## [LOGO]

# Stock market prediction

Submitted by: Edīte Rimkeviča

Student ID number: 190220655

Candidate Number: 522057

Supervised by:

Leonidas Anastasakis Viktor Pekar

This dissertation is submitted in partial fulfilment of the requirements of Aston University for the degree of MSc Business Analytics.

Submission date: July 2021

#### **Declaration**

I declare that I have personally prepared this report and that it has not in whole or in part been submitted for any other degree or qualification. Nor has it appeared in whole or in part in any textbook, journal or any other document previously published or produced for any purpose. The work described here is my/our own, carried out personally unless otherwise stated. All sources of information, including quotations, are acknowledged by means of reference, both in the final reference section and at the point where they occur in the text.

#### **Acknowledgements**

I would like to express my gratitude and appreciation for Barbara S. whose guidance, support and encouragement has been invaluable throughout this study.

# Leonidas Anastasakis Viktor Pekar

I also wish to thank the team an CIL who have been a great source of support.

## **TABLE OF CONTENTS**

ABSTRACT	4
INTRODUCTION	4
ABBREVIATIONS	4
LITERATURE REVIEW	5
THE STOCK MARKET	5
STOCKS	6
STOCK TYPES	6
WHAT ARE THE BENEFITS AND RISKS OF STOCKS?	7
THE EFFICIENT MARKET HYPOTHESIS	8
STOCK PREDICTION TECHNIQUES	10
STATISTICAL	10
PATTERN RECOGNITION	11
MACHINE LEARNING	11
SENTIMENT ANALYSIS	11
Hybrid	11
THEORY/METHOD/METHODOLOGY	12
MAIN BODY OF WORK	15
CRISP-DM	15
1. Business Understanding	15
2. Data Understanding:	15
3. DATA PREPARATION:	15
4. MODELING:	16
5. EVALUATION:	16
6. DEPLOYMENT	17
CONCLUSIONS	17
RECOMMENDATIONS	17
REFERENCES	18
A RANDOM WALK DOWN WALL STREET: THE TIME-TESTED STRAT	TEGY FOR SUCCESSFUL INVESTING
	19
APPENDICES	20

## **Abstract**

A maximum of one page summarising the aims of the project, the analysis and the main conclusions and recommendations (if appropriate).

## Introduction

A clear, concise statement describing the subject of the report, its aims and objectives together with a description of the context/background/setting of the project and a plan of the project. The introduction may contain more than one section.

## **Abbreviations**

API - an application programming interface

CPI - consumer price index

EMH - the efficient market hypothesis

GDP – gross domestic product

### Literature review

A description and discussion of literature sources that have made a significant contribution to your study. It should demonstrate a critical, considered selection of ideas, concepts, models and data analysis methods used to support and inform your work.

#### The Stock Market

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. Such financial activities are conducted through institutionalized formal exchanges or over-the-counter (OTC) marketplaces which operate under a defined set of regulations. There can be multiple stock trading venues in a country or a region which allow transactions in stocks and other forms of securities. (Chen, 2021)

The first **stock markets** appeared in Europe in the 16th and 17th centuries, mainly in port cities or trading hubs such as Antwerp, Amsterdam, and London. These early stock exchanges, however, were more akin to bond exchanges as the small number of companies did not issue equity. In fact, most early corporations were considered semi-public organizations since they had to be chartered by their government in order to conduct business.

#### (Smith, 2004)

In the late 18th century, stock markets began appearing in America, notably the New York Stock Exchange (NYSE), which allowed for equity shares to trade. The honour of the first stock exchange in America goes to the Philadelphia Stock Exchange (PHLX), which still exists today. The NYSE was founded in 1792 with the signing of the Buttonwood Agreement by 24 New York City stockbrokers and merchants. Prior to this official incorporation, traders and brokers would meet unofficially under a buttonwood tree on Wall Street to buy and sell shares (Hayes, 2021).

The advent of modern stock markets ushered in an age of regulation and professionalization that now ensures buyers and sellers of shares can trust that their transactions will go through at fair prices and within a reasonable period of time. Today, there are many stock exchanges in the U.S. and throughout the world, many of which are linked together electronically. This in turn means markets are more efficient and more liquid. (Hayes, 2021).

Though it is called a stock market or equity market and is primarily known for trading stocks/equities, other financial securities - like exchange traded funds (ETF), corporate bonds

and derivatives based on stocks, commodities, currencies, and bonds - are also traded in the stock markets. (Chen, 2021)

**Stock exchanges** are secondary markets, where existing owners of shares can transact with potential buyers. It is important to understand that the corporations listed on stock markets do not buy and sell their own shares on a regular basis (companies may engage in stock buybacks or issue new shares, but these are not day-to-day operations and often occur outside of the framework of an exchange). So, when you buy a share of stock on the stock market, you are not buying it from the company, you are buying it from some other existing shareholder. Likewise, when you sell your shares, you do not sell them back to the company—rather you sell them to some other investor.

The NYSE and Nasdaq are the two largest exchanges in the world, based on the total market capitalization of all the companies listed on the exchange. (Hayes, 2021)

For example, larger exchanges may require that a company has been in operation for a certain amount of time before being listed, and that it meets certain conditions regarding company value and profitability. In most developed countries, stock exchanges are self-regulatory organizations (SROs), non-governmental organizations that have the power to create and enforce industry regulations and standards.15 The priority for stock exchanges is to protect investors through the establishment of rules that promote ethics and equality. Examples of such SRO's in the U.S. include individual stock exchanges, as well as the National Association of Securities Dealers (NASD) and the Financial Industry Regulatory Authority (FINRA). (Hayes, 2021)

#### **Stocks**

A stock or share (also known as a company's "equity") is a financial instrument that represents ownership in a company or corporation and represents a proportionate claim on its assets (what it owns) and earnings (what it generates in profits).

Stock ownership implies that the shareholder owns a slice of the company equal to the number of shares held as a proportion of the company's total outstanding shares. (Hayes, 2021)

#### Stock types

There are two main kinds of stocks, **common** stock and **preferred** stock.

**Common** stock entitles owners to vote at shareholder meetings and receive dividends.

Preferred stockholders usually don't have voting rights but they receive dividend payments before common stockholders do, and have priority over common stockholders if the company goes bankrupt and its assets are liquidated.

Common and preferred stocks may fall into one or more of the following categories:

- Growth stocks have earnings growing at a faster rate than the market average. They
  rarely pay dividends and investors buy them in the hope of capital appreciation. A startup technology company is likely to be a growth stock.
- **Income stoc**ks pay dividends consistently. Investors buy them for the income they generate. An established utility company is likely to be an income stock.
- Value stocks have a low price-to-earnings (PE) ratio, meaning they are cheaper to
  buy than stocks with a higher PE. Value stocks may be growth or income stocks, and
  their low PE ratio may reflect the fact that they have fallen out of favor with investors
  for some reason. People buy value stocks in the hope that the market has overreacted
  and that the stock's price will rebound.
- Blue-chip stocks are shares in large, well-known companies with a solid history of growth. They generally pay dividends.

Another way to categorize stocks is by the size of the company, as shown in its market capitalization. There are large-cap, mid-cap, and small-cap stocks. Shares in very small companies are sometimes called "microcap" stocks. The very lowest priced stocks are known as "penny stocks." These companies may have little or no earnings. Penny stocks do not pay dividends and are highly speculative.

#### What are the benefits and risks of stocks?

Stocks offer investors the greatest potential for growth (capital appreciation) over the long haul. Investors willing to stick with stocks over long periods of time, say 15 years, generally have been rewarded with strong, positive returns.

But stock prices move down as well as up. There's no guarantee that the company whose stock you hold will grow and do well, so you can lose money you invest in stocks.

If a company goes bankrupt and its assets are liquidated, common stockholders are the last in line to share in the proceeds. The company's bondholders will be paid first, then holders of preferred stock. If you are a common stockholder, you get whatever is left, which may be nothing.

Even when companies aren't in danger of failing, their stock price may fluctuate up or down. Large company stocks as a group, for example, have lost money on average about one out of every three years. If you have to sell shares on a day when the stock price is below the price you paid for the shares, you will lose money on the sale.

Market fluctuations can be unnerving to some investors. A stock's price can be affected by factors inside the company, such as a faulty product, or by events the company has no control over, such as political or market events.

The risks of stock holdings can be offset in part by investing in a number of different stocks. Investing in other kinds of assets that are not stocks, such as bonds, is another way to offset some of the risks of owning stocks. (U.S. Securities and Exchange Commission, n.d.)

## **The Efficient Market Hypothesis**

The efficient market hypothesis (EMH) is introduced by Fama in 1964 (Fama, 1965). Fama demonstrated that stock price movement predictions in the short-time would yield low accuracy, considering that the market is efficient, new information would correct price instantly. The stock is priced accurately until a new piece of information is introduced. The Efficient Market Hypothesis exists in three forms: weak EMH, semi-strong EMH, and strong EMH.

The week EMH – and that stock prices move randomly while price changes are independent of each other hence, it is not possible to beat the market by earning abnormal returns on the basis of technical analysis. Week EMH utilises historical data to predict the stock price. In line with random walk hypothesis.

The semi-strong EMH - the semi-strong form which states that prices adjusted rapidly according to market and public information such as dividend, earnings announcements, and political or economic events, hence it is not possible to earn abnormal returns on the basis of fundamental analysis. Besides historical data, semi-strong EMH utilises all the current public information to predict the stock price.

The strong EMH - the strong form which states that prices reflect market, public, and private information as such no investor has monopolistic access to information (Naseer and Tariq 2015). In the strong EMH all the data is used to predict the stock price, including public and private information.

The Efficient Market Hypothesis implies that investors at best would gain consistent average profits while basing investing decisions on technical or fundamental analysis.

#### Random walk

On the other hand, the random walk hypothesis states that stock prices do not depend on past stocks (Fama, 1995)

Thus, these are not patterns to be exploited since the historical data do not reflect the pattern of the current stock price. the random walk hypothesis states that stock prices do not depend on past stocks. Thus, these are not patterns to be exploited since the historical data do not reflect the pattern of the current stock price.

The impact of Eugene Fama's, results has extended beyond the field of research. For example, his results influenced the development of index funds. Nobel

For many of us, the rise and fall of stock prices symbolizes economic development. In the 1960s, Eugene Fama demonstrated that stock price movements are impossible to predict in the short-term. In the early 1980s, however, Robert Shiller discovered that stock prices can be predicted over a longer period, such as over the course of several years. In contrast to the dominant perception, stock prices fluctuated much more than corporate dividends. Robert Shiller's conclusion was therefore that the market is inefficient.

#### Stock market prediction techniques

Two conventional approaches used for stock market prediction are technical analysis and fundamental analysis.

The fundamental analysis relies on information gathering, analysing and interpreting as main steps in assessing stock price. The trading opportunity arises between an event and the markets response to the event.

Fundamental analysis consists of three aspects:

- Macroeconomic analysis which approximate the future profit of company affected by the macroeconomic environment. The most common indexes are Gross Domestic Product (GDP) and Consumer Price Index (CPI).
- Industry analysis which evaluate the value of the company based on industry status and prospect
- Company analysis which estimates internal value of the company and examine the financial status and the current operations (Hu, et al., 2015)

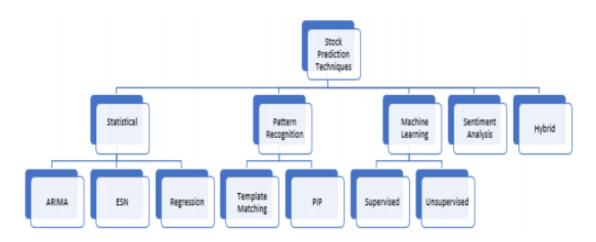
Technical analysis is a numerical time series approach to predict stock markets based on historical data using charts as the primary tool [6]. This approach tries to mine information from the historical data to recognize the pattern, sometimes referred as mining the time series [2].

(Vui, et al., 2013) argues that conventional approaches as fundamental and technical analysis are becoming subpar compared to computers, where with recent rise in the

computational power, computers can analyse larger data sets accurately and faster. Machine learning and computational intelligence still utilise these approaches as foundation for new artificial intelligence approaches.

The stock market has nonlinear, nonstationary ,dynamic, nonparametric, noisy, and chaotic nature, which makes predicting stock market movements and price behaviours is extremely challenging (Abu-Mostafa & Atiya, 1996)

In addition, the stock market is affected by many highly interconnected factors that include economic, political, psychological, and company-specific variables (Zhong & Enke, 2017).



(Shah, et al., 2019)

**Stock Prediction Techniques** 

**Statistical** 

Arima

Autoregressive integrated moving average (ARIMA). The ARIMA model is a widely used technique for stock market analysis (Hiransha, et al., 2018).

ARMA combines the autoregression (AR) models which try to explain the momentum and mean reversion effects often observed in trading markets and Moving Average (MA) models which try to capture the shock effects observed in time series. Compared with ARMA, ARIMA is capable of dealing with non-stationary time series, by introducing an initial differencing step, which is referred as the integrated part in the model (Jiang, 2020).

ARIMA is a univariate time series prediction and therefore it is not able to identify the non-linearities present in the data, meanwhile, neural network applied to the same data are able to identify the non-linear trends present in the data (Hiransha, et al., 2018).

Linear regression (LR). Linear regression is a classical linear model that

tries to \_t the relationship between the predicted target and the input variables with a linear model, in which the parameters can be learned in the least squares approach.

Generalized autoregressive conditional heteroskedasticity (GARCH). GARCH is also a generation of the autoregressive conditional heteroscedasticity (ARCH) model, which describes the error variance as a function of the actual sizes of the previous time periods' error terms. Instead of using AR model in ARCH, GARCH assumes an ARMA model for the error variance, which generalizes ARCH.

**ESN** 

Regression

#### **Pattern Recognition**

Template matching

PIP

#### **Machine Learning**

Supervised

Unsupervised

#### Sentiment analysis

Hybrid

Sentiment represents the behaviours of various market participants. Flow-of-funds is a type of indicator used to investigate the financial status of various investors to pre-evaluate their strength in terms of buying and selling stocks, then, corresponding strategies, such as short squeeze, can be adopted. Raw data include stock price series and price patterns such as K-line diagrams and bar charts. Trend and momentum are examples of price-based indicators, trend is used for tracing the stock price trends while momentum is used to evaluate the velocity of the price change and judge whether a trend reversal in stock price is about to occur. Volume is an indicator that reflects the enthusiasm of both buyers and sellers for investing, it is also a basis for predicting stock price movements. The cycle is based on the theory that stock prices vary periodically in the form of a long cycle of more than 10 years containing short cycles of a few days or weeks. Finally, volatility is often used to investigate the fluctuation range of stock prices and to evaluate risk and identify the level of support and resistance. (Hu, et al., 2015)

## Theory/method/methodology

Approach, frameworks/models and reasons for choosing these methods of collecting and analysing data. This may be a separate chapter or may be included at appropriate points in other parts of the report.

#### **Benchmark**

If positive, invest 10 000, hold one day. sells instantly if can make 2% profit (Abu-Mostafa & Atiya, 1996).

#### **Data sources**

In this research, we hypothesize that combining disparate online data sources with traditional time-series and technical indicators for a stock can provide a more effective and intelligent daily trading expert system.

There are several commercial financial expert systems that can be used for trading on the stock exchange. However, their predictions are somewhat limited since they primarily rely on time-series analysis

In Phase I, we scrape four sets of data from online resources. These datasets include: (a) publicly available market information on stocks, including opening/closing prices, trade volume, NASDAQ and the DJIA indices, etc.; (b) commonly used technical indicators that reflect price variation over time; (c) daily counts of Google News on the stocks of interest; and (d) the number of unique visi- tors for pertinent Wikipedia pages per day. (Weng, et al., 2017)

#### 1. Formulas for the generated features

In this study, we generated seven different types of features or Wikipedia traffic data and Google news data. The formulas are nown below. In the formula, n means the time periods,  $V_t$  means ne data point at period t.

1. Moving average:

$$MA(n)_t = \frac{V_t}{n} + \frac{V_{t-1}}{n} + \cdots + \frac{V_{t-n+1}}{n}$$
 (1)

2. Exponential moving average:

$$EMA(n)_t = (V_t - MA(n)_{t-1}) \times (\frac{2}{n+1}) + MA(n)_{t-1}$$
 (2)

3. Disparity:

$$Disparity(n)_t = \frac{V_t}{MA(n)_t} \times 100$$
 (3)

4. Momentum1:

$$Momentum 1_t = \frac{V_t}{V_{t-5}} \times 100$$
 (4)

5. Momentum2:

$$Momentum2_t = (V_t - V_{t-5}) \times 100$$
 (5)

6. Rate Of Change:

$$ROC_t = \frac{V_t}{Momentum 2_t} \times 100$$
 (6)

7. Relative Strength Index:

$$RSI(n) = 100 - \frac{100}{1 + \frac{AverageGain(n)}{AverageLoss(n)}}$$
(7)

Weng, B Stock market one-day ahead movement prediction using disparate data sources

#### Structured events

- (1) Structured events are more useful representations compared to bags-of-words for the task of stock market prediction.
- (2) A deep neural network model can be more accurate on predicting the stock market compared to the linear model.
- (3) Our approach can achieve stable experiment results on S&P 500 index prediction and individual stock prediction over a large amount of data (eight years of stock prices and more than 550,000 pieces of news).
- (4) The quality of information is more important than the quantity of information on the task of stock market prediction. That is to say that the most relevant information (i.e. news title vs

news content, individual company news vs all news) is better than more, but less relevant information. (Ding, et al., 2014)

In case of the absence of the news for the individual company on the previous day, model cannot predict as it does not utilise long-term and mid-term news (Luss & d'Aspremont, 2015) (Ding, et al., 2015)

(Ding, et al., 2014) argues that it can reduce profitability.

(Ding, et al., 2015) derives that there is no necessary to use full news article. In fact, the titles are prove to be more useful for prediction compared to the news content.

(Ding, et al., 2015) Building on the previous work, it was proven that events are better features than words for stock prediction. The proposed model delivers relatively better results for the companies with lower fortune ranking, with lower volume of news available.

## **Main body of work**

As many sections or chapters as necessary to cover the specifics of the investigation to include the data collection, modelling, data analysis and evaluation.

## **CRISP-DM**

#### 1. Business Understanding

: The situation is assessed from a business perspective and objectives for the data project are determined

#### 2. Data Understanding:

The initial data is collected, described and explored in detail and the quality of data is verified.

#### 3. Data Preparation:

The data is selected, cleaned, constructed, integrated and formatted appropriately to fit with the objectives of the project

**Reddit r/worldnews** top30 for the last 24hours daily. Top rankings are determinate by posts upvotes. Title of the post is used. Author obtained data straight from Reddit, with Python using Reddit API PRAW 09/03/2021.

To increase period of the analysis, previous Reddit headlines was web scraped from the https://www.redditsearch.io from 01/01/2021 to 08/03/2021. However, there are problems with data between 01/02/2021 and 22/02/2021. There are present posts with low counts of upvotes and at the present removed by moderators from r/worldnews. In addition, when plotting subjectivity for periods, there is no variability in maximum and mean subjectivity of the posts. Author expects predictions to improve as training sets predominately consist of the manually web scraped data.

## Development of subjectivity score

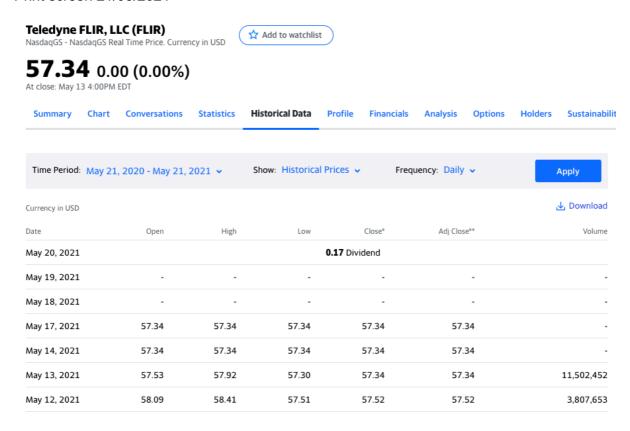


**Reddit r/wallstreets** bets for the last 24 hours, count of the stocks mentioned in discussion boards. Sentiment analysis applied to gauge market outlook towards the stocks.

#### Wikipedia SP500 stocks and changes

In case of FLIR being acquired by S&P 500 constituent Teledyne Technologies Inc on 14<sup>th</sup> of May 2021, Wikipedia listing was updated after daily data collection was done for 19<sup>th</sup> of May 2021. Yahoo finance reported no trading volume on the acquisition date.

#### Print screen 21/05/2021



**Yahoo finance:** Open, close, high, low, close, adjusted close volume. For SP500 component stocks and SP500 index itself. Dividends are not taken into profit calculation.

#### 4. Modeling:

A modeling technique is selected and the model is built and assessed

#### 5. Evaluation:

the results of the analysis are reported and evaluated

To verify the statistical significance of our earnings, we perform a randomization test [Edgington and Onghena, 2007] by randomly buying or shorting for 1000 trials. The mean profit over the randomized test is -\$9,865 and the performance of our model is significant at the 1% level. (Ding, et al., 2015)

#### 6. Deployment

A plan for deployment, monitoring and maintenance is set out and the final report is produced

## **Conclusions**

Summary of overall findings, discussion of possible courses of action, constraints and limitations.

## **Recommendations**

Actions recommended on the basis of the findings and conclusions. Conclusions and recommendations may be in the same chapter (not all projects require recommendations).

## References

- Abu-Mostafa, Y. & Atiya, A., 1996. Introduction to financial forecasting. *Applied intelligence,* Issue 6(3), pp. 205-213.
- Chen, J., 2021. Stock Market. [Online]

  Available at: <a href="https://www.investopedia.com/terms/s/stockmarket.asp#what-is-the-stock-market">https://www.investopedia.com/terms/s/stockmarket.asp#what-is-the-stock-market</a>

  [Accessed 01 June 2021].
- De Bondt, W. & Thaler, R., 1985. Does the stock market overreact?. *The Journal of finance,* Volume 40(3), pp. 793-805.
- Ding, X., Zhang, Y., Liu, T. & Duan, J., 2014. Using structured events to predict stock price movement: An empirical investigation.. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Volume October, pp. 1415-1425.
- Ding, X., Zhang, Y., Liu, T. & Duan, J., 2015. Deep Learning for Event-Driven Stock Prediction.

  Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Volume June.
- Fama, E., 1965. The behavior of stock-market prices. *The journal of Business*, 38(1), pp. 34-105.
- Fama, E., 1995. Random walks in stock market prices. *Financial analysts journal,* Issue 51(1), pp. 75-80.
- Hayes, A., 2021. *How Does the Stock Market Work?*. [Online]
  Available at: <a href="https://www.investopedia.com/articles/investing/082614/how-stock-market-works.asp">https://www.investopedia.com/articles/investing/082614/how-stock-market-works.asp</a>
  [Accessed 1 June 2021].
- Hiransha, M., Gopalakrishnan, E. A., Menon, V. K. & Soman, K. P., 2018. NSE stock market prediction using deep-learning models. *Procedia computer science,* Issue 132, pp. 1351-1362.
- Hu, Y. et al., 2015. Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review. *Applied Soft Computing*, Volume 36, pp. 534-551.
- Jiang, W., 2020. Applications of deep learning in stock market prediction: recent progress. arXiv preprint arXiv:2003.01859.
- Lavrenko, V. et al., 2000. Mining of concurrent text and time series. *In KDD-2000 workshop on text mining. Citeseer,* Volume 2000, pp. 37-44.
- Luss, R. & d'Aspremont, A., 2015. Predicting abnormal returns from news using text classification. *Quantitative Finance*, Volume 15(6), pp. 999-1012.

- Malkiel, B., 2003. The efficient market hypothesis and its critics. *Journal of economic perspectives*, Volume 17(1), pp. 59-82.
- Shah, D., Isah, H. & Zulkernine, F., 2019. Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, Volume 7(2), p. 26.
- Smith, B., 2004. A history of the global stock market: from ancient Rome to Silicon Valley. In: A Financial Revolution. s.l.:University of Chicago press, pp. 13-17.
- U.S. Securities and Exchange Commission, n.d. What are stocks?. [Online]

  Available at: <a href="https://www.investor.gov/introduction-investing/investing-basics/investment-products/stocks#Kinds">https://www.investor.gov/introduction-investing/investing-basics/investment-products/stocks#Kinds</a>
  [Accessed 1 Jun 2021].
- Vui, C. S. et al., 2013. A review of stock market prediction with Artificial neural network (ANN). Penang, Malaysia, IEEE, pp. 477-482.
- Weng, B., Ahmed, M. A. & Megahead, F. M., 2017. Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications*, Volume 79, pp. 153-163.
- Wirth, R. & Hipp, J., 2000. *CRISP-DM: Towards a standard process model for data mining.* London, UK, Springer-Verlag.
- Zhong, X. & Enke, D., 2017. Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, Volume 67, pp. 126-139.

# A random walk down Wall Street : the time-tested strategy for successful investing

Burton G. Malkiel

Year of publication: 2016; Completely revised and updated, Norton paperback

# **Appendices**

Supporting material, often detailed but not essential for the reader to make sense of the main 'story'. No appendix material should be included unless it is referred to in the main text.