PREDICTING FUTURE BEHAVIOUR OF S&P 500 STOCK MARKET INDEX

A machine learning-based approach leveraging MarketPsych sentiment indicators

Ángeles Blanco Fernández

A Thesis for the Degree of Master in Data Science

October 2020

Author Note

This thesis has been tutored by Felipe Alonso Atienza.

Alberto de la Fuente and Antonio Aita in BBVA Asset Management SGIIC are acknowledged for their advice on the business side of this work.

Contact: angels.blanc.fdz@gmail.com

Abstract

Successful investment strategies need to be ahead of stock market movements. Machine learning paves the way for the development of financial theories that can forecast those movements. In this work an application of the Triple-Barrier Method and Meta-Labeling techniques is explored with XGBoost for the creation of a sentiment-based trading signal on the S&P 500 stock market index. The results confirm that sentiment data have predictive power, but a lot of work is to be carried out prior to implementing a strategy.

Keywords: Stock market prediction; Ensemble learning; Extreme gradient boosting; News sentiment; Trading strategy; Meta-labeling

Table of Contents

Chapter I.

Introduction

Financial markets are one of the most fascinating inventions of our time, playing a key role in the contemporary economy. Predicting stock price behaviour in an accurate fashion has been a long-standing area of interest in the fields of finance and economics, both for researchers and investors. It is, nonetheless, an extremely challenging endeavour due to the intrinsic complexity of financial markets. Numerous interrelated factors influence price movements, such as supply and demand, global economic conditions, political events, investors' sentiment towards traded companies, and so on and so forth. Noted quantitative researcher Prof. Marcos López de Prado provides an ingenious perspective on this matter in his book *Machine Learning for Asset Managers* (2020) [1]:

> Imagine if physicists had to produce theories in a universe where the fundamental laws of nature are in a constant flux; where publications have an impact on the very phenomenon under study; where experimentation is virtually impossible; where data are costly, the signal is dim, and the system under study is incredibly complex… (Chapter 1, p.20).

He then goes on to state his admiration for the advancements that financial academics have accomplished in the face of paramount adversity. Stock market prices are by and large dynamic, non-parametric, non-linear, chaotic in nature and with an utterly low signal-to-noise ratio. This highly volatile nature of the share market makes investments risky and, therefore, advanced knowledge in future price movements is essential to minimize the associated risk involved in a trade. Taking the example of stocks trading, if the value of a stock is expected to increase in the future, it is more likely to be bought. Conversely, investors will typically sell or refrain from buying shares whose value is expected to undergo a foreseeable fall. So, fundamentally, there is a manifest need in forecasting price behaviour to maximize capital gains and minimize losses.

Several theories on the feasibility of beating the markets have been conceived over the years. One such popular and most debated theory proposed by Eugene Fama (1970) [2] is the **Efficient Market Hypothesis** (EMH) which states that at any point in time, the market price of a stock already reflects all information about that stock. According to EMH, price changes are unpredictable and forecasting a financial market is a hopeless effort [3]. However, criticism to this theory has given rise to an increasing number of scientific papers that contradict and reject the validity of the EMH basis -- there are also famous examples of funds that have successfully beaten the market over the years, like Renaissance Technologies's flagship Medallion fund [4] --, introducing new and successful approaches that combine classical technical analysis indicators with methodologies that range from traditional econometrics to more novel applications -- namely data mining, sentiment analysis and machine learning [5]. **Machine Learning** (ML) has a significant ability to identify valid information and detect patterns in the data. Even so, it is important for investors to recognize that ML is not a substitute for economic theory, but rather a most powerful tool for building modern economic theories [1].

Reasonably, it follows that in order for prediction algorithms to be efficacious, they need to incorporate at least some of the myriad influential factors mentioned at the beginning of this section. These information flows may often come in the form of professional news hitting major newswires like Thomson Reuters or Bloomberg, and might also be present in social media content, such as in the interactions that take place between the users of social networking services like Twitter. Modern **Natural Language Processing** (NLP) algorithms can "interpret" news, converting them from unstructured text to numbers that indicate sentiment of a news item regarding a given asset, among other features. This manageable numerical input can then be processed by machine learning models and algorithmic trading strategies [6].

Purpose and structure

In this work, the viability of creating a trading strategy based on sentiment data from Thomson Reuters MarketPsych Indices (TRMI) to beat the Standard & Poor's 500[1] (hereinafter S&P 500) stock market index is explored. The learning algorithm applied is Extreme Gradient Boosting (XGBoost), a tree-based ensemble technique which has been significantly employed by top data scientists in competitions. The particularities of the methodology and the description of the financial model are explained in subsequent sections, highlighting the implementation of novel ideas for quantitative research introduced for the first time in the book *Advances in Financial Machine Learning* (López de Prado, 2018), such as the Triple-Barrier Method (TBM) and Meta-Labeling [7]. The ultimate business goal of this dissertation is to assess the usefulness of the mentioned sentiment indicators and their incorporation into investment and trading decision processes as those made by asset managers when administering investment portfolios and performing the necessary asset allocation[2]; specifically, portfolios that have broad exposure to equity markets[3] by means of funds that track financial stock indices like the S&P 500 -- i.e., index funds --, or by using any other type of financial asset[4]. Investment managers can handle several multi-billion dollar funds, thus yielding their decisions critical for avoiding grave losses and delivering net profitable returns.

The rest of the thesis is structured as follows: Chapter II digs a little deeper into some financial concepts lightly introduced in this section which are pivotal for the reader to be knowledgeably aware of the proceedings. This chapter additionally considers the state-of-the-art proceedings in financial machine learning for this work. Chapter III discusses the data utilized in the construction of the financial model, also accounting for the description of the

[1] spglobal.com/spdji/en/indices/equity/sp-500/#overview

[2] See Chapter II, section *On asset allocation*, for a more detailed explanation on asset allocation and the particular need for implementing a strategy with sentiment data.

[3] The terms stock market, share market and equity market are used interchangeably throughout the document.

[4] Also see Chapter II for the full definition of a financial asset and further basic financial terms.

variables, their preparation and the selection process of relevant predictors. Chapter IV introduces the TBM and the Meta-Labeling notions as part of the section *Labeling*. It also illustrates the intuition behind the XGBoost algorithm and reviews the special model architecture created. Chapter V presents the results obtained for various different scenarios. Finally, Chapter VI summarizes the findings, gives the conclusions achieved and mentions further lines of work.

Chapter II.

A business perspective

This chapter introduces some concepts on the underlying business need of this work in the section *Financial world*, and briefly reviews the financial machine learning state-of-the-art in the succeeding section, *Financial state-of-the-art*.

Financial world

Prior to continuing into the procedures of this study, it is of great interest to delve a little deeper into the financial building blocks that conform the business need from which this work is derived; namely, investments, stock market indices, some relevant measures and, finally, asset allocation.

On investments and analysis

Quoting Feng and Palomar in their book *A Signal Processing Perspective on Financial Engineering* (2016) [8]: an investment is the commitment of resources, which in financial markets usually take the form of money, in the expectation of reaping future benefits. Thus, the investment is the present commitment of money in order to reap (hopefully more) money later. The carriers of money in financial markets are usually referred to as financial assets. There are various classes of financial assets, namely, equity securities (e.g., common stocks), exchange-traded funds (ETFs), commodities, fixed-income securities, derivatives (e.g., options and futures), etc; and a detailed description of each kind of asset is well documented, e.g., [9, 10]. For different kinds of assets, the key quantities of interest are not the same; for example, in equity securities the quantities of interest are the compounded returns or log-returns, which

will later be introduced in this chapter; but in contrast, for fixed-income securities, they are the changes in yield to maturity.

In regard to investment analysis, there are by and large three families of investment philosophies: fundamental analysis, technical analysis, and quantitative analysis. **Fundamental analysis**, in the example of shares, uses financial and economical measures, such as earnings, dividend yields, expectations of future interest rates[5], and management, to determine the value of these assets. Warren Buffett is probably the most famous practitioner of fundamental analysis[6]. **Technical analysis** is essentially the search for patterns in one dimensional charts of the prices of a stock, though it is generally implemented in an anecdotal way with a low predictive power. **Quantitative analysis** applies quantitative (namely scientific or mathematical) tools to discover the predictive patterns from financial data. And to put this in perspective with the previous approach, Feng and Palomar also give an ingenious analogy: "Technical analysis is to quantitative analysis what astrology is to astronomy."

On indices and the S&P 500 index

An index is a method to track the performance of some group of assets in a standardized way. Indexes typically measure the performance of a basket of securities[7] intended to represent a certain area of the market. These may be broad-based to capture the entire market such as the S&P 500, or more specialized such as indexes that track a particular industry or segment. Often, indices in financial markets serve as benchmarks against which to evaluate the performance of a portfolio's returns[8].

[5] An interest rate is the amount charged by a lender and due per period, as a proportion of the total amount lent, deposited or borrowed.
[6] Warren Buffett: How He Does It: investopedia.com/articles/01/071801.asp
[7] A security is a tradable financial asset.
[8] In finance, a portfolio is a collection of investments; and returns are profits on those investments.

The history of the inception of the S&P 500 index dates back to the first half of the 20th century. The S&P 500 is a financial market index that measures the stock performance of 500 of the top publicly traded companies in leading industries of the United States economy, selected by a committee that assesses their compliance with certain criteria, and it is reconstituted quarterly. For instance, a company's market capitalization, i.e., the market value of a company's outstanding shares, must be greater than or equal to USD 8.2 billion to be included[9].

The S&P 500 accounts for 83% of the U.S. total stock market value. Because it reflects nearly all of the largest stocks in the U.S., it is often regarded as synonymous with "the market" as a whole. Figure 1 shows how the S&P 500 reflects critical events for both the U.S. and the global economy.

In terms of its estimation, the S&P 500 is a capitalization-weighted index; that is, companies are weighted in proportion to their market capitalization. The formula to calculate the S&P 500 index value may be written as:

$$S\&P\ 500\ Index\ = \frac{\Sigma\,(P_i\,Q_i)}{Divisor},$$

where P_i is the price of each stock in the index and Q_i is the number of shares publicly available for each stock. The divisor is proprietary information of S&P and is not released to the public, although, by definition, it is a figure that is adjusted to keep the value of the index consistent despite corporate actions that affect market capitalization[10]. Each index has its own calculation methodology, and in most cases the relative change of an index is more important than the actual numeric value representing it.

In this work, the S&P 500 was selected because it is one of the world's best-known indices and one of the most commonly used benchmarks for the stock market, but other indices

[9] S&P 500 The Gauge of the Market Economy: spglobal.com/spdji/en/documents/additional-material/sp-500-brochure.pdf
[10] S&P U.S. Indices Methodology: spglobal.com/spdji/en/documents/methodologies/methodology-sp-us-indices.pdf

may apply for the business purposes introduced in the previous chapter -- mainly asset allocation --, as portfolios may be constructed with positions into several indices.

Figure 1. S&P 500 and some paramount historical milestones.

The S&P 500 index is shown here with some relevant events in the economy. As a single gauge for the stock market, it can be seen how index prices undergo major falls after crucial events that trigger economic recessions.

On returns and volatility

Besides the price of a stock, another quantity of interest to measure profits on investments is the **rate of return** or, simply, return. Let p_t be the price of an asset at a point in time *t* which can denote any arbitrary period such as days, months, 5-minute intervals, etc. The simple return (also known as linear return or net return) over an interval from time *t - 1* to *t* is:

$$R_t = \frac{p_t - p_{t-1}}{p_{t-1}}.$$ [11]

Returns present the benefit of normalization, as they enable the measure of all assets in a comparable metric, thus allowing for the evaluation of analytic relationships despite originating from price series of unequal values. The time series of returns shows irregular fluctuations of positive and negative peaks, corresponding to relative increases and decreases of the price of

[11] Note that in the case of the existence of dividends (part of the profits paid to shareholders), the return can be adjusted by simply adding the computation of dividends to the numerator of this equation.

the represented asset. Figure 4 in Chapter II shows a calculation of returns for the S&P 500 index. It is also worth noting that this signal presents heteroscedasticity, as the standard deviation (volatility) of its values is non-constant in time.

At this point, special mention should be made to two important notions in asset allocation: the expected return and the amount of risk or volatility taken on an investment. The expected return of a single security, for instance, can be approximated by the sample mean, that may be written in a simple approach as: $\hat{\mu} = (1/T)\ \Sigma_{t=1}^{T}\ r_t$, where T is a time period and r_t is the continuously compounded return or log-return[12], $r_t = log(1 + R_t)$. **Volatility** is a statistical measure of the dispersion of returns for a security or a market index. In most cases, the higher the volatility, the riskier the security. It is often measured as the standard deviation $\hat{\sigma}$ or variance between returns from that same security or market index:

$$\hat{\sigma} = \sqrt{\Sigma_{t=1}^{T}\ (\ r_t\ -\hat{\mu}\ )^2\ /\ (T\ -\ 1)}\ .$$

It is important to emphasize that in this work mean return and volatility are computed with exponentially weighted moving windows, as it will be shown in succeeding sections. Formulas for their calculation are thus slightly different, but can be found in [11].

Risk and volatility are popular terms used in financial markets. The concept of volatility is mathematical while risk may be psychological, as it refers to the possibility of adverse effects, such as a loss. Therefore, volatility does not directly imply risk of loss. It merely refers to the price action (how rapidly or severely it may change), and some investments may be more volatile than others.

[12] Log-returns have several statistical benefits over simple returns, e.g., [12]..

On asset allocation

In Chapter I, it was suggested that sentiment indicators could be a useful tool to incorporate on asset allocation decisions. Let us begin by introducing a formal definition of what asset allocation entails: asset allocation is the name given to the diversification of investments between different products or markets, in order to improve performance and control the risk of the aggregate[13]. In other words, asset allocation attempts to balance risk versus reward by optimizing the percentage of each asset in an **investment portfolio**. This is executed according to the desired risk tolerance, goals and investment time horizon; and with a focus on the overall portfolio.

There are several strategies taken into account on the construction of an investment portfolio, but in general terms they could be classified as either long-term or short-term approaches. The former approach could be deemed as being more strategic, integrating fundamental analysis and expert assessment with the primary goal of creating a combination of assets that seek to provide the optimal balance between expected risk and return for a long-term investment. This kind of optimization tends to be based on historical information -- classical mean-variance portfolio selection applies here [13]. A short-term approach is regarded as a **tactical asset allocation**. In this strategy, while an original asset combination is formulated much like a strategic approach, an investor takes a more active style that tries to position the portfolio into those assets, sectors, markets, or individual stocks that show the most potential for perceived gains -- and get out of risky positions on the contrary case of perceived losses. It is here, in this tactical approach, where sentiment-derived information can take a pivotal role and exert major influence.

[13] bbva.es/diccionario-economico/a/asset-allocation.html

As an example guided by this work, imagine a portfolio that has a certain percentage of its exposure on U.S. equities, either by index funds or any other types of investment vehicles. In the advent of a foreseeable fall for the U.S. stock market predicted by a sentiment-derived signal, active asset managers are likely to diminish their exposure by closing their long positions[14] on U.S. equities, thus avoiding serious risk and possible capital losses, and optimizing the overall portfolio.

Financial state-of-the-art

Financial problems pose a particular challenge to legacy methods of classical statistics, because economic systems exhibit a degree of complexity that is beyond the grasp of classical statistical tools [1]. As a consequence, ML plays an increasingly important role in finance. The state-of-the-art in financial machine learning is taken here as the contributions of Prof. López de Prado [1, 7]. Also, a review on the latest applications of prediction techniques for stock market analysis can be found here [3]. Some other papers analyzing news or sentiment indicators and stock market prices may be found in [6, 14, 15].

[14] Taking a long position stands for the action of buying a security. The opposite term, taking a short position, is consequently to sell such a security.

Chapter III.

Exploratory data analysis

This chapter first offers a description of the data used for the generation of this report. As a remainder for the reader, the aim is to construct a predictive model with sentiment information as a form of input, being the future behaviour of the S&P 500 index the target to predict (later on the disambiguation of "future behaviour" will be given). Lastly, this chapter then goes through the different phases of data preparation.

Data description

The dataset is a historical time series of news sentiment-related and financial variables. It comprises daily observations from 2001/08/31 up to 2020/08/31. The dependent variable or target variable is a special discretization of the S&P 500 index following a labeling created by the combination of the previously mentioned TBM and Meta-Labeling methods. The goal is not to simply predict whether the price is going to go up or down, but rather to create a signal of predictions on whether to Buy or Sell, as part of a trading strategy. The *Labeling* section offers a deep explanation of this special labeling methodology. The focus of this section is a preliminary description of the datasets.

Sentiment indicators

As stated earlier in this document, the sentiment indicators used are the **Thomson Reuters MarketPsych Indices** (TRMI). TRMIs analyze news and social media in real-time to convert the volume and variety of professional news and the internet into manageable information flows that drive sharper decisions[15]. Indicators can be generated for various types of assets;

[15] MarketPsych's User Guide: old.marketpsych.com/guide/

but, for this research, indices related to Companies, Currencies and Countries are selected. Besides, three classes of indicators are provided:

- **Emotional indicators** such as Surprise, Fear and Joy (see Figure 2 for two examples of these metrics).
- **Macroeconomic metrics** like Earnings Forecast, Interest Rates Forecast, Monetary Policy Loose Vs Tight.
- **Buzz metrics** on the asset level, i.e., Buzz -- which represents a sum of entity-specific words and phrases used in TRMI computations --, and **on market-moving topics** for that asset, such as Litigation, Regulatory Crackdown, Mergers and Volatility.

TRMIs are asset-level scores on a collection of content. There are two time-related metrics that determine TRMI scores: the window length, which determines what range of content is scored in generating a set of TRMIs; and the update frequency, which determines the time between consecutive TRMI scores. For this work, the selected combination exhibits a window length of 24 hours and a daily update frequency.

Selected TRMIs are evaluated on the combined content of news and social media, and only English-language text is used. News content is restricted to Reuters news and Internet news narrowed down to those from top international and business news sources, top regional news sources, and leading industry sources. In respect of social media content, this generally includes tweets and financial blogs within the top 20%, measured by incoming links on popularity ranks. Nonetheless, it also includes content from hundreds of less-popular asset-specific blogs and forums.

Consequently, sentiment information is firstly structured in three datasets, from where asset-specific information is filtered: for Countries dataset, only the group of companies assembled according to market capitalization oriented by the S&P 500 is selected; for the

Currencies dataset, only indicators related to the United States dollar are chosen; and for the Countries dataset, only indicators for the United States are picked. Note that these specifications apply solely for this work, as the initial data downloaded from Reuters consists of substantially more information -- e.g., there are 186 countries or regions besides the US in the Countries dataset. After this first filtering and the elimination of system variables, the datasets are merged into one, comprising a total of 90 sentiment features (see Appendix A for more details about the sentiment features) and 1 date feature index.

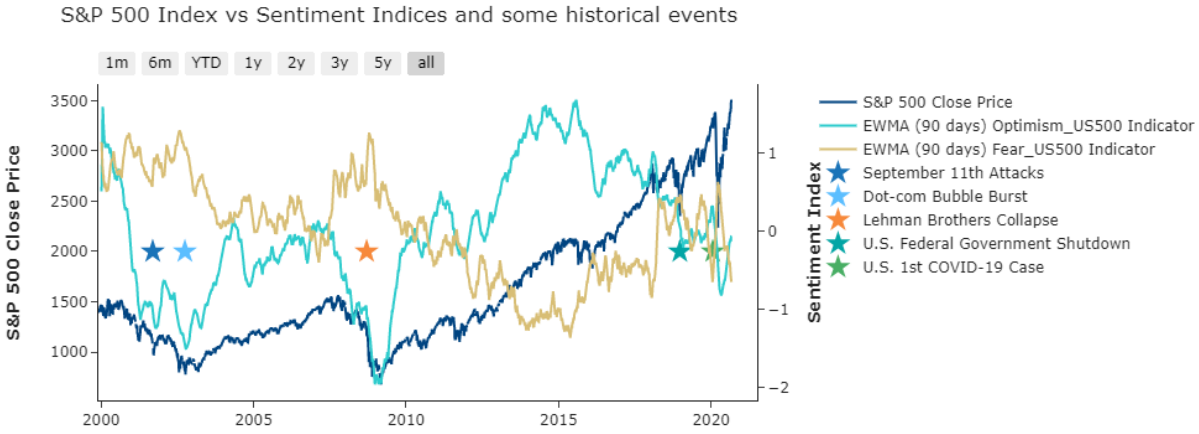


Figure 2. S&P 500 Index vs emotional indicators and milestones.

As an analogy with Figure 1, here the S&P 500 and some major historical events are represented with 2 sentiment features of the emotional class: Optimism on S&P 500 companies, and Fear on S&P 500 companies. One can observe that there exists a sizeable degree of proportionality both between the almost opposing emotional features (with a Pearson-correlation coefficient of -0.93[16]), and between each of them and the S&P 500 index (with a Pearson-correlation coefficient of 0.92 for the Optimism feature, and a value of -0.84 for the Fear variable). Both the plot and the Pearson coefficients have been produced after dispersion and scale transformations, and with an EWMA[17] of 90 days applied to the sentiment indicators.

[16] Measures computed on the training set.

[17] Later on this chapter, in the *Smoothing* section, a definition of EWMA will be given.

In this merged sentiment dataset, weekend values are then eliminated and added to Monday values as part of a weighted average. The reason for this action is to behave in accordance to market data. The stock market is closed on weekends and hence no price data is available during that time. The premise for the weighted average is that the closing index on Monday is the result of accumulated opinions that investors and traders have formed during that day and along the weekend -- e.g., news about a company can of course be released while the market is closed, shifting what investors are willing to pay to own a share of the company.

Lastly, the dataset is divided into train and test sets for the subsequent data preparation, with ratios of 79% (from 2001/08/31 to 2016/08/31) and 21% (from 2016/09/01 to 2020/08/31), respectively.

Financial variables

Financial data for the S&P 500 index is extracted from Yahoo Finance using their Python API. The initial financial dataset consists of 4 selected variables and 1 date feature index, constituting daily bar chart data. The data structures used to contain trading information are often referred to as bars. One daily time bar represents the set of quotes (High, Low, Open, Close)[18] for one trading day (see Figure 3). Adjusted Close -- which is the Close adjusted for stock splits[19] and dividends -- is also provided, but since the S&P 500 is an index, this variable is just the same as Close and thus is rejected. Volume is also provided, but it has been found to be unreliable[20], so it is also discarded.

[18] Over the time period covered by a bar, High and Low mean the maximum and minimum price reached, respectively. Open is the first price of the time period, and Close is the last price.

[19] A stock split is an increase in the number of shares of a corporation's stock without a change in the shareholders' equity. Companies often split shares of their stock to make them more affordable to investors. Unlike issuing new shares, a stock split does not dilute the ownership interests of existing shareholders. For example, if an individual owns 100 shares of a company that trades at $100 per share and the company declares a two-for-one stock split, the individual will own 200 shares at $50 per share immediately after the split. If the company pays a dividend, the dividends paid per share also will fall proportionately [16].

[20] elitetrader.com/et/threads/yahoo-spx-volume-is-just-weird.319962/page-2

Figure 3. S&P 500 candlestick bar chart.

This image shows an example of a typical bar chart extracted from Yahoo Finance and constructed with candlesticks, which represent the set of quotes (High, Low, Open, Close) for one trading day. Vertical bars on the bottom of the graph represent Volume. On the upper right corner, a scheme of candlesticks is given.

Two more features are added at this state: Daily Return and Daily Volatility (see Figure 4). Daily Return is the percent change of the index between adjacent days and Daily Volatility is computed here as an exponentially weighted moving standard deviation of Daily Return, with a span of 22 days (approximately one business month).

Lastly, as with sentiment data, the dataset is divided into train and test sets for the subsequent data preparation, with the ratios expressed in the previous subsection.

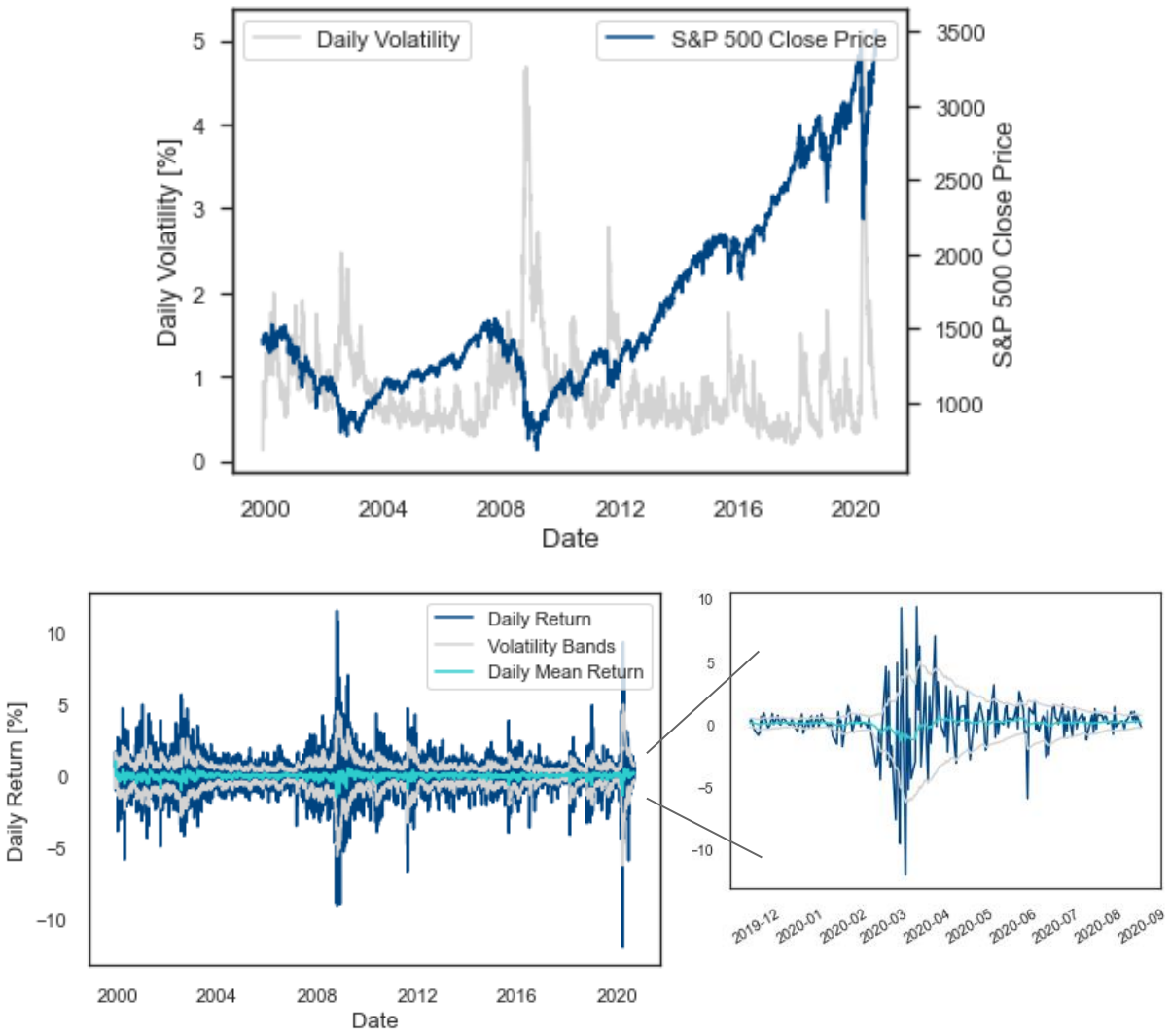


Figure 4. S&P 500 index, daily returns and daily volatility.

On the upper half, daily values of S&P 500 index with absolute volatility estimates of daily returns. Note the higher volatility peaks during bear market[21] periods. On the left bottom half, daily return calculations and daily volatility bands at $\hat{\mu} \pm 1\hat{\sigma}$*, with a zoom into the last 10 months of data on the right. Note that this is not intended as a representation of volatility bands derived from Bollinger Bands[22]; this is solely a chosen form of visualization for the calculations under consideration.*

[21] A bear market is a general decline in the stock market over a period of time. It includes a transition from high investor optimism to widespread investor fear and pessimism. One generally accepted measure of a bear market is a price decline of 20% or more over at least a two-month period. The opposite concept is known as a bull market.

[22] bollingerbands.com/bollinger-bands

Data preparation

Both the sentiment dataset and the financial dataset are subjected to notable data preparation, including imputation of missing values, smoothing, feature engineering and statistical transformations. All implementations are applied first on each of the train sets and then on the test sets. After this separate manipulation, the sentiment data and the financial data are joined together into final train and test sets.

Missing values

Missing values have low rates, yet they are present in both datasets. The whole sentiment dataset has 1% of missing cells when accounting for the total number of cells, with a median of 0% and a mean of 2% across all variables. From the total of 90, 36 variables have missing values, with 25 under a 2% rate and 2 of them exceeding a rate of 20% -- carryTrade_USD[23] (68%) and currencyPegInstability_USD[24] (43%). These last ones are dropped. On the other hand, under a timeline perspective, missing rates are higher during the first years, as can be seen in Figure 5. This is most certainly due to the fact that in 2005 the archive began including Internet news. Also, in 2009, tweets were included. For the remaining variables, imputation is done via forward filling with the previous registered value. For understanding this decision, one has to comprehend the underlying mechanism that triggers the presence of missing values in the first place.

As mentioned earlier in this Chapter, all TRMIs are based on relevant text collected over a window of content. If over that window there was no relevant text identified for a

[23] Carry Trade is a strategy used in the forex trading market whereby an investor sells a certain currency with a relatively low interest rate and buys a different one with a higher interest rate. It is a form of financial arbitrage that is designed to exploit the result of differences between central bank interest rates.

[24] A currency peg is a policy in which a national government sets a specific fixed exchange rate for its currency with a foreign currency or a basket of currencies. Pegging a currency stabilizes the exchange rate between countries. (Investopedia)

particular indicator, then the correct index value is NA[25], not zero, and the index will appear blank.

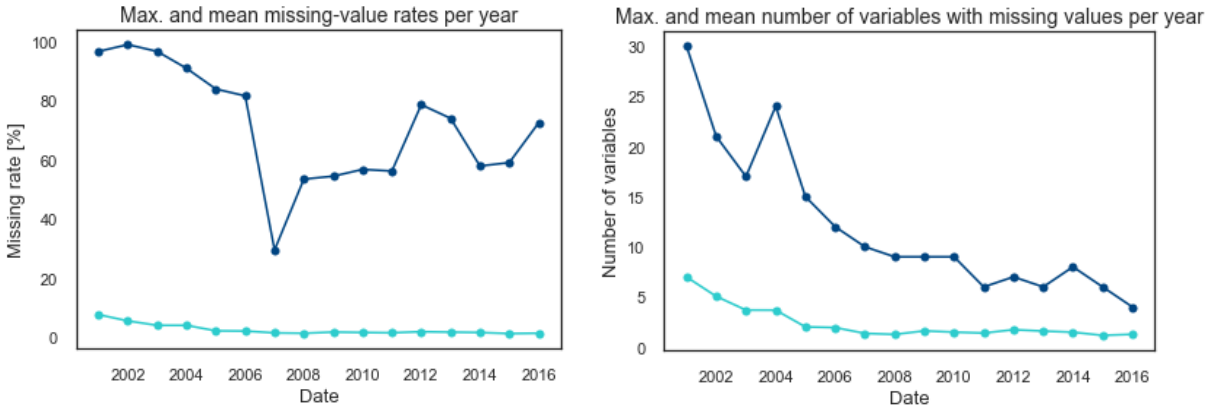


Figure 5. Evaluation of missing values per year.

On the left, the dark blue line represents the maximum missing rate found per year; the aquamarine line represents the mean missing rate across all variables per year. On the right, the dark blue line shows the maximum number of variables that have missing values in the corresponding year; the aquamarine line indicates the mean number of variables having missing values.

Since, as seen in Figure 5, the probability for a value of being NA is highly dependent on time, which is an *observed* variable, one could confidently state that it is not a case of missing completely at random (MCAR) values for any of the variables. As to further confirm under the Rubin (1976) [17] classification, whether they are cases of missing at random (MAR) or missing not at random (MNAR), it is a slightly more complex matter and the situation is different for every variable[26]: there are some metrics, like the discarded carryTrade_USD, which are highly specific and technical, that will have less propensity of appearing on news and thus be less likely to have a measurement.

[25] NA differs in meaning from true zero in that true zero represents the presence of text corresponding to positive and negative values that add up to zero. In other words, a zero value reflects that relevant text was found and its sentiment implications net to zero. In contrast, NA represents the absence of any relevant text and of any resultant measurement.
[26] iriseekhout.com/missing-data/missing-data-mechanisms/assuming-a-missing-data-mechanism/

Notwithstanding, the implementation of a forward filling for the remaining variables comes pretty straightforward and is built under the assumption that in a day where no relevant information has been found, previous registered content will still have an impact and accordingly be the prevailing information.

By contrast, for the financial dataset, no missing data is found at random. All observed missing values belong to stock market holiday closure days for U.S. stock exchanges[27] and account for 2% of the total cells in the dataset and a 4% rate per variable. Instead of being dropped, this "missing data" is imputed by means of a forward cubic spline interpolation on the premise that it will improve the results. Weekends are not included in the imputation as they were prior eliminated.

Smoothing

Sentiment data has a very noisy nature, with large fluctuations between consecutive days. The presence of noise in the training data generally reduces the accuracy of the learned predictors. In general, noise is present in both the predictors and in the target variable. From these two kinds of noise, typically, the latter has a more pronounced misleading effect than the former. The explanation of this observation is that, generally, noise in the target values causes a large distortion of the regularity patterns that are exploited for prediction. By contrast, noise in the input variables tends to simply blur these patterns. However, sizable levels of noise in the features can have as adverse an effect as target noise [18]. Because of this, sentiment variables have been denoised or smoothed by applying an **exponentially weighted moving average** (EWMA) with a span of 22 days (approximately one business month). For clarification, the EWMA differs from a simple moving average in that the latter regards each point in the data

[27] A stock exchange is a facility where stockbrokers and traders can buy and sell securities, such as shares of stock, bonds, and other financial instruments.

window to be equally important when calculating the average (smoothed) value, and the former places more emphasis on the most recent data. In a dynamic system, the most current values tend to better reflect the state of the process. Hence, a low-pass filter like the EWMA that applies weighting factors that decrease exponentially is therefore more useful. Figure 6 shows an example of this transformation on variable Sentiment_US500. On the other hand, no smoothing was carried out on financial data.

Figure 6. Comparison of raw and smoothed sentiment data for a 1-year time window.

On the left, index at Close is represented with one raw variable from the sentiment data, Sentiment_US500. On the right, the index at Close is represented with Sentiment_US500 after undergoing dispersion and scale transformations, and applying the exponentially weighted moving average with a span of 22 days. Note the better appreciation of the underlying sentiment pattern on the transformed variable.

Feature engineering

Principal Component Analysis (PCA) with 3 dimensions was performed on sentiment data. Briefly, PCA is a processing technique that reduces data by geometrically projecting them onto lower dimensions called principal components (PCs), with the goal of finding the best summary of the data using a limited number of PCs. The first PC is chosen to minimize the total distance between the data and their projection onto the PC. By minimizing this distance, the variance of

the projected points is also maximized. The second (and subsequent) PCs are selected similarly, with the additional requirement that they be uncorrelated with all previous PCs [19].

First resultant PC for sentiment data retains a variance of 42%, with the second and third PCs totaling a 15% and a 7%, respectively. These principal components were adjoined to the sentiment dataset as part of the predictors, adding up to 93 input variables. However, it is noteworthy that the variance retained by the PCs is low.

At this point, besides PCs, two more variables were added to this dataset by analogy with a popular trading strategy developed by technical analysis: **crossovers**. In stock investment, as previously introduced in Chapter II, technical analysis seeks to predict price movements by examining historical data. One classical and simple tool is to use moving averages (MA). There are various related strategies with moving averages, but crossovers are one of the main ones. The process for a moving average crossover is to apply two moving averages to a chart: one longer and one shorter. When the shorter-term MA crosses above the longer-term MA, a Buy signal is produced (+1), as it indicates that the trend is shifting up. Meanwhile, when the shorter-term MA crosses below the longer-term MA, a Sell signal is triggered (-1), as it indicates that the trend is shifting down. In adapting this fundamental technique to the sentiment data, instead of a conventional MA, the previously proposed EWMA is used. Regarding the spans, for the shorter-term EWMA, the smoothed variables themselves are taken (remember that a previous EWMA has been applied as a low-pass filter); for the longer-term EWMA, variables were newly smoothed out with a span of 60 days (approximately 3 business months). Crossovers were calculated for predictors Sentiment_US500 and StockIndexSentiment_USA, which account for overall positive references, net of negative references, in the Companies and Countries assets, respectively. The selection of these two predictors is linked to business judgement, as they were found to be meaningful in prior analysis to this work. Finally, adding these new crossover variables makes a total of 95

predictors from the sentiment dataset. An example of crossovers can be seen in Figure 7. In an ensuing section it will be shown how this "tuned" crossover sentiment strategy proves to be exceptionally useful for the objectives of this thesis.

In regard to the financial variables, several **technical indicators** widely used by traders to check for bearish or bullish signals were added: Relative Strength Index (RSI), Rate Of Change (ROC), Stochastic Oscillator, Williams %R, Moving Average Convergence Divergence (MACD) and Average Directional Movement Index (ADX). Description of the indicators can be consulted in Appendix B. Calculation formulas can be consulted in [20]. In addition, a binary variable was added for signaling whether Close price is higher or not than Open price for each of the observations. Adding these indicators accounts for a total of 11 variables in the financial dataset.

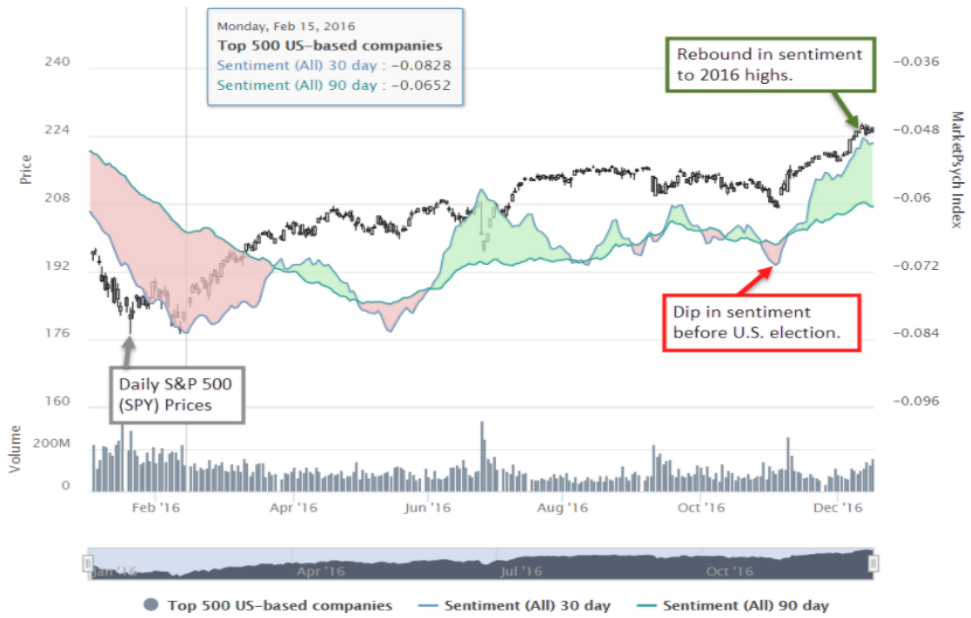


Figure 7. 2016 S&P 500 Media Sentiment.

This image is taken from the MarketPsych Newsletter[28]*. It depicts their financial media-derived sentiment data versus the S&P 500 (SPY) daily prices in 2016, and it is a visual representation of crossovers. Note that prices*

[28] marketpsych.com/newsletter/62/the-seasons-of-social-change-s-p-500

are represented here as a candlestick chart. Attention should also be drawn to the fact that this is not the index itself, but an ETF[29] called SPY that tracks the index. As a curiosity, SPY is the largest ETF in the world ranked by assets under management.

Transformations

All variables from both datasets were subjected to transformations to equal dispersion and scale. Firstly, the **Yeo-Johnson power transform** was applied to stabilize variance, make the data more gaussian-distribution-like and improve the validity of measures of association. Secondly, all variables were standardized[30]. In many models, this is a crucial step, as values of variables with much larger scale than the others tend to predominate in the results, masking the influence of the other variables. In the case of XGBoost though, this is not an issue as it is not affected by monotonic transformations of variables. **Standardization** was applied nevertheless for facilitating the comparison of distributions amongst variables.

Sampling observations

CUSUM filter was applied to the dataset for eliminating irrelevant observations. The CUSUM filter is a quality-control method, designed to detect a shift in the mean value of a measured quantity away from a target value, thus falling under the category of event-based sampling. The target value or threshold used is the daily volatility of the series computed in a previous section, and the measured quantity is the log-return of the price series of the S&P 500 Close. Details of this method can be consulted in [7]. The premise for using this method is that ML algorithms achieve highest accuracy when they attempt to learn from relevant examples. This

[29] See Chapter II for a previous introduction of ETFs.

[30] As there is sometimes a misconception with this terminology, just to clarify, standardization for any variable here results in a mean of 0 and a standard deviation of 1.

also follows business intuition, as portfolio managers typically place a bet after some event -- like the release of some macroeconomic statistics or a spike in volatility -- takes place.

Chapter IV.

Building the model

After the comprehension of the underlying business needs, the exploration of data and the execution of the necessary cleansing and adaptation both for the training and test datasets, it is time for the next step in this data science project, the construction of a final model. Chapter IV addresses the initial issues involved in this aim, namely labeling and variable selection; and eventually describes the chosen set of algorithms to deal with the prediction task.

Labeling

It was mentioned in Chapter III that the intended target which is to be predicted is the future behaviour of the S&P 500, using a special discretization of the index that follows some innovative methods that are described in this section. It is now the moment to elucidate.

First and foremost, it is noteworthy that the forecasting problem is treated as a classification problem, thereby being solved by supervised learning algorithms. This was decided in order to minimize forecasting error -- for instance, it may be harder to forecast the S&P 500 return of tomorrow than its sign --, and because a discrete signal that tells whether to Buy or Sell is easier to incorporate into asset management decisions.

Consequently, before delving into the supervised learning algorithm, assessment of the labeling of the original target variable (the S&P 500 price at daily Close) is to be done. As previously mentioned, the labeling methodology here followed is based on the ideas proposed by López de Prado in his book *Advances in Financial Machine Learning* (2018) [7].

Dealing with data structures and statistical properties

A common approach in ML papers related to finance is to label their observations with the **classical fixed-time horizon**, whereby an observation whose return over a fixed time window surpasses a certain threshold is given a label from the set {-1, 1} based on the sign of the return, and 0 otherwise -- in many cases not even a threshold is used, which leads to many labels referring to non-meaningful, small price changes. This method is regarded by López de Prado as overly simplistic in a real setting, signaling various main flaws:

- First, the use of a constant threshold regardless of the observed volatility, will in some cases cause labels to be 0 even when return is predictable and statistically significant. This may be resolved by using a dynamic threshold with daily volatility estimates (as computed in the section *Data Description*, subsection *Financial variables*).
- Second, every investment strategy has stop-loss limits, designed to limit an investor's loss on a security position when prices are falling; and may also have the contrary limits, designed to take profits before they vanish. These orders are set to automatically close the position when the price reaches either limit, and are not accounted for in the fixed-time horizon method.

Furthermore, there is a popular saying that states that a model is only as good as the data it is fed, so understanding the nature of the data is paramount. The fixed-time horizon method is frequently used with time bars (remember the definition of bar from subsection *Financial variables*). Time bars are obtained by sampling information at fixed time intervals, e.g., once every day, and although they are the most commonly used in literature (note that they are also used in this study), they are neither in accordance with the rhythm of markets in terms of activity nor have adequate statistical properties. To deal with the **heteroscedasticity** and non-stationarity of returns -- as stationarity is a necessary attribute for inference --, the use

of volume bars or dollar bars instead of time bars is recommended. Volume bars circumvent the problem by sampling data every time a certain amount of a security's units (i.e., shares, futures contracts[31], etc) has been exchanged, whereas dollar bars sample observations every time a certain market value is exchanged. Returns computed from dollar bars typically express greater stationarity. Notwithstanding, the use of any of these proposed bar types is out of the scope of this work, as the S&P 500 index is not a tradable financial asset by itself. Data on futures contracts or any other type of financial instrument constructed to indirectly invest or speculate on the S&P 500 would be needed.

At this point, it is relevant to mention that there is also a trade-off between stationarity and memory in time series. Considering degrees of differentiation with an original series of prices, prices have zero differentiation and returns represent a 1-step differentiated series. The greater the step, the greater the stationarity and the lower the memory. López de Prado proposes another method, named fractional differentiation, that aims to find the optimal balance between these opposing factors; the minimum non-integer differentiation necessary to achieve stationarity. This retains the maximum amount of information in the data [21]. However, the use of fractional differentiation is also outside the scope of this work.

The Triple-Barrier Method and Meta-Labeling

Coming back to the task of labeling, the conclusion is that an alternative labeling method is needed. López de Prado thereupon introduces the **Triple-Barrier Method** (TBM), combining the necessity for labeling with real market behaviour. TBM labels observations by using three barriers or thresholds: two horizontal barriers defined by stop-loss and profit-taking limits, which will be simply computed as multiples of the estimated volatility (in this work, they are

[31] A futures contract is a legal agreement to buy or sell a particular security at a predetermined price at a specified time in the future. For instance, S&P 500 futures allow an investor to hedge with or speculate on the future value of various components of the S&P 500 index.

symmetrical and just equal to the volatility, i.e., the value of the multipliers is 1); and a third barrier, a vertical one, which will be put at the end of a time horizon. Whenever the return reaches the upper barrier first, the corresponding observation will be labeled as a 1, because a profit was made. If the lower barrier is touched first, the label will be -1, and losses will be locked in. If the purchase times out before either limit is broken, observations can be labeled as either 0 or as the sign of the return. The approach taken in this work is the latter. Consequently, TBM will compute the labels -1 in the case of a Sell signal and 1 for a Buy signal. The return reached when a barrier has been touched will be called target return. TBM can be visualized in Figure 8.

Figure 8. Visualization of a configuration for the triple-barrier method as extracted from [7]..

Here the triple-barrier method is represented with a symmetric configuration for the horizontal barriers. This is a case where no horizontal barrier has been touched first. The sign of the target return is positive, so this could be labeled as a 1 (Buy).

Hitherto, setting the side of a bet or trade (telling whether to Buy or Sell) has been dealt with, but investors are also faced with the situation of having to adjust the size of that bet, which includes the possibility of no bet at all. Once a model has been trained for setting the

side of a trade with the TBM labels, a secondary model can be trained to set the size. This accepts the primary exogenous model as input. Data must then be labeled again, via a method that López de Prado calls **Meta-Labeling**. This second strategy assigns labels to trades of either 1 (if the bet is taken) or 0 (if the bet is not taken); that is, the ML algorithm will be trained to decide whether to take the bet or pass, another purely binary prediction. This will be done by evaluating the forecasted labels from the previous model and assessing whether the target return (the one achieved when touching a barrier) for that observation is positive or negative. If the target return is positive and the predicted outcome is 1, or if the target return is negative and the predicted outcome is -1, then the decision is to take the trade (1). Otherwise, if the sign of the target return is not the same as the predicted label, then the decision will be to not act on that bet (0), as this means that the first prediction is wrong. When the predicted label is 1, the probability of this second decision can be used to set the size of the trade, where the side (sign) of the position has already been set by the primary model. An initial naive approach may be to discretize this probability within ranges that fall into the set of categories {Strong Buy, Buy, Sell, Strong Sell}, but setting the size up until that point is not contemplated in this work.

To recapitulate, the final combined decision of the models will be the side of a possible trade, and whether an investor should act or pass on the presented opportunity, i.e., {Buy (1), Do Nothing (0), Sell (-1)}. See Figure 9 for a visual representation.

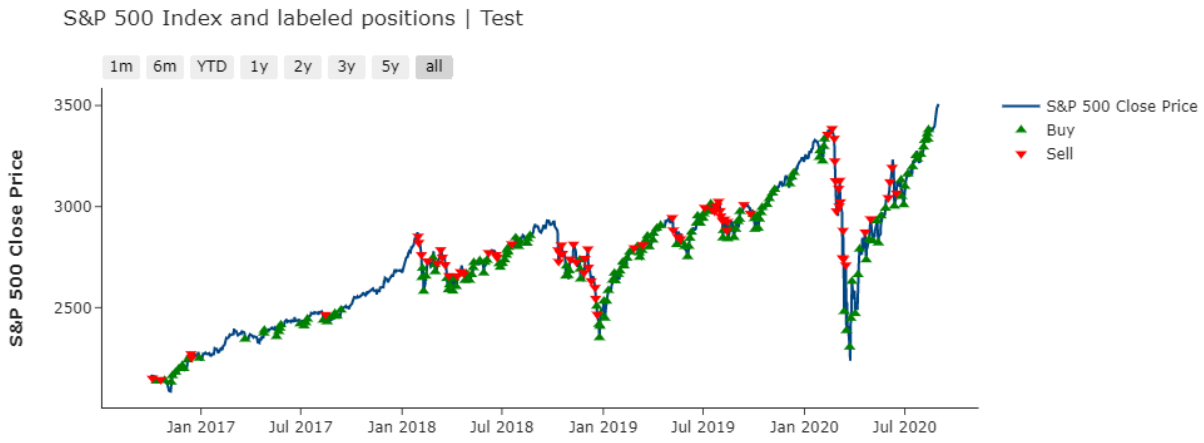


Figure 9. Labeled final positions for the test set.

*Example of final desired positions derived for the test set with the combined labels TBM * Meta-Labeling in a scenario with no profit-taking and stop-loss limits, a horizon of 14 natural days and a minimum return (threshold) of 0.5%. Note that only the positions where a bet is taken are shown. This represents the ideal outcome of the strategy in that scenario.*

Meta-labeling ML models can deliver more robust and reliable outcomes than standard labeling methods. Furthermore, the justification for the use of a meta-labeling strategy as a second layer in the creation of a model with the conditions exposed in previous sections can be summarized in a set of quantifiable items:

• First, meta-labeling is particularly helpful for achieving higher F1-scores (a discussion on scores and model evaluation methods is offered in the next chapter), as it will filter out false positives.

• Second, it provides the possibility of building an ML system on top of a fundamental or technical model. That is, the primary model can be an ML model like in the case of this work, but it could also be an events-based strategy where the portfolio manager generates the side signal, for instance.

• Third, the effects of overfitting are limited, since it will only decide the size, not the side.

• Fourth, meta-labeling allows for building sophisticated strategy structures by decoupling the side prediction from the size prediction. For instance, the explanatory variables driving a Buy signal might differ from the ones indicating a Sell. In that case, an ML strategy may be developed exclusively for long positions, and another ML strategy specifically for short positions, based on entirely different primary models.

• Lastly, in the words of López de Prado, "as important as it is to identify good opportunities is to size them properly", since, for example, achieving high accuracy on small bets and low accuracy on large bets will result in great capital losses.

In general, and as a final conclusion for this section, meta-labeling is definitely a powerful tool to have in the arsenal.

Labeling was created for **two scenarios**: a scenario with profit-taking and stop-loss limits and a scenario with none of these limits (the multipliers of the volatility for setting the width of the horizontal barriers are zero). The maximum time period or time window for the trades is 14 natural days. That is, predictions will be made for a maximum of 14 natural days if no limits are reached or imposed. Also, a minimum return threshold of 0.5% required to run the search for triple barriers was applied. Labeling was achieved with the use of the Python library mlfinlab[32] by Hudson and Thames[33] and the code provided in López de Prado's book [7].

[32] mlfinlab.readthedocs.io/en/latest/index.html
[33] hudsonthames.org

Variable selection

ML models benefit from the selection of a reduced number of variables in terms of interpretability and performance. A collection of predictors might very well contain non-informative variables or noise and this might impact performance to some extent. Feature selection helps in reducing the generalization error of the model by removing these irrelevant features. Moreover, as demonstrated by Max Khun and Kjell Jhonson in their book *Feature Engineering and Selection: A Practical Approach for Predictive Models* (2019) [22], the effect of feature sets can be much larger than the effect of different models, so choosing the right set of relevant features for the prediction task at hand is of great importance. Besides, in terms of computational complexity, feature selection will improve efficiency. Similarly, dependence over a smaller set of predictors or explanatory variables is easier to describe and interpret.

In this work, variable selection has been attained by means of a sequential feature selection algorithm, which uses greedy search to reduce an initial d-dimensional feature space to a k-dimensional feature subspace where k < d. Concretely, **Sequential Forward Floating Selection** (SFFS) has been used. Sequential Forward Selection (SFS) algorithms add one feature at the time based on the classifier performance until a feature subset of the desired size k is reached. The *floating* variant, SFFS, can be considered as an extension of SFS. This variant has an additional inclusion step to remove features once they were included, so that a larger number of feature subset combinations can be sampled. It is important to emphasize that this additional step only occurs if the resulting feature subset is assessed as "better" by the criterion function after addition of a particular feature [23]. The number of k features (a parameter of the algorithm) was selected to be in the range (10, 30), and 5-fold cross-validation was used. The classifier used here to assess performance of feature subsets prior to applying them on a final model is a Random Forest of 40 base estimators and a maximum depth of 2 nodes, and the criterion function is the negative log-loss. An explanation of the Random Forest algorithm

and the negative log-loss function is given in the next sections. The labels used in the Random Forest algorithm were that of the primary model (the ones obtained with TBM for setting the side of the trades).

Two initial feature d-dimensional spaces were evaluated: a feature space consisting of only the predictors from the sentiment dataset, and a feature space consisting of the entire final dataset, containing both sentiment and financial data. Also two scenarios were considered, one with the existence of profit-taking and stop-loss limits, and another without them (which is somewhat an analogous configuration to the previously introduced fixed-time horizon). As a result, four k-dimensional feature subspaces were obtained, where k is in the set {11, 13}.

Extreme Gradient Boosting

In the section *Labeling*, labels for learning the side of a trade and labels for learning the size of such trade were obtained. The same ML algorithm was applied for learning both binary classification labels: **Extreme Gradient Boosting** (hereinafter XGBoost).

XGBoost is a popular ensemble learning algorithm used in prediction tasks, so to better introduce XGBoost, let us first bring in the concept of ensemble learning methods. Ensemble methods combine a set of weak learners[34] (e.g., a learner with high bias[35]), all based on the same learning algorithm, in order to create a (stronger) learner that performs better than any of the individual ones. They are broadly employed nowadays for its predictive performance progress, and can help reduce bias and/or variance[36] [37]. Commonly used algorithms for creating an ensemble of weaker learners are bootstrap aggregating (bagging) and boosting.

[34] Terms earner and estimator are used interchangeably throughout the document.

[35] This error is caused by unrealistic assumptions. When bias is high, the ML algorithm has failed to recognize important relations between features and outcomes. In this situation, the algorithm is said to be "underfit."

[36] This error is caused by sensitivity to small changes in the training set. When variance is high, the algorithm has overfit the training set, and that is why even minimal changes in the training set can produce wildly different predictions. Rather than modelling the general patterns in the training set, the algorithm has mistaken noise with signal.

[37] A review of the bias-variance tradeoff can be found in scott.fortmann-roe.com/docs/BiasVariance.html.

Bagging is perhaps one of the earliest and simplest ensemble methods. In a classification task, for example, it works as follows: first, N bootstrapped replicas of the training data are generated. That is, N different training data subsets are randomly drawn – with replacement – from the entire training dataset. Second, each training data subset is used to train a different classifier of the same type. Third, the N resulting individual base classifiers are then combined by taking a simple majority vote of their decisions. When the base estimator can make forecasts with a prediction probability, the bagging classifier may derive a mean of the probabilities. **Random forests** (RF) [24] -- used in the previous section for variable selection -- are combinations of decision trees (hence the name forest), which predict a target value by learning easy decision rules formed from the data features. RF shares some similarities with bagging, in the sense of training independently individual estimators over bootstrapped subsets of the data. But they incorporate a second level of randomness: when optimizing each node split in a tree, only a random subsample (without replacement) of the attributes will be evaluated, with the purpose of further decorrelating the base estimators. i.e., increasing diversity, which in turn will increase effectiveness of the ensemble model.

The concept of **boosting**, on the other hand, is to sequentially train weak learners to correct their past performance. **Adaboost**, for example, one of the first boosting algorithms, works as follows, as described by López de Prado: first, a training set is generated by bootstrapping, according to some initial sample weights. Second, an estimator is fit using that training set. Third, if the single estimator achieves an accuracy greater than the acceptance threshold (e.g., 50% in a binary classifier, so that it performs better than chance), the estimator is kept, otherwise it is discarded. Fourth, more weight is given to misclassified observations, and less weight to correctly classified observations. Fifth, the previous steps are repeated until N estimators are produced. Lastly, the ensemble forecast is the weighted average of the individual forecasts from the N models, where the model weights are determined by the

accuracy of the individual estimators. By contrast, **Gradient Boosting** [25] fits a new estimator of the residual errors (made by the prior estimator) using gradient descent to find the failure in the predictions of the previous learner, minimizing a given loss function (i.e., log-loss), and all trees are scaled by the same weight.

Boosting's main advantage is that it reduces both variance and bias in forecasts (variance by combining a set of learners, and bias by correcting for the predictions of each prior learner in the sequence), but it may be prone to overfitting. **XGBoost** [26] is an ensemble tree method that was introduced for better speed and performance, and its trees are somewhat characteristic. In-built cross-validation ability, efficient handling of missing data, regularization for avoiding overfitting, or parallelized tree building are common advantages of the XGBoost algorithm.

Delving more into the inner workings of these models is not in the scope of this project, but mathematical descriptions and details for all the algorithms can be found in [18, 26, 27]. Also, the reader may want to check the fantastic YouTube channel *StatQuest*[38], whose mission is to break down major methodologies into easy-to-understand pieces, and, for our interest here, does a great job at explaining boosting algorithms.

Model architecture

As explained in previous sections, the final model of this work is a combination of two ML models, a result of dividing the prediction task into two problems: setting the side of a bet or possible trade (i.e., predicting either 1 or -1), and setting the size -- referred here as to whether to take such a bet or not (i.e., predicting either 1 or 0).

[38] https://www.youtube.com/c/joshstarmer/featured

XGBoost was used for learning both tasks, with 3-fold cross-validation built with class `sklearn.model_selection.`**`TimeSeriesSplit`**`()` for hyperparameter tuning. Tuned parameters were the number of estimators, the maximum depth of the trees, the learning rate[39]and the number of early stopping rounds[40]. Ad-hoc configurations were also provided for parameters min_child_weight (minimum sum of instance weight[41] needed in a child) with a value of 2, and scale_pos_weight (which controls the balance of positive and negative weights). This last parameter was set to the value of `sum(negative instances or zero instances) / sum(positive instances)` to control for a (very) moderate class imbalance. Notwithstanding, setting this balance control on weights came with the cost of lowering recall or sensitivity of the positive class, over a higher recall on the negative class and a higher global AUC (these metrics will be defined in the next chapter). Negative log-loss or negative cross-entropy loss was used as the evaluation metric. This metric is defined on probability estimates and for a binary classifier with a true label $y \epsilon \{0,1\}$ and a probability estimate $p = Pr(y=1)$, the log loss per sample is the negative log-likelihood[42] of the classifier given the true label:

$$L_{loss}(y,p) = -log\ Pr(y \mid p) = -(y\ log(p) + (1-y)\ log(1-p)).$$

With the total log-loss being the sum of all the individual scores divided by the number of samples. Log-loss will return high absolute values for bad predictions (low negative values) and low absolute values for good predictions (high negative values). This metric is a good ally

[39] In this context, the learning rate is the step size shrinkage used in update to prevent overfitting. After each boosting step, one can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative.

[40] Early stopping is a form of regularization used to avoid overfitting when training a learner with an iterative method, such as gradient descent. Such methods update the learner so as to make it better fit the training data with each iteration. Up to a point, this improves the learner's performance on data outside of the training set. Past that point, however, improving the learner's fit to the training data comes at the expense of increased generalization error. Early stopping rules provide guidance as to how many iterations can be run before the learner begins to overfit.

[41] If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning. The larger min_child_weight is, the more conservative the algorithm will be.

[42] The reader may find clarifying to check: en.wikipedia.org/wiki/Likelihood_function#Log-likelihood.

for investment applications, since it extra penalises bad predictions that have a high confidence on the prediction --- a situation that could incur in great capital losses for an investment strategy. Accuracy (the percentage of correctly classified observations) -- often used on classification tasks --, on the other hand, accounts equally for an erroneous prediction with either high or low probability.

As mentioned in the section *Labeling*, forecasts generated with the primary model for setting the side were used as an extra input for the meta-model besides the previously selected features.

Chapter V.

Results and discussion

Using the predictions result from the models, a decision on the possibility of a trade can be made. As stated previously, two scenarios were considered, one with profit-taking and stop-loss limits, and another without them. The time horizon for evaluating the strategy is 14 natural days. In the scenario without limits, it follows that the time horizon for holding the trades is held constant for all observations. Conversely, when limits are implemented, time exits for the trades vary across observations. This is the reason why an "unreal" scenario without profit-taking or stop-loss limits was also created for assessing the usefulness of sentiment data, as an initial speculation is that this added irregularities further complicate the task of prediction.

First and foremost, various metric concepts need to be clearly defined:

• **Accuracy**: as mentioned before, measures the portion of all testing samples classified correctly.

• **Recall** (also known as sensitivity): measures the ability of a classifier to correctly identify positive labels.

• **Specificity**: measures the classifier's ability to correctly identify negative labels.

• **Precision**: measures the proportion of all correctly identified samples in a population of samples which are classified as positive labels.

• **F1-score**: this is the harmonic mean of precision and recall (measures efficiency).

• **AUC**: when using normalized units, $\epsilon[0, 1]$, the area under the curve (often referred to as simply the AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'). The "curve" is the receiver operating characteristic (ROC) curve, which is a plot of the recall as a function of fall-out or false

positive rate (1 - specificity). In other words, AUC measures how good the classifier is in separating the two classes.

Technical data is used as a sort of benchmark for comparison. Notwithstanding, the created technical indicators strongly depend on the periods taken for calculation, thus more tests could be performed. For an "unreal" scenario with no profit-taking or stop-loss limits and thus constant time horizon across observations of 14 days, the selection of features that contains both sentiment data and financial data achieved the best overall results (see Table 1). The effect of meta-labeling can also be fully appreciated, increasing accuracy from 50% to 68% and AUC from 0.53 to 0.73. Results with low specificity and very high recall were obtained prior to addressing class imbalance by changing sample weights (as explained in the previous section). This adjustment resulted in higher accuracies and AUC values. It can be derived from AUC values that the incorporation of sentiment data increases the ability of the classifier to discriminate between classes. Also, differences in the confusion matrices and in the recall and specificity between subsets of explanatory variables suggest that a well-suited option for future exploration could be to create separate specialized models for predicting each class. This situation was previously explained in the fourth item of the list of advantages that a meta-labeled strategy poses (Chapter IV, section *Labeling*).

In the scenario with profit-taking and stop-loss limits, and thus varying holding periods for the trades, as expected, scores are mostly inferior (see Table 2), with no better than random results for the primary model. The sentiment subset achieved a higher AUC in the meta-model than the rest of feature subsets, mostly due to a higher recall. One could speculate that in this context of varying periods, sentiment data reaches higher performance because of its rapidly changing nature and a quick adjust to price movements.

Table 1. Test results for scenario “fixed-time horizon” for setting side (upper half) and size (lower half).

Metric	Selection from sentiment features	Financial or technical features	Selection from entire dataset
Accuracy	49%	30%	50%
F1-score (weighted by support)	51%	14%	52%
AUC	0.53	0.46	0.53
Confusion matrix	39 38 91 87	77 0 178 0	47 30 98 80
Accuracy	61%	37%	68%
F1-score (weighted by support)	60%	53%	67%
AUC	0.62	0.52	0.73
Confusion matrix	59 70 30 96	26 152 10 67	103 25 57 70

Table 2. Test results for scenario with profit-taking and stop-loss limits for setting side (upper half) and size (lower half).

Metric	Selection from sentiment features	Financial or technical features	Selection from entire dataset
Accuracy	52%	37%	43%
F1-score (weighted by support)	52%	24%	43%
AUC	0.43	0.48	0.45
Confusion matrix	26 64 60 108	88 2 160 8	51 39 109 59
Accuracy	61%	39%	44%
F1-score (weighted by support)	58%	24%	40%
AUC	0.64	0.53	0.58
Confusion matrix	46 78 24 110	5 157 1 95	29 119 25 85

It is also interesting to compare the different feature importances and feature subsets shown in Figure 10 and Figure 11. For instance, sentiment on volatility related to the USD plays a major role in this second setting, whereas sentiment on fundamental strength (positivity about accounting fundamentals, net of references to negativity about accounting fundamentals) is of primary importance for the first setting. Both may be reasonable outcomes: in the second setting, labels have been derived from limits imposed with multipliers to daily volatility, which may have a certain correlation with volatility on the USD (as seen in Chapter II, volatility merely refers to the price action: how rapidly or severely it may change). In the case of the first setting, where observations are labeled generally for a longer term (maximum horizon, i.e., 14 natural days), fundamentals may come into play. However, note that these assumptions are purely speculative and further analysis, tests and research on the nature of the explanatory variables and their interplay is needed. A study on the distributions of trading periods could be performed as well. At the same time, another interesting result is that, for both settings, the engineered crossover related to stock index sentiment in the USA is of great significance in deciding whether to take the bets or not. Figure 12 may be exemplifying for this.

In comparing results between train and test sets, no signs of overfitting were found. For illustrative purposes, in the context with no horizontal limits imposed where best AUC was obtained, train accuracy (computed with cross-validation) and test accuracy is 50% on both datasets for the primary model, and 57% and 68% (respectively) for the meta-model. This implies that generalization is obtained.

It is also noteworthy, as mentioned in the section *Labeling*, that López de Prado's intention towards meta-labeling is to create a primary model that reaches a high recall, even if the precision is not very high. This is particularly helpful for achieving higher F1-scores. Low precision is corrected by applying meta-labeling to the positives predicted by the primary model. Notwithstanding, and in the experience of this work, this can lead to models where no

identification of Sell opportunities (if Sell is the negative class) is obtained. Depending on the business strategy or need under consideration, this may not be adequate. Be that as it may, having a high recall at least assures that the first model is really good at identifying Buy betting opportunities (in the case where Buy is the positive class), and then the meta-model can boost a low precision. However, it can be hypothesized that this application might be another sign for building separate models for each of the positions at stake -- i.e., long or short. Finally, as an extra note, the reader may also find appealing to check another implementation of López de Prado's devised strategies; e.g., [28], which is from Hudson and Thames.

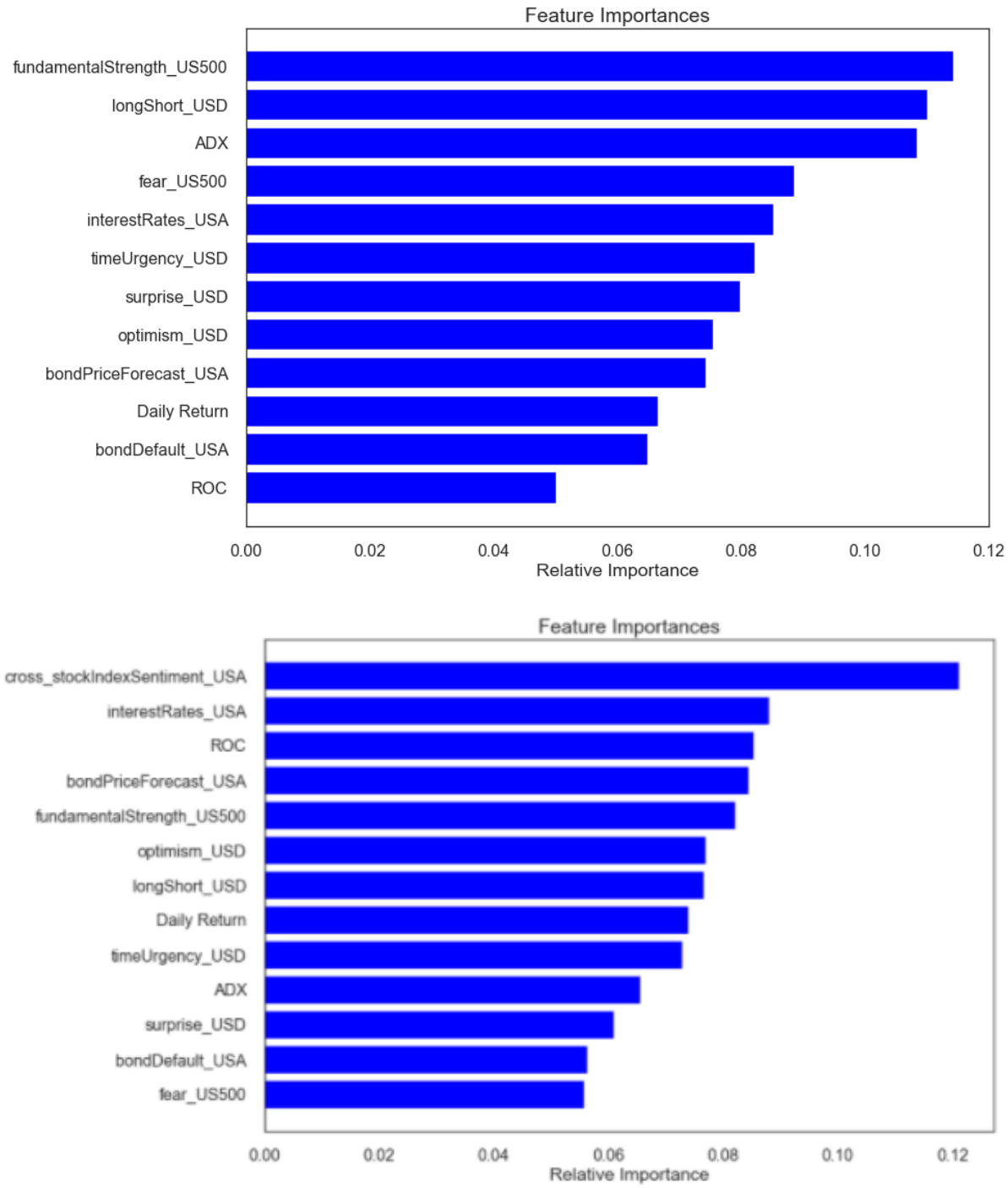
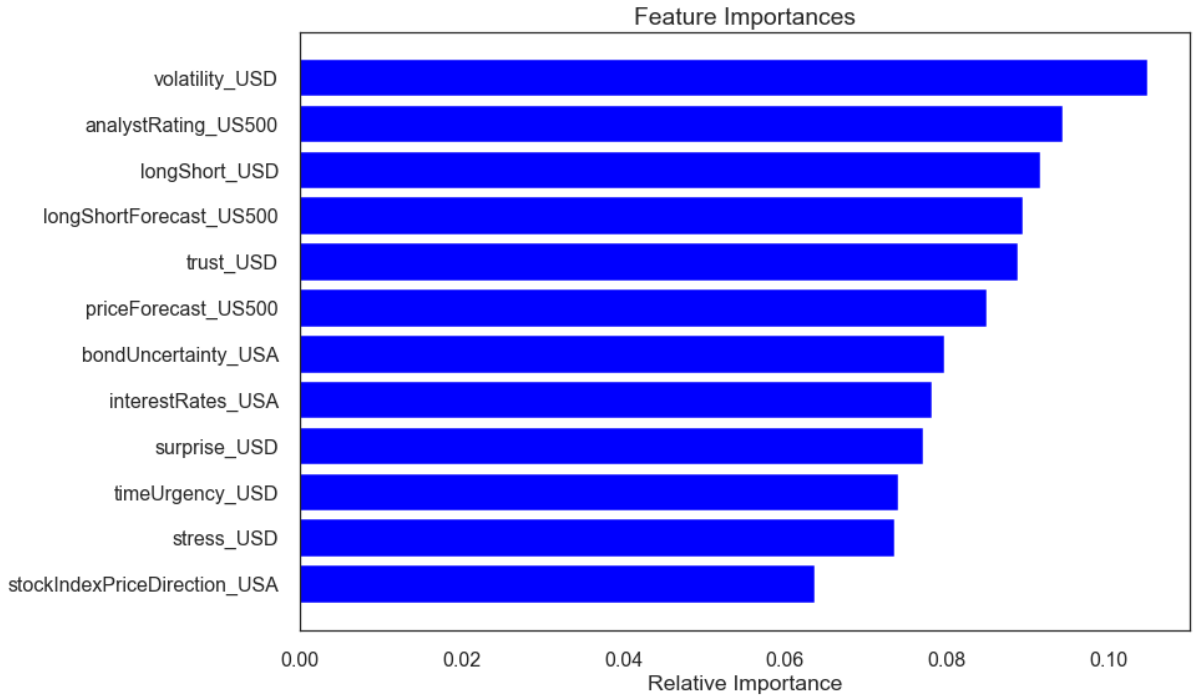


Figure 10. Feature importances (relative to the weight or number of times a feature has been used for splitting nodes) for the scenario without profit-taking or stop-loss limits and a feature subset from the entire dataset. *The first graph corresponds to the primary model and second graph to the secondary model or meta-model.*



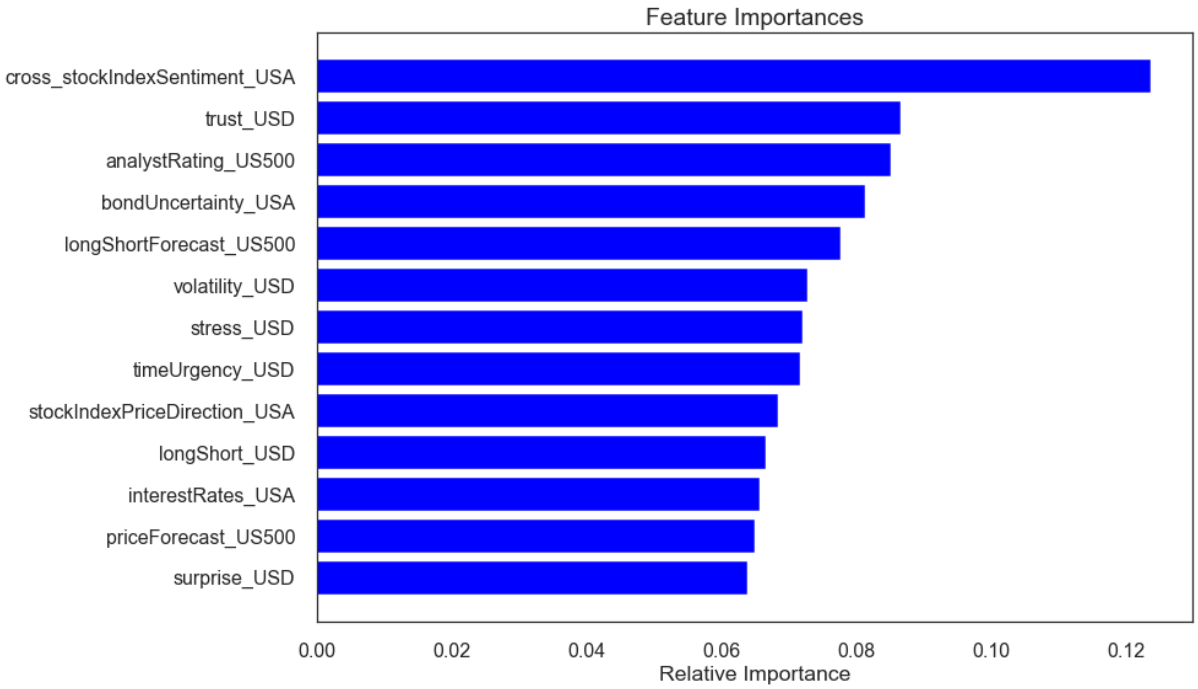


Figure 11. Feature importances (relative to the weight or number of times a feature has been used for splitting nodes) for the scenario with profit-taking or stop-loss limits and a feature subset from the sentiment dataset. *The first graph corresponds to the primary model and second graph to the secondary model or meta-model.*

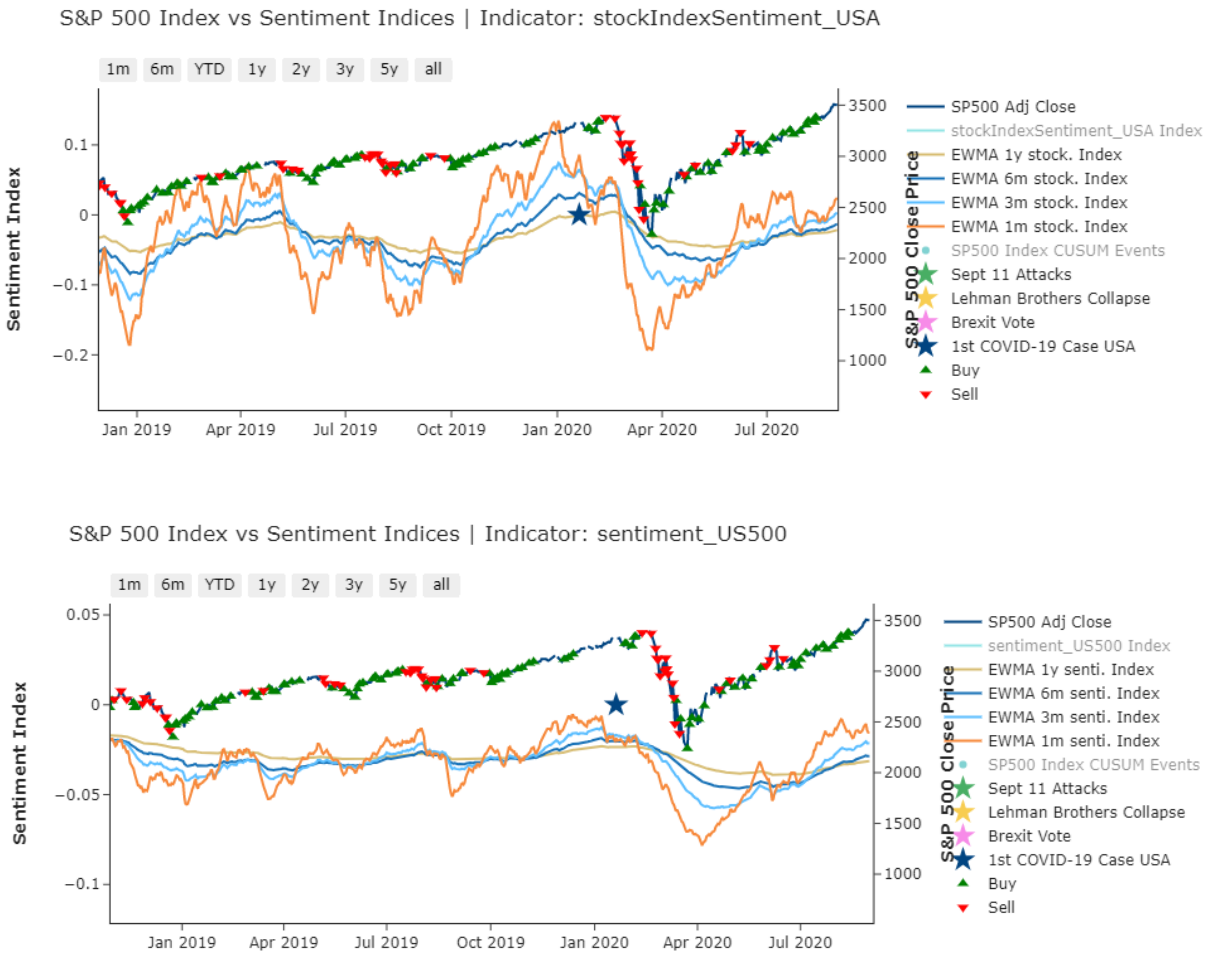


Figure 12. Engineered crossovers from sentiment data.

A representation of test final ideal decisions over possible trades and the engineered crossovers is shown here. Remember that when the fast EWMA crosses over the low EWMA, an uptrend signal is generated (1), and a downward signal otherwise (-1). It can be seen that in some cases, crossovers are ahead of a market trend. Events that appear on the legend like the Lehman Brothers Collapse[43] but take place further back in time (these graphs are extracted from an interactive visualization created with the Python library Plotly) have been used during the exploration phase of sentiment data.

[43] The reader may find this interesting: en.wikipedia.org/wiki/Bankruptcy_of_Lehman_Brothers

Chapter VI.

Conclusions and future lines of work

Financial markets are complex adaptive systems, composed of manifold dynamic networks of interactions. This intricate nature of markets renders the process of accurately predicting their future behaviour an arduous task. Recent progress in machine learning techniques has proven effective in the quest for financial prediction, but developing realistic financial strategies further complicates the issue. Financial ML should be regarded as a subject in its own right. Additionally, to develop theories for prediction, factors involved in the movement of markets should be considered, such as sentiment towards their components. In this work, the usefulness of developing a strategy with sentiment data on the S&P 500 stock market index has been tested with the ideas of TBM and Meta-labeling, and the efficacy of adding a meta-layer to correct for the predictions of a previous model has been shown.

Sentiment data can and should be a part of investment strategies as shown in the results, but intense and careful studies on the interplay and character of explanatory variables ought to be performed for this aim, together with appropriate data preparation. Also, the findings show that sentiment does not rapidly dilute in the market. A test for obtaining the optimal prediction horizon could further be performed. In addition, tree-based models can be of help in deciphering what features are behind a given event, as slightly seen with the importance plots.

The lines of future work are countless and diverse. Further research includes, but is not restricted to, using volume or dollar bars for preparing financial data, better exploiting technical indicators with optimised periods, trying up-sampling techniques in the existence of unbalanced classes instead of applying extra weight in the model, optimizing the smoothing spans in decay coefficients for sentiment data EWMAs, determining optimal multipliers for the horizontal volatility barriers of TBM, exploring other algorithms and model architectures,

investigating the "hidden" sentiment variables involved in a phenomenon in question (such as the market crash caused by the COVID-19), testing extension to other market indices such as the EURO STOXX 50, and using futures contracts or other financial instruments for assessing profitability with strategy performance metrics -- like the annualized return or the Sharpe ratio[44] -- which, in the end, is one of the main objectives in the creation of an investment strategy.

As a final conclusion, *Advances in Financial Machine Learning* and *Machine Learning for Asset Managers* by Prof. Marcos López de Prado should be key pieces to accompany investment professionals in their development of financial strategies, always keeping in mind that ML is a tool that needs to be steered. In his very own words:

> The goal of financial ML ought to be to assist researchers in the discovery of new economic theories. The theories so discovered, and not the ML algorithms, will produce forecasts. (Chapter 1, p.19) [1]

[44] investopedia.com/terms/s/sharperatio.asp

Appendix A.

Sentiment dataset

An explanation on scoring ranges extracted from MarketPsych's User Guide and the total content of features of the sentiment dataset are given here.

Score ranges

As extracted from Thomson Reuters, the indices are marked as ranging from either -1 to 1 or 0 to 1, corresponding to bipolar and unipolar indices, respectively. In practice, those denoted as "unipolar" can in fact range below 0, although not below -1. This occurs because unipolar indices reflect the orthogonal nature of many single emotions and topics. A negative comment such as, "I don't enjoy owning this stock" is not emotively equivalent to, "I am pessimistic about the stock's prospects" or "I am angry with the company's management." The initial statement is specifically one of negative Joy, which decreases the overall Joy index for assets related to that company. When there are many such negative Joy comments for an asset, the Joy index itself may show negative values. Nonetheless, in practice unipolar indices are positive over 90% of the time, because language typically reflects positive assertions. Thus, in the sections below, this range is marked as "0 to 1*".

Features

Table 3. Companies sentiment indicators - US500

Index	Description: references in news and social media to...	Range
analystRating	upgrade activity, net of references to downgrade activity	-1 to 1
anger	anger and disgust	0 to 1*
buzz	buzz	NA
conflict	disagreement and swearing net of agreement and conciliation	-1 to 1
cyberCrime	Cyber attacks and data breaches	0 to 1*

debtDefault	debt defaults and bankruptcies	0 to 1*
dividends	dividends rising, net of references to dividends falling	-1 to 1
earningsForecast	expectations about improving earnings, less those of worsening earnings	-1 to 1
emotionVsFact	all emotional sentiments, net of all factual and topical references	-1 to 1
fear	fear and anxiety	0 to 1*
fundamentalStrength	positivity about accounting fundamentals, net of references to negativity about accounting fundamentals	-1 to 1
gloom	gloom and negative future outlook	0 to 1*
innovation	innovativeness	0 to 1*
joy	happiness and affection	0 to 1*
laborDispute	labor unrest and work stoppages	0 to 1*
layoffs	staff reductions and layoffs	0 to 1*
litigation	litigation and legal activity	0 to 1*
longShort	buying, net of references to shorting or selling	-1 to 1
longShortForecast	forecasts of buying, net of references to forecasts of shorting or selling	-1 to 1
loveHate	love, net of references to hate	-1 to 1
managementChange	changes in a company's management team, net of references to stability in the management team	-1 to 1
managementTrust	trust expressed in a company's management team, net of references to reports of unethical behavior amongst the management team	-1 to 1
marketRisk	positive emotionality and positive expectations net of negative emotionality and negative expectations. Includes factors from social media found characteristic of speculative bubbles – higher values indicate greater bubble risk. Also known as the “Bubbleometer.”	-1 to 1
mergers	merger or acquisition activity	0 to 1*
optimism	optimism, net of references to pessimism	-1 to 1
priceDirection	price increases, net of references to price decreases	-1 to 1
priceForecast	forecasts of asset price rises, net of references to forecasts of asset price drops	-1 to 1
sentiment	overall positive references, net of negative references	-1 to 1

stress	arousal and intensity, weighted towards distress	0 to 1*
surprise	unexpected events and surprise	1 to 1*
timeUrgency	urgency and timeliness, net of references to tardiness and delays	-1 to 1
trust	trustworthiness, net of references connoting corruption	-1 to 1
uncertainty	uncertainty and confusion	0 to 1*
violence	violent crime, terrorism, and war	0 to 1*
volatility	volatility in market prices or business conditions	0 to 1*

Table 4. Countries sentiment indicators - US

Index	Description: references in news and social media to...	Range
bondBuzz	sum of all references to the country's bonds and debt (excluding corporate debt) in that country	NA
bondDefault	debt defaults, late payments, and bankruptcy	-1 to 1
bondFear	fear and anxiety	0 to 1*
bondOptimism	optimism, net of references to pessimism	-1 to 1
bondPriceDirection	bond price increase, net of references to price decrease	-1 to 1
bondPriceForecast	forecasts of bond price rises, net of references to forecasts of asset price drops	-1 to 1
bondSentiment	overall positive references, net of negative references	-1 to 1
bondStress	arousal and intensity, weighted towards distress	0 to 1*
bondSurprise	unexpected events and surprise	0 to 1*
bondTrust	trustworthiness, net of references connoting mistrust	-1 to 1
bondUncertainty	uncertainty and confusion	0 to 1*
bondVolatility	volatility in bond and debt values	0 to 1*
centralBank	country central bank references	0 to 1*
debtDefault	debt defaults and bankruptcies in a country	0 to 1*
interestRates	interest rates rising, net of references to rates falling	-1 to 1
interestRatesForecast	forecasts of interest rates rising, net of forecasts of rates falling	-1 to 1
monetaryPolicyLooseVsTight	monetary policy being loose, net of references to monetary policy being tight	-1 to 1

ratesBuzz	sum of all references underlying the centralBank, debtDefault, interestRates, interestRatesForecast, and monetaryPolicyLooseVsTight TRMI	NA
stockIndexBuzz	sum of all relevant references feeding into the TRMI	NA
stockIndexFear	fear and anxiety	0 to 1*
stockIndexMarketRisk	positive emotionality and positive expectations net of negative emotionality and negative expectations. Includes factors from social media found characteristic of speculative bubbles – higher values indicate greater bubble risk. Also known as the “Bubbleometer.”	-1 to 1
stockIndexOptimism	optimism, net of references to pessimism	-1 to 1
stockIndexPriceDirection	stock price increases, net of references to price decreases	-1 to 1
stockIndexPriceForecast	forecasts of stock price rises, net of references to forecasts of asset price drops	-1 to 1
stockIndexSentiment	overall positive references, net of negative references	-1 to 1
stockIndexStress	arousal and intensity, weighted towards distress	0 to 1*
stockIndexSurprise	unexpected events and surprise	0 to 1*
stockIndexTrust	trustworthiness, net of references connoting mistrust	-1 to 1
stockIndexUncertainty	uncertainty and confusion	0 to 1*
stockIndexVolatility	volatility in stock market prices	0 to 1*

Table 5. Currencies sentiment indicators - USD

Index	Description: references in news and social media to...	Range
anger	anger and disgust	0 to 1*
buzz	buzz	NA
carryTrade	carry trade	0 to 1*
conflict	disagreement and swearing net of agreement and conciliation	-1 to 1
currencyPegInstability	the instability of a currency peg, net of references to the stability of a currency peg	-1 to 1
emotionVsFact	all emotional sentiments, net of all factual and topical references	-1 to 1
fear	fear and anxiety	0 to 1*
gloom	gloom and negative future outlook	0 to 1*

joy	happiness and affection	0 to 1*
longShort	buying, net of references to shorting or selling	-1 to 1
longShortForecast	forecasts of buying, net of references to forecasts of shorting or selling	-1 to 1
loveHate	love, net of references to hate	-1 to 1
marketRisk	positive emotionality and positive expectations net of negative emotionality and negative expectations. Includes factors from social media found characteristic of speculative bubbles – higher values indicate greater bubble risk. Also known as the “Bubbleometer.”	-1 to 1
optimism	optimism, net of references to pessimism	-1 to 1
priceDirection	price increases, net of references to price decreases	-1 to 1
priceForecast	forecasts of asset price rises, net of references to forecasts of asset price drops	-1 to 1
priceMomentum	currency price trend strength, net of references to trend	-1 to 1
sentiment	overall positive references, net of negative references	-1 to 1
stress	arousal and intensity, weighted towards distress	0 to 1*
surprise	unexpected events and surprise	1 to 1*
timeUrgency	urgency and timeliness, net of references to tardiness and delays	-1 to 1
trust	trustworthiness, net of references connoting corruption	-1 to 1
uncertainty	uncertainty and confusion	0 to 1*
violence	violent crime, terrorism, and war	0 to 1*
volatility	volatility in market prices or business conditions	0 to 1*

Appendix B.

Technical indicators

Technical indicators regarding momentum and trend have been included. The period for calculation is 10 business days, and the rest of parameters are general settings. TA Python library [20] was used. The following descriptions are also extracted from this library.

• **Relative Strength Index** (RSI): compares the magnitude of recent gains and losses over a specified time period to measure speed and change of price movements of a security. It is primarily used to attempt to identify overbought[45] or oversold[46] conditions in the trading of an asset. The former condition is generally interpreted as a sign that the stock is overvalued and the price is likely to go down. The latter is a result caused due to panic selling. RSI ranges from 0 to 100 and generally, when RSI is above 70, it may indicate that the stock is overbought and when RSI is below 30, it may indicate the stock is oversold. [29]

• **Rate Of Change** (ROC): same concept as simple return, but computed here with a period of 10 business days.

• **Stochastic Oscillator**: presents the location of the closing price of a security in relation to the high and low range of the price of a security over a period of time. In other words, it follows the speed or the momentum of the price. As a rule, momentum changes before the price changes.

• **Williams %R**: reflects the level of the close relative to the highest high for the look-back period. In contrast, the Stochastic Oscillator reflects the level of the close relative to the lowest low.

[45] A stock is said to be overbought when the demand unjustifiably pushes the price upwards.

[46] A stock is said to be oversold when the price goes down sharply to a level below its true value.

• **Moving Average Convergence Divergence** (MACD): is a trend-following momentum indicator that shows the relationship between two moving averages of prices.

• **Average Directional Movement Index** (ADX): determines both the direction and strength of a trend. The Plus Directional Indicator (+DI) and Minus Directional Indicator (-DI) are derived from smoothed averages of these differences, and measure trend direction over time. These two indicators are often referred to collectively as the Directional Movement Indicator (DMI). The Average Directional Index (ADX) is in turn derived from the smoothed averages of the difference between +DI and -DI, and measures the strength of the trend (regardless of direction) over time.

References

[1] López de Prado, M. (2020). Machine Learning for Asset Managers (Elements in Quantitative Finance). Cambridge: Cambridge University Press, 1-22. DOI:10.1017/9781108883658

[2] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25, 383–417.

[3] Shah, D.; Isah, H.; Zulkernine, F. (2019) Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. Int. J. Financial Stud., 7, 26.

[4] Moreno Mendienta, M. (16 Nov 2019). El fondo que convierte 100 dólares en 400 millones en 30 años. Cinco Días. El País Economía. https://cincodias.elpais.com/cincodias/2019/11/15/mercados/1573831472_956898.html

[5] Arévalo, Rubén, Jorge García, Francisco Guijarro, and Alfred Peris. (2017). A dynamic trading rule based on filtered flag pattern recognition for stock market price forecasting. Expert Systems with Applications 81: 177–92.

[6] Dijkstra, Margot & Borovkova, Svetlana. (2018). Deep Learning Prediction of the EUROSTOXX 50 with News Sentiment. 10.13140/RG.2.2.12318.23367.

[7] López de Prado, M. (2018). Advances in Financial Machine Learning. Wiley. 1-393 ISBN: 978-1-119-48208-6

[8] Y. Feng and D. P. Palomar. (2015) A Signal Processing Perspective on Financial Engineering. Foundations and Trends in Signal Processing, vol. 9, no. 1-2, pp. 1–231.

[9] Z. Bodie, A. Kane, and A. J. Marcus. (2013) Investments. Tata McGraw-Hill Education, 10th edition.

[10] J. C. Hull. (2014) Options, Futures, and Other Derivatives. Pearson Education India, 9th edition.

[11] Pandas Docs, User Guide: Computational tools, Exponentially weighted windows: https://pandas.pydata.org/pandas-docs/stable/user_guide/computation.html#stats-moments-exponentially-weighted

[12] Quantivity, Uncommon Returns through Quantitative and Algorithmic Trading (11 Feb 2011) Why Log Returns. https://quantivity.wordpress.com/2011/02/21/why-log-returns/

[13] Markowitz, H.M. (1952). "Portfolio Selection". The Journal of Finance. 7 (1): 77–91. DOI:10.2307/2975974. JSTOR 2975974.

[14] Atkins A., Niranjan M., Gerding E. (2018) Financial news predicts stock market volatility better than close price. The Journal of Finance and Data Science, vol 4, issue 2

[15] Jiao P, Veiga A., Walther A. (2016) Social media, news media and the stock market. Department of Economics, Discussion Paper Series. ISSN 1471-0498

[16] U.S. Securities and Exchange Comission. An official website of the United States Government. Introduction to Investment: Glossary. https://www.investor.gov/introduction-investing/investing-basics/glossary/

[17] Rubin, D. B. (1976). Inference and missing data. Biometrika 63(3): 581-592.

[18] Sabzevari M. (2019) Ensemble learning in the presence of noise. Ph.D. Thesis. Universidad Autónoma de Madrid. 1-25

[19] Lever, J., Krzywinski, M. & Altman, N. Principal component analysis. Nat Methods 14, 641–642 (2017). https://doi.org/10.1038/nmeth.4346

[20] TA Docs. Technical Analysis Library in Python. https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html

[21] Cavallaro, A. (23 Aug 2018) Introduction to "Advances in Financial Machine Learning" by López de Prado. Quantopian.

[22] Kuhn, M., Johnson, K. (2019) Feature Engineering and Selection: A Practical Approach for Predictive Models. https://bookdown.org/max/FES/

[23] Mlxtend User Guide. Sequential Feature Selector. http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/

[24] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[25] Friedman, Jerome H. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29 (2001), no. 5, 1189--1232. doi:10.1214/aos/1013203451. https://projecteuclid.org/euclid.aos/1013203451

[26] Chen, T., Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. DOI: 10.1145/2939672.2939785

[27] XGBoost Docs. Introduction to Boosted Trees. https://xgboost.readthedocs.io/en/latest/tutorials/model.html

[28] Joubert, J., Singh, A. (30 Apr 2019) Does Meta-Labeling add to signal efficacy? http://hudsonthames.org/does-meta-labeling-add-to-signal-efficacy-triple-barrier-method/

[29] Khaidem, L. et al. (2016) Predicting the direction of stock market prices using random forest. To appear in Applied Mathematical Finance. https://arxiv.org/abs/1605.00003