BABEŞ-BOLYAI UNIVERISTY CLUJ-NAPOCA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
SPECIALIZATION COMPUTER SCIENCE

**Diploma Thesis**

# STOCK MARKET PREDICTION

Supervisor:
**Lect. Dr. Mihoc Tudor**

Author:
**Flaviu-Andrei Jurj**

2020

# Abstract

Stock market has been a studied domain for a long time especially due to the attention received from investors. More specifically, stock market prediction is known for its complexity and volatility. Initially, the task was done solely by humans, but lately with the technological evolution more and more operations are automated including the actual prediction.

In the beginning of artificial intelligence, basic methods were used to complement statistical methods. Moving forward through the timeline more methods were tested and as with most domains in which artificial intelligence was used the current standard involves more or less machine learning.

This thesis will approach the stock market prediction problem as a time series forecasting one. It will be seen from the two classic perspectives: as a classification problem predicting only a general trend such as increase, decline or stagnation and as a regression problem predicting a set of prices during different time horizons. The problem will be solved from a deep learning perspective experimenting with different architectures which will mainly consist of specific variations of recurrent neural networks which have proven their effectiveness in other problems from other domains: long short-term memory networks and gated recurrent unit networks.

# Contents

# List of Tables

# List of Figures

# List of Algorithms

# Chapter 1

# Introduction

The stock market can be defined as a loose network of economic transactions of stocks (also called shares) which represent ownership claims on businesses. The act of investing has changed a lot during recent years becoming more and more automated through various electronic platforms which even try to predict the actual movement of the market.

There are many strategies which may be adopted by an investor, but ultimately they reduce to the most important task: the capability of predicting to some degree the movement of the market. In this thesis, we will evaluate the reliability of machine learning for tackling this problem. Technical analysis seeks to determine the future price of a stock based only on the previous trends of the price without taking into account details and fundamentals of the company.

# Chapter 2

# Theoretical Fundamentals

## 2.1 Stock Market

### 2.1.1 Stock Market Fundamentals

The stock market refers to the collection of markets and exchanges where regular activities of buying, selling and issuance of shares of publicly-held companies take place. It was created a long time ago in order to facilitate the raise of capital of companies, promoting a transparent way of trading regarding company assets. Nowadays, its main purpose is to regulate the exchange of stocks or other financial assets, ensuring a fair environment for both investors and corporations (whose stocks are traded in the market). It can be seen as the staple of the global financial system. The stock market created a dynamic system encouraging permanent innovation and improvement in every domain.

From an investor perspective, each action is the result of an investment strategy which contains a set of rules, behaviours or procedures. Although, there are many investment strategies, most of them can be classified in two separate groups: fundamental analysis and technical analysis. Fundamental analysis represents the analysis of a company's past performance as well as the credibility of its accounts through various factors and indicators based on the financial statements, business trends and general economic conditions. On the other hand, technical analysis is not concerned with any of the company's financial prospects and seeks to determine the future price of a stock based only the trends of the past price (a form of time series analysis). For the following reasons, fundamental analysis is usually seen as a long-term strategy, while technical analysis is seen as a short-term one.

This thesis will tackle the topic of stock market prediction more from a technical analysis perspective being more suitable for automated non-subjective decisions. We will further evaluate the evolution of different methods and techniques reaching the latest trends of machine learning strategies, more specifically deep learning which at the moment is not well researched for financial time series forecasting research. [6]

### 2.1.2 Historical evolution techniques in stock market prediction

Even though the stock market has a long history, only in the last couple of decades we have seen real development and research in terms of techniques to automate or aid the investor in the process of stock market prediction. At our current state, we can group all techniques in the following categories: statistical, pattern recognition, machine learning, sentiment analysis and hybrid.[7] At their core, they can be classified as mainly technical analysis, but they can also

borrow some aspects from fundamental analysis.

Prior to the emergence of machine learning, statistical techniques were used which they often assumed linearity. However, they were mostly replaced step by step with other techniques which are more and more researched especially with the exponential growth in computational power. All the other categories may be considered to some degree as machine learning, but they are split accordingly due to the fact that they usually have different goals in mind. While pattern recognition works mostly on predicting certain figures or shapes which repeat themselves with unknown periodicity, machine learning category covers many techniques which have some degree of similarity and not being different enough to represent an entire category. Sentiment analysis represents also a trending methodology which have gained a lot of momentum lately even in financial time series analysis, but it is still mostly researched for recommendation systems.

One subcategory of machine learning which haven't been intensively researched due to its only recent success in other domains is represented by deep learning. Deep learning has been a major breakthrough in many domains such as object detection, speech recognition, natural language processing and so on. This thesis will approach the problem of stock market prediction as a time-series problem using the company historic data to predict on a short-term horizon the price or the trend. We will compare the results from both classical perspectives as a regression and as a classification.

### 2.1.3 Difficulties and controversy

Stock market has been a studied domain for a long time generating debates and controversy regarding whether it is possible or not to consistently predict its movement. The prediction problem is still an open problem due to its complexity taking into account the volatility of the stock market.In the following paragraphs, we will discuss about two theories which have generated controversy among people.

Chaos theory is a branch of mathematics focusing on the study of chaos - states of dynamical systems whose apparently random states of disorder and irregularities are often governed by deterministic laws that are highly sensitive to initial conditions. One underlying principle of chaos, also called the 'butterfly effect', describes how a small change in one state of a deterministic nonlinear system can result in large differences in a later state. Chaotic behaviour exists and is mostly characterized in natural systems such as weather, climate or even heartbeat irregularities. For this reason, many consider that the weather forecast for example can be considered accurate only in the following 2-3 days.

Coming back to our domain of interest, chaos theory is seen only as a spontaneous occurrence in some systems with artificial components such as the stock market and road traffic. While there exists some research studying the effects of chaos theory in economic and financial systems, the empirical literature that tests for chaos in economics and finance presents very mixed results.[2] There is little consensus that the chaos theory may only illustrate sudden shocks and crashes of the market which happen very rarely and can be considered insignificant. During the stock market history, there existed unpredictable events called 'black swan' which were characterized by their extreme rarity and severity impact, but no real linkage with the chaos theory has been done.

The second hypothesis of which we are going to discuss is more closely related to the stock market domain and is called the 'efficient-market hypothesis'. The 'efficient-market hypothesis'

is a hypothesis in financial economics that states that asset prices reflect all available information at a given time. This would basically imply that all publicly known information about a company, which obviously includes its price history, would already be reflected in the current price of the stock. Accordingly, changes in the stock price reflect release of new information, changes in the market or random movements around the value that reflects the existing information set. Burton Malkiel, in his influential 1973 work 'A Random Walk Down Wall Street', claimed that stock prices could therefore not be accurately predicted by looking at price history. As a result, Malkiel argued, stock prices are best described by a statistical process called a "random walk" meaning each day's deviations from the central value are random and unpredictable.

However, investors and researchers have disputed the hypothesis both empirically and theoretically. For example, Warren Buffet who is considered by many one of the most successful investors in the world rebutted this hypothesis in its speech in 1984.[3] Moreover, there are accepted events which are considered 'stock market anomalies' by the hypothesis since they are violations in which consistently abnormal returns could have been earned by some investment strategies that are constructed based on potential market inefficiencies.Not lastly, there are imperfections in the financial markets which are attributed by behavioural economists to a combination of cognitive biases such as overreaction, overconfidence, information bias and many others.

## 2.2 Scientific Problem

### 2.2.1 Problem definition

Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. In terms of a general scientific category, the problem will be evaluated from the perspective of a time series. Time series analysis and time series forecasting are the two linked activities which are done over a time series. However, while time series analysis is mostly linked to statistics, time series forecasting has evolved as being mostly referred with machine learning in mind. In comparison with other problems, time series require more attention regarding the processing of data since it has a natural time ordering which must not be broken in any way.

The problem of stock market prediction may also be split depending on what information we want to obtain: regression if we want a continuous output variable such as the closing price of some stock or classification if we only want a hint regarding the future trend. This thesis will try to use both categories for different time horizons such as 1-day, 3-day, 5-day etc. For regression we will have two main versions:

- single point in future of the closing price for a company stock
- set of points corresponding to each day in the time horizon - allowing us to see the general flow chart of the company stock price

For the classification category, we will split it in three or five categories (not decided yet): major decline, minor decline, stagnation, minor increase, major increase. Based on all this variants, we are going to compare and decide how should we interpret all results into an automated algorithm which decides whether to buy or sell.

### 2.2.2 Data

Because we are not going to predict intraday evolution of prices (which is usually done in fast markets as Forex), we are going to use the classic data which is provided for evaluating a company stock value. It is available daily and contains five measurements abbreviated usually as OCHLV:

1. Open - the value of a single company stock at the beginning of the daily trading session
2. Close - the value of a single company stock at the end of the daily trading session
3. High - the maximum value of a single company stock obtained during the entire daily trading session
4. Low - the minimum value of a single company stock obtained during the entire daily trading session
5. Volume - the number of shares that were traded during the entire daily trading session

Based on this data, we could theoretically have a complete technical analysis taking into account only the company's past data. However, we will also add at least one stock market index representative for the market from which the company comes from. Stock market indexes can be seen as powerful indicators for global and country-specific economies. Because we are going to use stock data from popular American companies, we'll choose from S&P500, DowJonesIndustrial and NasdaqComposite. Using stock market indexes should help the network getting a grip of the general economic trend in addition to only the company's past data which may or may not reflect a lot the past economic trend of the general economy. In terms of actual data that is obtained from these stock market indexes, we are talking about the same five measurements mentioned earlier, but at the level of that stock market index (which is usually seen as a conglomerate of the most important companies of that market). It should be noted that is totally possible to trade only stock market indexes as opposed to trading individual company stocks

### 2.2.3 Proposed approach

Financial time forecasting has been the top problem of computational intelligence for both academic financial researchers and industry investors due to its broad implementation areas and impact. As mentioned previously, many techniques and methodologies have evolved throughout the recent decades alongside the exponential growth in computational power. Machine learning has become one of the headlines in computer science and more recently a sub-domain of machine learning, deep learning, has started to receive the most attention of the industry. Even though deep learning has been recognized as the next breakthrough in many problems such as speech recognition, image recognition, natural language processing and many more, in terms of time series forecasting and more specifically stock market prediction the research is still continuously developing. The current state of art is a lot harder to define since there are many different approaches to what is going to be forecast and how those results should be incorporated in a strategy that would maximize the profit and minimize the loss. Consequently, this thesis will bring a clearer comparison between similar approaches in terms of output results and strategies, but keeping a consistent approach in terms of the actual deep learning model used.

Deep learning at its core is more of an abstract concept representing a class of machine learning that uses multiple layers to progressively extract higher level features from the raw

input. In terms of architectures, there are many variants which have seen success in one or more domains: deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks. For the specific domain of stock market prediction, there has not been decided a definitive architecture which should be seen as the starting stone. However, based on the fact that the problem of stock market prediction can be seen as a time series forecasting problem where the past data should influence the future trend, there is a growing trend in using variations of recurrent neural networks since they were created exhibiting temporal behaviour in mind and some sort of 'memory'.

Based on the systematic literature review [6], we can draw many conclusions from which will orientate our search. We are going to discuss about several topics: subtopics of financial time series prediction considered and categorization of techniques and network structures that were used.

Firstly, let us begin with the subtopics of financial time series which were considered in the literature review. The paper research included not only stock market dependent activities, but also other markets which have a certain connection with it such as commodity (oil, gold etc.), cryptocurrency or forex markets. Even though, there might be certain differences between them, at their core in all applications the same underlying dynamics occur. As mentioned also in this thesis before, all categories can be clustered in two main groups based on their expected outputs price prediction and trend (movement) prediction. Although, price prediction is definitely the harder problem of the two increasing the difficulty exponentially from classification to regression, in terms of financial and economics interests the actual improvements are not seen as a major importance taking into consideration also the risk management. For that main reason, many researchers consider trend prediction a more crucial study area than the actual price prediction considering the actual implications. However, from the statistics of the literature review [6] it can be observed that trend forecasting represent less than 40% of the actual research considered. Regarding the classification categories which were used two methodologies accounted for almost any study:

- two-class approach as in buy or sell which basically represent an upward or downward trend
- three-class approach which represent one of the following three patterns: decline, increase or stagnation

Other types of classifications are relatively hard to promote or justify for some simple reasons. First, it will be hard to come up with new and justified categories which would not blur the lines between them. Secondly, new categories will imply in theory some percentage margins based on which you would decide for example whether the increase is small or large. Those percentage might also be adjusted depending on how long the time horizon is for the actual predictions. Lastly, more categories would increase the actual difficulty of the problem and can be justified only if you find along a method which would be able to reward the improvements which were attempted. In terms of regression, the main distinctions between solutions without mentioning the time horizon is whether the problem is approached as a single-point or multi-point regression. The difference stems from the actual output of the problem which can be only a single point in the future based on the decided horizon or a set of points for each day in the future until the decided horizon. While the second category is on a whole another level in terms of difficulty, it is also more justified in terms of trying to improve and solve the original classification problem. The first category would eventually be used alongside a investment strategy and a risk management literally in the same way as the classification one, while

the actual set of points may offer more hindsight regarding the market future information and changing an actual investment strategy as a whole. In this thesis, we will try compare the results from multiple approaches and attempt to link some of them into an investment strategy algorithm.

Secondly, we are going to summarize the trending techniques of architectures used in financial time series problems. According to [6] recurrent neural networks dominated in terms of used architectures. The main reason for this occurrence is the actual nature of our problem which implies data across a timeline with time-dependent components. Even though, recurrent neural networks accounted for more than half of the models, we should mention that many variations of recurrent neural networks have been included in this category and classic recurrent neural networks weren't nearly as present as more popular choices (as in other domains) such as long short-term memory (LSTM) or more recently gated recurrent unit (GRU). LSTM networks have been successfully used in domains such as speech recognition, music composition or even time series prediction giving them a stronger sense of security. Besides recurrent neural networks, other techniques were also used such as deep multi layer perceptron (DMLP) due to its acceptance in the past of its 'older brother', the multi layer perceptron (MLP), convolutional neural networks (CNN) or deep reinforcement learning. Taking everything into account, this thesis will use variations composed of LSTM and GRU networks for all the previously mentioned variants of problems due to their recognized worth in another domain and their large flexibility in terms of both classification and regression problems

### 2.2.4 LSTM and GRU

In the following subsection, we are going to present the details regarding the structure and characteristics of long short-term memory and gated recurrent unit layers and networks. The discussion will progressively evolve from classic recurrent neural networks to LSTM which were first proposed over two decades ago and finally to GRU which is more of a newer concept being more 'light-weight' than LSTM, but lacking some abilities.

Recurrent neural networks (RNN) have been created with series in mind such that neighbouring inputs might have an influence on each other in some way.A RNN remembers a portion of its past (previous inputs) and influences the results of future inputs. In order to achieve this behaviour, RNN in addition to 'vanilla' neural networks contain a hidden state which is updated and used for each input received in the series. This hidden state can be thought of as a context based on prior inputs. Moreover, this structure allows to process variable length sequences of inputs, making them applicable for unsegmented tasks such as handwriting recognition. In addition, when speaking of recurrent neural networks we can introduce the topic of 'parameter sharing' which has been successfully in image classifying convolutional neural networks. However, a major difference should be noted in the actual definition of the neighbour between RNN (elements in a series) and CNN (neighbouring pixels in an image).

Even though, RNN were a significant improvement in the theory of machine learning, their classic versions didn't see much success initially. Practical difficulties have been reported in training recurrent neural networks to perform tasks in which the temporal contingencies present in the input/output sequences span long intervals [1]. Based on the previous citation, it was demonstrated that it exists a trade-off between efficient learning by grading descent and maintaining relevant context information for long periods of time. Establishing on these aspects, further research was done to search for stricter structures and architectures which would improve the previous state of the art. Long short-term memory has been a major breakthrough in finding structures that would outperform previous networks, being proposed in 1997. As a testimony,

even today many applications in different domains representing state of the art are using LSTM architectures under the hood.

WORK IN PROGRESS ...

# Chapter 3

# Application Development

# Chapter 4

# Final Conclusions

# Bibliography


[1] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult". In: *IEEE transactions on neural networks* 5.2 (1994), pp. 157–166.

[2] Chris Brooks. "Chaos in foreign exchange markets: a sceptical view". In: *Computational economics* 11.3 (1998), pp. 265–281.

[3] Warren Buffet. *Here's What Warren Buffet Thinks About The Efficient Market Hypothesis*. URL: https://www.businessinsider.com/warren-buffett-on-efficient-market-hypothesis-2010-12.

[4] Klaus Greff et al. "LSTM: A search space odyssey". In: *IEEE transactions on neural networks and learning systems* 28.10 (2016), pp. 2222–2232.

[5] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. "An empirical exploration of recurrent network architectures". In: *International conference on machine learning*. 2015, pp. 2342–2350.

[6] Omer Berat Sezer, M Ugur Gudelek, and Ahmet Murat Ozbayoglu. "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019". In: *Applied Soft Computing* (2020), p. 106181.

[7] Dev Shah, Haruna Isah, and Farhana Zulkernine. "Stock Market Analysis: A Review and Taxonomy of Prediction Techniques". In: *International Journal of Financial Studies* 7.2 (2019), p. 26.