on the stock demand as well.

Trends are pertinent to the above, except they are not incidental but, instead, they have to do with a company gathering the momentum a stock price may entail, or with a company suffering from a potential reversion the trend itself may have caused.

Demographics serve as technical data as well. To be more specific that kind of demographics are addressing to the valuation multiples in correlation with the amount of the middle-aged investors. The higher the number of middle-aged investors, the higher the valuation multiples will be. The exact opposite happens when one of them is low as the equation is proportional.

Eventually, liquidity and liquidity discount are extremely valuable. In order for a stock to be easily transacted and create demand it has to convey high liquidity. Large capital stocks show high liquidity comparing to small capital stocks.

Stock Market sentiment data are essentially important nowadays and have to be exploited. This particular term includes all psychological and social factors that have an impact on the market stakeholders. In simple words describes the crowd psychology that is revealed through diverse stock trades occurred within a specific day in the market. It actually explains and deploys behavioral finance and social science as it refers to attitude investors may demonstrate. An example is the investors' persistence to mistakes, their tendency in demonstrating greater pain when they sustain losses contrary to when they have gains as well as data over emphasizing.

Market sentiment when it comes to stocks is something between the fundamentals and technical data using indicators such as price changes created by the investors. It has two categories, the bearish when stock prices go downward and the bullish when prices go upward. The rationale behind bulls and bears is to sell when a particular market is overbought (bullish) and buy when it is oversold (bearish). Simply put it is all about emotions and not business performance.

PESTEL analysis data are those data that stem from the macro-environment analysis. The macro-environment contains all the political, economic, social, technological, environmental and legal factors that affect the various markets. Since, the fact that these factors cause surprising results on stock price fluctuations is quite evident that trying to understand, assess and predict them could prove rather advantageous. Financial instability, unemployment, elections, extreme weather phenomena, technological developments and energy resources are some of these factors that provide data so that we can utilize mining techniques in order to attempt on extracting valuable information. We have witnessed, among other macro-environmental changes, a serious collapse of the stock exchanges all over the world when financial crisis and recession occurred, or a country’s debt has been extremely enlarged or a powerful currency had sustained depreciation.

4 Design

In this chapter we analyze the various parameters and data that are used for this project. To be more specific, several aspects regarding data such as data description, quality and origin for stocks are separately taken into consideration. Consequently, the process is to construct and format these data, thus, to be applicable on implementing the algorithm. Additionally, a data randomness test will take place regarding their efficiency to the stock market. Eventually, we will attempt to explain the solution to the problem.

4.1 The Stock Market

4.1.1 Stock Data Analysis

The objective at this part is to determine the most suitable datasets depending on their origin, quality and description thus to construct the excess returns time series so that they will be compatible before tested for randomness.

4.1.2 Stock Data Collection

DataStream International [5] is the data source used for the project. More specifically the London and the New York stock markets are taken into consideration along with their respective indices FTSE-500 and the S&P-500. The formation of the data series requires the following time series: FTSE-100 index, Treasury-Bill Rates of UK, S&P-500 index, Treasury-Bill Rates of US. The FTSE-500 data consist of 3122 daily observations of the index and the T-Bill rates for the same period and frequency. The UK T-Bill rates are annually scaled with one-month maturity. The S&P-500 data refer to the value of the index on daily basis and provide 3125. The US T-Bill rates cover the same period and frequency however they are scaled with a maturity of thirteen weeks.

4.1.3 Stock Data Description

This chapter contains detailed description of each one of these series as presented on the following table as well as a graph of FTSE 500 below including its values over time.

Table 1: FTSE 500, UK T-Bill rates, S&P 500 and US T-Bills metrics

Metrics	**FTSE 500**	**S&P 500**	**UK T-Bills %**	**US T-Bills %**
Standard Error	11.83135	7.99835	0.05842	0.028012
Mean Error	1856.1023	781.0135	7.95838	5.44365
Median Error	1661.4	546.04	6.7204	5.24
Standard Deviation	681.0056	471.004578	3.22598	1.510578
Sample Variance	459892.75	221068.2248	9.98048	2.190034
Standard Range	2459.98	1643.32	10.8675	6.56
Minimum Range	986.12	282.12	5.0024	2.89
Maximum Range	3389.85	1921.25	14.8945	9.34
Observations	3122	3125	3122	3125

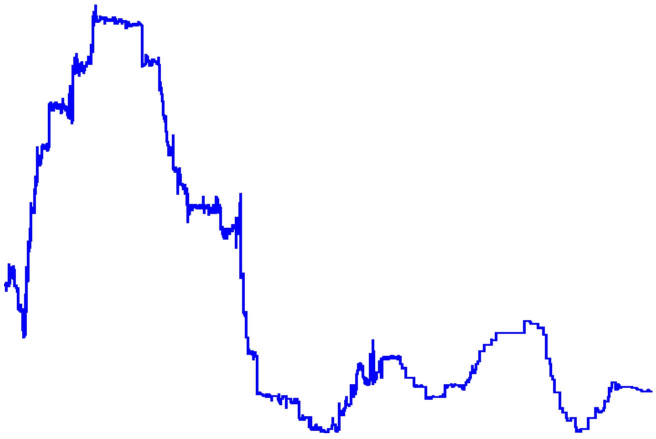


Figure 2: FTSE 500 time series.

Regarding the FTSE 500 and S&P 500 it is obvious that they show an upward trend while the T-Bill rates show a backward one suggesting that there is a correlation (interest rates decrease causes stock increase) among the stock market value and T-Bill rates.

4.1.4 Stock Data Quality

Data quality examines whether a dataset contains missing values or not. These missing values are the so-called *outliers* [6] and are different from the general ones and behave inconsistently.

In order to check for *outliers,* we calculate the first (Q1) which is the 25-th and the third (Q3) which is the 75-th quartile of the distribution our data. To continue with an xp value is called the k-th percentile of a given distribution if P(X<xp)=k/100, where X is a random variable [16] meaning that the 25-th value that creates two dataset subsets which contain 25% and 75% of the mass of the samples respectively while, the 50-th value is the *median* value.

Each set requires a calculation of the Q1 and Q3 values in order to calculate the quantity Q3-Q1 while considering any value greater than Q3+3(Q3- Q1) or lower than Q1-3(Q3-Q1) an *extreme outlier*.

4.1.5 Stock Data Construction

Here we try to construct the data according to the stock market returns provided from the indices. The equation that defines the returns is the following:

$p_t - p_{t-1} / p_{t-1}$ **(1)**, where p_t refers to day t.

Furthermore, another equation that is needed is: $b_t = 1/100 * rate_{t-1} / 360$ **(2)**, where the ratet-1 refers to the annual value of the T-Bill rates on day t-1.

As a next step we set $1/100 * rate_{t-1}$ equal to c_{t-1} in order to combine the two equations presented above and form a third one which is

$b_t = c_{t-1}/360$ **(3)**. As a result, by combining equations **(1)** and **(3)** we get the excess return which is $R_t = r_t - b_t$.

Due to the fact that the r_t value demonstrates both positive and negative values it means that a slight change should take place by adding 1 to r_t scale value. That change leads to a new equation which is:

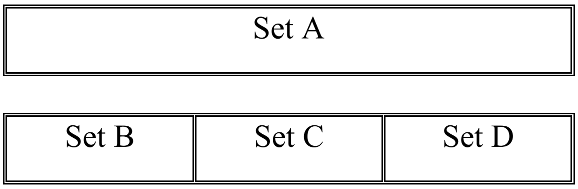
$$1 + r_t = 1 + p_t - p_{t-1} / p_{t-1} = p_t / p_{t-1} \textbf{ (4)}$$

As such the logarithmic and the T-Bill rates results are $\ln(p_t / p_{t-1})$ and $\ln(c_{t-1}/360+1)$ respectively providing us with a fifth equation which follows:

$$y(t) = \ln(p_t / p_{t-1}) - \ln(c_{t-1}/360+1) \textbf{ (5)}$$

4.1.6 Stock Data Formation

Due to the nature of the *Neural Networks* parameters a further data split is required meaning that there is need in categorizing it into three subsets as the figure below indicates.



Set B will be the *Training I* set, Set C will be the *Validation I* set, and Set D will be the *Validation II* set respectively.

4.1.7 Stock Data Randomness

Regarding to randomness there is need in taking into consideration potential random sequences. Such sequences prevent the system under use from the ability to predict the underlined sequence. Hence, there is a point in proving that stock data fluctuations are not randomly generated in order to present a feasible prediction.

As excluded from the literature *theoretical* and *empirical* tests took place to test randomness. The first category deals with number groups of a sequence and evaluate specific results-values while, the second category exploits the sequence according to the recurrence rule.

Below an empirical test is given with the use of a run test which is a group of consecutive symbols of one kind preceded and followed by symbols of another kind [7].

Next step is to calculate the median value of the *excess returns* of stocks assigning the symbol '+' for values above the median and '-' for the ones below. As a result, a new series S will be created containing '+'s and '–'s values. Moreover, we define as $+r$ and $-r$ the number of runs that contain '+' and '-' respectively and r to be equal to $+r + -r$. According to research a random series S, r is approximately normally distributed with mean:

$$E(r)=m+1 \quad \textbf{(6)}$$

and variance:

$$var(r)=m(m+1)/2m-1 \cong 1/4\ (2m-1) \quad \textbf{(7)}$$

The FTSE and S&P indices *training sets* contain 2986 and 2988 samples respectively. For the FTSE *training set* we found rFTSE =1467 and mFTSE =1493, while for the S&P we found rS&P=1489 and mS&P =1495. Hence, equations **(6)** and **(7)** indicate that the number of runs in a sequence between 2986 and 2988 will be normally distributed presenting a mean of 1495 and variance of 1489.

5 Methods

The Methods Description chapter attempts to describe the methods and the process that was deployed in order to test the algorithms as well as the functions and other popular techniques. This part is a detailed description and explanation of all the steps which were incrementally made regarding the several forecasting models.

5.1 Neural Networks

Neural Networks as their name suggests imitate human brain function consisting of interconnected processing units which are called *neurons* or *nodes*.

These neurons use a number of inputs and provide a distinct output which is shown in the figure 3 below:

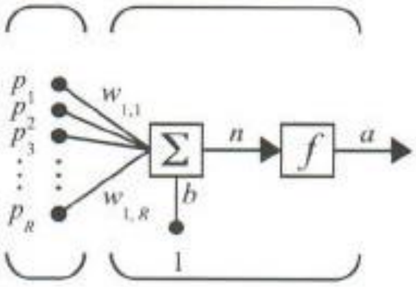


Figure 3: Neuron with R inputs.

Figure 3 presents a single neuron with R inputs p_1, p_2, …, p_R, where each input is weighted with a value w11, wl2 , …, wlR respectively while the output of the neuron *a* equals to *f*(w11 p1 + w12 p2 + … + w1R pR).
Furthermore, a *transfer function* called *f* characterizes each *neuron*. The most commonly used transfer functions are the *hardlimit*, the *pure linear*, the *sigmoid* and the *tansigmoid* function.

- *Hardlimit function* maps all values that belong within $(-\infty, +\infty)$ into $\{0,1\}$ range values mostly applied in classification:

$$f(x)=1, x \geq 0$$

$$f(x)=0, x < 0$$

$$f(x) \in \{0,1\}$$

- *Purelinear function* returns all real values and is utilized for output neurons:

$$f(x)=x$$

$$f(x) \in (-\infty, +\infty)$$

- *Sigmoid function* maps all values to the range of $[0,1]$:

$$f(x)=1/1+e^{-x}$$

$$f(x) \in [0,1]$$

- *Tansigmoid function* maps all values to the range of $[-1,1]$:

$$f(x)=2/1+e^{-2n}-1$$

$$f(x) \in [-1,1]$$

- Layers

Another important characteristic of the *neurons* is that they are distributed into *layers* and more specifically one *input layer* and one *output layer*. Input samples are bound to the number of inputs while, the number of outputs are bound to how many neurons are in the output layer. In this particular case one neuron is needed in order to represent the next day's *returns*.

- Weights

Neurons are also characterized by weights which have to be adjusted within the specific training set. Given the fact that the adjustment was accomplished, the exploitation of the network's unseen data will be feasible. In this topic there will an attempt to measure the generalization ability among the *training set* and the *test set*.

- Error Function

The *error function* or *cost function* is used to train *Artificial Neural Networks* by exploiting the chain rule. These back-propagation functions are recursive and iterative and are used to calculate the weights updates until improvement is reached in such a level that there is no further need in reducing error and the respective task can be optimally performed.

- Gradient Descent

Gradient descent is an optimization function that adjusts weights according to the error they caused.

A gradient represents how two variables relate to each other such as the correlation between the network's error and a single weight trying to identify the weight that will produce the least error. The correlation among the network *error* and each of those *weights* is a derivative, *dE/dw*, which measures to what degree a slight change in a weight leads to a slight change in the error.

This task is achieved with the use of the *chain rule* which retrains the inputs and outputs of the network thus to find the weight in question and the way it is linked to overall error:

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}. \quad \textbf{(8)}$$

In a feedforward network such as the Neural Networks, the net's error correlation with a single weight is presented in the equation below:

$$\frac{dError}{dweight} = \frac{dError}{dactivation} * \frac{dactivation}{dweight} \quad \textbf{(9)}$$

The interpretation is that given the *error* and *weight* variables mediated by a third *activation* variable it is possible to calculate how a change in *weight* affects a change in *error* calculating how a change in *activation* affects a change in *error*, and how a change in *weight* affects a change in *activation* in the first place. The ultimate goal is to adjust the model's weights until the point that the error cannot be reduced further.

5.2 Traditional Time Series

Such kind of forecasting series are widely applied in econometrics and they are known for their ability in dealing with linear tasks using the following function:

yt= f(x1t, x2t,…, xkt)= á + â1x1t + â2x2t + … + âkxkt + rst for t=1,…,N **(8)**

where, xi are the *explanatory variables,* y is the *explained variable,* âi for i=1,..,k are the *coefficients,* y1, …, yN, is the under prediction time series and rst is an *independent and identically distributed noise component.*

Traditional Time Series aim to provide a sample of N examples {(x1t,…,xkt, yt), t=1,…,N}, returning a function g that approximates f in the least vector E=(e1,…,et) error. Each et is defined as ei=e(g(x1t,…, xkt), yt) and refer to an arbitrary error function. According to the above the returning function g is formed as following:

y = g(x1t, x2t,…, xkt)=_*a*+_1 *b* x1t +_2 *b* x2t + … +_*k* *b* xkt for t=1,…,N **(9)**

where, *a*,_*i* *b* for i=1,..,k are the *estimators of the coefficients* and *y*, is the prediction for y_t.

6 Experimentation and Evaluation

As one of the most important parts of the project this section checks the effectiveness of the algorithms and assesses the work done in comparison with existing data mining methods. The algorithm will be assessed for its efficacy on Stock Market applications.

6.1 Experimentation -Testing

Here we test several algorithms in order to serve the present study. Testing is of paramount importance as it demonstrates the algorithms functionality and issues which may potentially occur and will need to be fixed, thus, to improve the algorithm.

- Testing with Neural Networks

The first test deals with stock returns time series through *Neural Networks*. As mentioned previously Neural Networks are known for their variables complexity contrary to other models.

Initially, we used a genetic algorithm with specific metrics such as *TheilA*, *TheilB*, *TheilC* and Mean Absolute Error (*MAE)* which is a variant metric, in order to generate the most suitable networks, a procedure iterated three times to provide optimal results.

Secondly, we utilize the output of the first step by training and validating the network using the *training I set* and the *validation I set*. Moreover, the network was tested on unseen data exploiting the *validation II set*. Due to the complex nature of the Neural Networks the test was repeated several times, so that it would facilitate the selection based on the results of standard deviation values which ultimately provided four different structures.

Lastly, the four structures were trained in half splits for the *training set* and the *validation set* before retraining the structures for the complete *training set* and testing all the metrics regarding the standard deviation and mean values each resulted in.

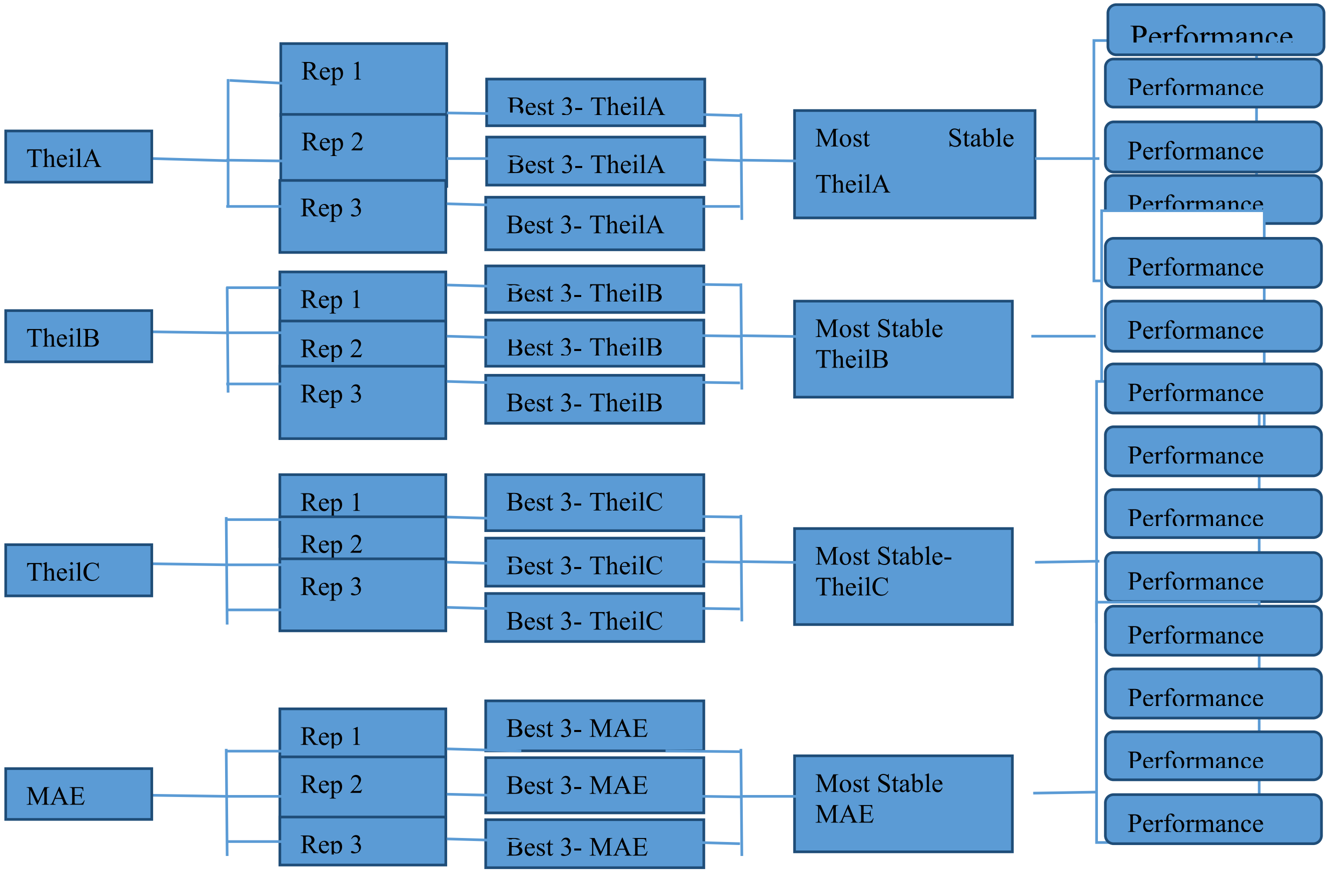


Figure 4: Neural Networks Test

Regarding each genetic algorithm specific maximum values were given to each three of the axes, in particular *xmax=25*, *ymax=25* and *zmax=30* in order to decrease potential loss of optimal structures.

To continue with, another set of variables was incorporated. These are the following: $P_{reproduction}$, $P_{crossover}$, $P_{mutation}$, *maxGen* and *m*. The first three are the probabilities of selecting reproduction, crossover and mutation operation, while *maxGen* is the termination criterion and *m* is the number of chromosomes per generation. The values of these parameters are highly correlated [8] as the larger the population size, the less possible is for crossover and mutation to occur. In this task the values that were applied are *m=35* and *maxGen=25*, $P_{crossover}$=0.7, $P_{mutation}$=0.09 and $P_{reproduction}$=0.35 respectively.

Furthermore, the selection of the size of the sets (*Training I*, *Validation I* and *Validation II*) was performed by defining them as:

a=*size of (Training I Set) / size of (Training Set)*

b= size of *(Validation I Set) / size of (Training Set)*

c= size of *(Validation II Set) / size of (Training Set,)*

so that the equation a+b+c=1 is satisfied.

In order to measure the FTSE index using the genetic algorithm 3 repetitions took place exclude results for the TheilA metric, TheilB metric, TheilC metric and MAE.

The results from the three repetitions with the use of each of the metrics for the FTSE 500 are cited on the tables below:

	Rep 1	**Rep 1**	**Rep 2**	**Rep 2**	**Rep 3**	**Rep 3**
Generation	**first**	**last**	**first**	**last**	**first**	**last**
Average **Standard**	1.049023 0.049675	1.011208 0.023934	1.040058 0.045094	1.010995 0.021823	1.046815 0.044531	1.005997 0.013988

Table 2: Repetition 1,2 &3 for FTSE using TheilA metric

Generation	Rep 1 first	Rep 1 last	Rep 2 first	Rep 2 last	Rep 3 first	Rep 3 last
Average	0.788657	0.718305	0.769896	0.739118	0.792293	0.736364
Standard	0.049915	0.011039	0.045117	0.011994	0.149557	0.015388

Table 3: Repetition 1,2 &3 for FTSE using TheilB metric

Generation	Rep 1 first	Rep 1 last	Rep 2 first	Rep 2 last	Rep 3 first	Rep 3 last
Average	1.140043	1.009858	1.039758	1.011036	1.052894	1.004896
Standard	0.040385	0.018994	0.039985	0.016114	0.079533	0.013218

Table 4: Repetition 1,2 &3 for FTSE using TheilC metric

Generation	Rep 1 first	Rep 1 last	Rep 2 first	Rep 2 last	Rep 3 first	Rep 3 last
Average	0.009129	0.008996	0.009034	0.008867	0.009335	0.008959
Standard	0.000368	0.000159	0.000265	0.000087	0.001204	0.000233

Table 5: Repetition 1,2 &3 for FTSE using MAE

To continue with, in order to measure the S&P index using the genetic algorithm we iterated the process repeating three times for each one of the metrics, TheilA, TheilB, TheilC and MAE. The results for each one of the metrics for the S&P 500 are presented on the following tables:

Generation	Rep 1 first	Rep 1 last	Rep 2 first	Rep 2 last	Rep 3 first	Rep 3 last
Average	1.033235	1.009128	1.044998	1.000923	1.039975	1.039996
Standard	0.039115	0.039982	0.087794	0.019825	0.049733	0.024766

Table 6: Repetition 1,2 &3 for S&P using TheilA metric

Generation	Rep 1 first	Rep 1 last	Rep 2 first	Rep 2 last	Rep 3 first	Rep 3 last
Average	0.699823	0.710408	0.721158	0.698995	0.709745	0.698897
Standard	0.036574	0.019133	0.043394	0.024898	0.043966	0.022918

Table 7: Repetition 1,2 &3 for S&P using TheilB metric

Generation	Rep 1 first	Rep 1 last	Rep 2 first	Rep 2 last	Rep 3 first	Rep 3 last
Average	1.168234	1.003029	1.037558	1.003943	1.033865	1.001289
Standard	0.047166	0.029232	0.058193	0.021945	0.045163	0.019574

Table 8: Repetition 1,2 &3 for S&P using TheilC metric

Generation	Rep 1 first	Rep 1 last	Rep 2 first	Rep 2 last	Rep 3 first	Rep 3 last
Average	0.011345	0.012093	0.010893	0.012876	0.010664	0.010318
Standard	0.000397	0.000145	0.000314	0.0002995	0.000468	0.000768

Table 9: Repetition 1,2 &3 for S&P using MAE

After the evaluation of the fittest structures was completed by the *Genetic Algorithm,* we determined the most suitable ones thus to exploit for both FTSE 500 and S&P 500 indices that are under examination. The outcome of this evaluation is presented on the next table:

Metrics	FTSE 500	S&P 500
TheilA	4-2-2-1	1-2-2-1
TheilB	6-4-2-1	1-3-1-1
TheilC	2-3-2-1	5-1-2-1
MAE	7-3-1-1	4-2-2-1

Table 10: Suitable structures for FTSE 500and S&P 500

- Testing with Auto-Regressive models

For this evaluation in order to construct the model, the only variable needed is the lag structure as well as the parameter adjustment within the complete *Training Set* which includes both the *Training I Set* and *Validation I Set* thus to proceed to the *Test Set* to measure the model's performance. The adjustment of the parameters generated the following results for the FTSE 500 and S&P 500 respectively:

$$AR_{FTSE}\,1\text{: } y_t = 0.0001125 + 0.098737\, y_{t-1} \quad \textbf{(10)}$$

and

$$AR_{S\&P}\,1\text{: } y_t = 0.0004198 - 0.0069721\, y_{t-1} \quad \textbf{(11)}$$

$$\begin{aligned} AR_{S\&P}\,7\text{: } y_t = {} & 0.0005176 - 0.0051695 y_{t-1} \quad \textbf{(12)} \\ & -0.016112 y_{t-2} - 0.042995 y_{t-3} \\ & -0.020708 y_{t-4} - 0.048967 y_{t-5} \\ & -0.022141 y_{t-6} - 0.044298 y_{t-7} \end{aligned}$$

Once again, the metrics that are going to be applied are: *TheilA*, *TheilB*, *TheilC* and MAE. In this particular case the evaluation depends upon the *Test Set*. The results for the two indices are cited in the following tables:

Metrics	FTSE 500 Lag (1)
TheilA	1.00016204516712
TheilB	0.71899971010445
TheilC	1.00013412203813
MAE	0.00864008078906

Table 11: AR_{FTSE} 1 Evaluation

Metrics	S&P 500 Lag (1)	S&P 500 Lag (7)
TheilA	1.00089956684997	1.00020900988165
TheilB	0.69997880671278	0.70077821675661
TheilC	1.00092460833382	1.00019234730079
MAE	0.01100721810125	0.01021113794560

Table 12: $AR_{S\&P}$ 1 and $AR_{S\&P}$ 7 Evaluation

Eventually, as far as the Auto-Regressive models are concerned we should check whether there are unseen data neglected by the model because of the linearity they convey. The BDS test is the most appropriate to check on unseen data utilizing the equations that were previously stated:

$$AR_{FTSE}\ 1: y_t = 0.0001125 + 0.098737\ y_{t-1} \quad \textbf{(10)}$$

and

$$AR_{S\&P}\ 1: y_t = 0.0004198 - 0.0069721\ y_{t-1} \quad \textbf{(11)}$$

$$\begin{aligned} AR_{S\&P}\ 7: y_t = {} & 0.0005176 - 0.0051695y_{t-1} \quad \textbf{(12)} \\ & -0.016112y_{t-2} - 0.042995y_{t-3} \\ & -0.020708y_{t-4} - 0.048967y_{t-5} \\ & -0.022141y_{t-6} - 0.044298y_{t-7} \end{aligned}$$

After having applied the BDS test for each of the equations we gained several and multiple results which are mentioned in separate tables for the evaluation of each one the indices (AR_{FTSE} 1, $AR_{S\&P}$ 1 and $AR_{S\&P}$ 7).

The following table presents the outcome delivered by the BDS test applied on the autoregressive model AR_{FTSE} Lag 1 with regards to FTSE 500 index:

å	M	BDS	å	M	BDS
0.25	2	4.3015	1.00	2	5.1399
	3	6.6962		3	8.0902
	4	8.4122		4	9.6299
	5	9.6554		5	10.9186
	6	7.2879		6	12.3345
	7	2.3142		7	14.2998
	8	-2.7798		8	16.3423
	9	-9.3509		9	18.0122
	10	-7.7247		10	20.1003
0.5	2	4.4898	1.25	2	5.8111
	3	7.4906		3	8.1286
	4	9.1822		4	10.2004
	5	10.8943		5	11.2998
	6	12.5987		6	12.5773
	7	15.6004		7	14.3008
	8	17.9082		8	15.6119
	9	18.6767		9	17.0058
	10	17.2945		10	18.6565
0.75	2	4.7887	1.5	2	6.6771
	3	7.7212		3	9.5993
	4	9.3454		4	11.0032
	5	10.4769		5	12.0104
	6	11.8901		6	13.2242
	7	14.0015		7	14.4452
	8	16.1114		8	15.5164
	9	20.6967		9	16.4885
	10			10	17.6007

Table 13: BDS Test on AR_{FTSE} 1

A BDS test functions using a normal distribution of (0,1). The results presented on the table strongly suggest that there are unseen data that were not considered by the model.

The following table presents the outcome delivered by the BDS test applied on the autoregressive model $AR_{S\&P}$ Lag 1 with regards to S&P 500 index:

å	M	BDS	å	M	BDS
0.25	2	4.9004	1.00	2	5.1989
	3	7.1087		3	7.2958
	4	8.4979		4	8.4987
	5	10.2994		5	10.5777
	6	11.9992		6	12.5948
	7	13.6785		7	14.7995
	8	9.8989		8	16.9994
	9	5.6021		9	19.4009

	10	-5.0009		10	22.2123
0.5	2	4.5103	1.25	2	5.4332
	3	6.5776		3	7.6657
	4	8.2006		4	8.8356
	5	10.4141		5	10.5505
	6	12.3425		6	12.3208
	7	14.9202		7	14.0996
	8	17.1122		8	15.9043
	9	20.0123		9	17.6655
	10	24.1909		10	19.7763
0.75	2	4.5978	1.5	2	5.8102
	3	6.6576		3	7.9003
	4	8.1134		4	8.9546
	5	10.3906		5	10.5675
	6	12.4995		6	11.9874
	7	15.0194		7	13.4545
	8	17.4225		8	14.8008
	9	20.3996		9	17.4997
	10	23.9676		10	

Table 14: BDS Test on AR$_{S\&P}$ 1

According to the table above it is evident that using the BDS test on AR$_{S\&P}$ 1 the values that were provided are more extreme in comparison with the AR$_{FTSE}$ 1. Consequently, this leads to the conclusion that there are missing patterns.

The following table presents the outcome delivered by the BDS test applied on the autoregressive model AR$_{S\&P}$ Lag 7 with regards to S&P 500 index:

å	M	BDS	å	M	BDS
0.25	2	4.9002	1.00	2	5.0008
	3	7.2234		3	7.1982
	4	8.9102		4	8.5654
	5	10.6978		5	10.4978
	6	14.3896		6	12.4325
	7	17.7289		7	14.5634
	8	14.0967		8	16.7656
	9	19.4989		9	19.1003
	10	-5.2224		10	22.0003
0.5	2	4.3978	1.25	2	5.2234
	3	6.4647		3	7.3342
	4	8.0998		4	8.8112
	5	10.2231		5	10.4747
	6	12.1134		6	12.0788
	7	14.4955		7	13.9443
	8	16.7222		8	15.5053
	9	19.4132		9	17.2228
	10	22.7979		10	19.10276
0.75	2	4.7007	1.5	2	5.4969
	3	6.8224		3	7.5153

4	8.4044	4	8.9987
5	10.4135	5	10.3698
6	12.5034	6	11.6957
7	14.8232	7	13.0923
8	17.1115	8	14.3144
9	20.0991	9	15.6995
10	23.6972	10	17.0199

Table 15: BDS Test on $AR_{S\&P}$ 7

There is no big difference in the results presented on this table for the BDS test on $AR_{S\&P}$ 7 comparing to those of $AR_{S\&P}$ 1.

As a conclusion we can suggest that the higher the lag is the greater the possibility of correlation among data. Moreover, it is obvious that according to the BDS test there are unseen patterns neglected by the methods used.

6.2 Benchmarking

Aforementioned, testing and evaluation of the algorithm will be applied comparing to other techniques. This is a way to gain insight of the differences the developed algorithm encapsulates when it comes to data mining techniques benchmarking.

6.2.1 Neural Networks Model Conclusions

The conclusions that we have reached by testing the *Neural Networks* method are according to the TheilA metric the model did not succeed in overcoming models that are using the *Random Walk* while, as well as models which suggest that the market’s tomorrow value will be as beneficial according to TheilC metric. Regarding the stock’s excess returns it appears that the model can overcome the respective *Random Walk* model.

6.2.2 Auto-Regressive Model Conclusions

On the other hand, the conclusion deducted from *Auto-Regressive* models is that they did not succeed in capturing unseen data as the BDS test proved. Taking into consideration that the *Auto-Regressive* models are not capable of examining non-linear patterns it proved to be correct to also test the *Neural Networks* model which are efficient enough for both linear and also non-linear structures.

6.3 Project Evaluation

The anticipated outcome is to provide greater insight on data mining methods to observe which ones perform better in predicting more accurately what was developed for. Therefore, a final project evaluation is necessary.

Taking into account the results from the previously stated experimentation we can safely suggest Neural Networks proved to be a better option due to their twofold usage contrary to *Auto-Regressive* models which struggle to deal with non-linear structures. It should also be mentioned that weights play a significant role on how *Neural Networks* perform and this is the reasoning behind the *standard deviation* and the *mean* selection.

7 Conclusions and Further work

7.1 Conclusions

In this project the attempted task was to predict the daily stock return values. In particular, an examination of the daily excess returns of London's FTSE 500 index as well as the daily excess returns of New York's S&P 500 index along with the exploitation of the Treasury-Bill rates took place.

In order to achieve that, we used lagged values while two randomness tests (Run and BDS tests) were applied thus to prove the feasibility of the project.

According to research made we concluded that the fit models to serve the underlying cause were the *Neural Networks* and *Auto-Regressive* models. Moreover, it should be mentioned that the lag value for the FTSE 500 is 1 while for S&P 500 it ranges from 1 to 7. The performance of the applied models was measured utilizing the *Mean Absolute Error (*MAE).

The conclusion of the findings suggests that even though *Neural Networks* models outperform *Auto-Regressive* models it is not feasible to improve their efficiency in prediction regardless of the variable selection. The reasoning behind this is that daily data use does not help the models to identify unseen data and patterns, meaning that the noise of these data is high, creating a difficult task to overcome.

7.2 Further Work

As far as data is concerned, the data input should be identified in such a way that their space dimension is reduced to the extent that they continuously provide the appropriate information to the models that are used in experimentation.

Furthermore, the data examined should range within narrow periods because of the internal and external factors that influence the way the stocks fluctuate and function on daily basis. In other words, the models we exploit should be able to adjust on the suitable per time variables, thus, to enable optimal predictions.

Finally, due to its complexity the data of the *Stock Market* are extremely noisy, not to mention on a day by day basis. As such, attempts should be made on reducing the intrinsic noise of these data by trying to determine whether data on weekly or even monthly basis respond more optimally or not.

Bibliography

[1] Malkei B. G. (1999, 7th ed.). *A random walk down wall street.* New York, London: W. W. Norton & Company.

[2] Maddala G.S. (1992). *Introduction to econometrics.* New York, Toronto: Macmillan Publishing Company.

[3] Azoff E. M. (1994). *Neural network time series forecasting of financial markets.* Chichester: John Wiley and Sons.

[4] Tsibouris G. & Zeidenberg M. (1996). Testing the efficient market hypothesis with gradient descent algorithms. In Refenes, A. P. *Neural networks in the capital markets*. England: John Wiley & Sons Ltd, pp 127-136.

[5] DataStream web site http://www.primark.com/pfid/index.shtml?/content/ datastream.shtml

[6] Han J. & M. Kamber (2001). *Data mining: concepts and techniques*. San Francisco: Academic Press.

[7] Lindgren B. W. (1976). *Statistical theory*. 3rd edition. N.Y., London: Macmillan

[8] Man F. K., Tang S. K. & Kwong S. (1999). *Genetic algorithms: Concepts and designs*. Heidelberg: Springer-Verlang.

[9] Christantonis K. & Tjortjis C. (2019). Data Mining for Smart Cities: Predicting Electricity Consumption by Classification, IEEE 10th Int'l Conf. on Information, Intelligence, Systems and Applications (IISA 2019).

[10] Papas D. & Tjortjis C. (2014). Combining Clustering and Classification for Software Quality Evaluation, LNCS 8445, pp. 273-286, Springer-Verlag