Expert Systems With Applications 184 (2021) 115497

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A machine learning algorithm for stock picking built on information based outliers

Emilio Barucci [a], Michele Bonollo [a,d], Federico Poli [b], Edit Rroji [c,*,1]

[a] Department of Mathematics, Politecnico di Milano, Italy
[b] Independent Researcher, Italy
[c] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy
[d] Lason Italy SRL, Italy

ARTICLE INFO

Keywords:
Finance
Stock picking
Technical analysis
Private information
Classification algorithm

ABSTRACT

Stock picking based on regularities in time series is one of the most studied topics in the financial industry. Various machine learning techniques have been employed for this task. We build a trading strategy algorithm that receives as input indicators defined through outliers in the time series of stocks (return, volume, volatility, bid-ask spread). The procedure identifies the most relevant subset of indicators for the prediction of stock returns by combining an heuristic search strategy, guided from the Information Gain Criterium, with the Naive-Bayes classification algorithm. We apply the methodology to two sets of stocks belonging respectively to the EUROSTOXX50 and the DOW JONES index. Performance is smoother than in the Buy&Hold strategy and yields a higher risk-adjusted return, in particular in a turbulent period. However, outperformance vanishes when transaction costs (5–15 basis points) are inserted. Asset return and return/volume serial correlation turn out to be the most relevant indicators to build the trading algorithm.

1. Introduction

In this paper we build a stock picking/trading algorithm based on indicators derived from time series of stocks (price, volume, bid-ask spread, min–max price). The analysis of time series is based on regularities that should be associated with the dissemination of information in financial markets. Financial markets theory provides several insights on the effects of private-asymmetric information in financial markets but the empirical evidence is limited. In what follows, we adopt an agnostic approach as we start from a wide set of indicators of the dissemination of information in financial markets and we consider a machine learning technique (wrapping procedure composed of an heuristic searching strategy and a classification algorithm) to select indicators to build the stock picking/trading algorithm. The capability of an indicator to reflect private information is evaluated through its capability to predict price movements in the short run and to build a successful trading algorithm.

As input, the algorithm receives indicators (dummy variables 0 −1) that signal the dissemination of private-asymmetric information in financial markets in agreement with the contributions of the financial markets theory. The indicators are identified as outliers of the financial time series with a probability threshold defined on the empirical distribution or with respect to a structural model. The indicators are based on the following time series: return, trading volume growth, bid-ask spread and volatility, serial correlation of return and trading volume. A subset of these indicators is used to identify a market signal for each security (BUY, NEUTRAL or SELL) to be confronted to stock return through a classification algorithm.

We have two main goals in our analysis: to test the capability of these market indicators to build a successful trading algorithm and to shed some light on the relevance of each indicator in predicting price movements. In this perspective, a machine learning algorithm provides a very useful tool as it allows to consider a wide set of indicators.

The paper is related to several strands of literature. First of all, it contributes to literature on asymmetric information in financial markets/insider trading, see Campbell, Grossman, and Wang (1993), Conrad, Hameed, and Niden (1994), He and Wang (1995), Llorente, Michaely, Saar, and Wang (2001), Wang (1994) and Biais, Bossaerts, and Spatt (2010) for theoretical analysis and Cornell and Sirri (1992), Meulbroek (1992) and Biais et al. (2010) for empirical analysis, see also

* Corresponding author.
E-mail addresses: barucci@polimi.it (E. Barucci), michele.bonollo@polimi.it (M. Bonollo), f.poli3@outlook.com (F. Poli), edit.rroji@unimib.it (E. Rroji).
[1] ORCID: 0000-0002-8663-2118

https://doi.org/10.1016/j.eswa.2021.115497
Received 10 February 2021; Received in revised form 8 May 2021; Accepted 25 June 2021
Available online 5 July 2021

Jeng, Metrick, and Zeckhauser (2003), Seyhun (1986) and Seyhun (1992) for trading strategies based on insiders' trades. As far as we know, this is the first paper that exploits a full collection of indicators on the potential dissemination of private information in financial markets to build a trading strategy. The peculiarity of our approach is that the selection of the indicators is done through an iterative machine learning algorithm without choosing a priori a time series anomaly to identify private information and a trading signal.

The paper is also related to the literature on stock picking exploiting time series regularities, in particular to the papers exploiting short memory trends, i.e., momentum strategies, e.g., see Jegadeesh and Titman (1993), Rachev, Jasic, Stoyanov, and Fabozzi (2007) and Taylor (2014), and to papers that evaluate the performance of technical analysis strategies, e.g., see Bajgrowicz and Scaillet (2012), Brock, Lakonishok, and LeBaron (1992), Fang, Jacobsen, and Qin (2014), Hsu, Hsu, and Kuan (2010), Kim, Lim, and Shamsuddin (2011), Lin (2018), Lo, Mamaysky, and Wang (2000), Neftci (1991), Sullivan, Timmermann, and White (1999), Arévalo, García, Guijarro, and Peris (2017), Cervelló-Royo, Guijarro, and Michniuk (2015), Muncharaz (2020a), Muncharaz (2020b) and Bianchi, Mercuri, and Rroji (2021). As far as the momentum strategy is concerned, we provide a richer analysis extending the set of time series that are used to build the trading strategy and we fully endogenize the choice of the signals used to build the trading strategy. As far as the technical analysis literature is concerned, we concentrate on stock picking considering a large set of stocks, while the above papers mostly concentrate on trading a stock index or a limited number of asset classes. Differently from the above papers, trading signals are built through outliers in financial time series that are associated with the dissemination of private information about the fundamental value. This feature represents a significant novelty with respect to the technical analysis literature providing a theoretical motivation to our algorithm that is lacking in many technical analysis models. Finally, our paper refers to recent machine learning applications to forecasting financial markets and portfolio selection (classification algorithms, genetic algorithms, neural networks, deep learning, support vector machines), see Allen and Karjalainen (1999), Ballings, Van den Poel, Hespeels, and Gryp (2015), Chen and Wang (2015), Chong, Han, and Park (2017), Gerlein, McGinnity, Belatreche, and Coleman (2016), Hsu et al. (2010), Hu, Feng, Zhang, Ngai, and Liu (2015), Huang (2012), Kaucic (2010), Lee (2009), Neely, Weller, and Dittmar (1997), Paiva, Cardoso, Hanaoka, and Duarte (2019), Leung, Daouk, and Chen (2000), Cervelló-Royo and Guijarro (2020) and Fisher and Krauss (2018).

The stock picking/trading algorithm works as follows. We start with a large set of predictors that are supposed to be useful to define the trading signal. A predictor is a market indicator computed on a specific moving window and for a specific confidence level. We rank the predictors using the Information Gain Criterium, see Abellán and Castellano (2017) and Kononenko (1994) for details. Then we build a hybrid approach (forward–backward) for the identification of most useful predictors for the stock return one step ahead. As we have a multiclass response variable, the procedure uses the Naive-Bayes classification algorithm that learns to predict the decision strategy (BUY, NEUTRAL or SELL) once observed the set of predictors. The Naive-Bayes classification algorithm is widely used in finance applications. For example, Holopainen and Sarlin (2017) show that the Naive-Bayes classifier results to be in the set of best performing algorithms for early-warning tasks. In Jadhav, He, and Jenkins (2018) the authors developed an algorithm, based on Information Gain and wrapper techniques for feature selection, using three different classification approaches for a credit scoring problem and get satisfactory predictive accuracy results in most of the countries under scrutiny when they adopt the Naive-Bayes classifier.

The trading algorithm is identified evaluating the prediction accuracy of all trained classification models in the validation set. A model is made up by a subset of indicators (defined according to confidence level and estimation window), thresholds in the response variable for the identification of the market signal and trained probabilities in the Naive-Bayes classifier. The best model represents the trading model generating the trading signal for an out of sample data point.

We apply the proposed methodology to two sets of stocks belonging to the EUROSTOXX50 and the DOW JONES index. We evaluate the performance considering a period characterized by a bull market and a period including the crisis associated with the COVID-19 pandemic. We compare the performance of the trading strategy generated by the trading algorithm to that of the Buy&Hold strategy and to a classical moving average trading rule.

Considering the Sharpe ratio as a performance metric, we observe that results are mixed. The trading strategy outperforms the moving average trading rule by far. Assuming no transaction costs, the trading strategy outperforms the Buy&Hold strategy in all the four out of sample subsets. Including transaction costs (5–15 basis points) outperformance vanishes. This result is in line with the literature, e.g., see Neely, Weller, and Ulrich (2009), Sullivan et al. (1999) and Taylor (2014), showing that profitability of technical trading rules is weak in recent times because markets are becoming more efficient as traders are using them. This is the claim of the so called "adaptive market hypothesis", see Lo (2004). However, the results are more positive than the recent literature on the profitability of technical analysis, e.g., in Bajgrowicz and Scaillet (2012) profitability of technical analysis trading strategies is weak even with no transaction costs. This outcome is maybe due to the fact that our trading signals have not been already considered extensively in the literature and by practitioners. We observe that the trading algorithm performance is smoother than the Buy&Hold strategy and is poor in a bull market while it is good in a turbulent period. This result agrees with evidence provided in Kaucic (2010), Kim et al. (2011) and Taylor (2014) showing that technical trading rules are more resilient than the Buy&Hold strategy in turbulent periods.

Our analysis provides information about the capability of time series regularities to predict future movements of the market. We do confirm that the actual weekly risk adjusted return is the most significant predictor as momentum strategies suggest. Contrary to large part of the literature on private information suggesting that a large trading volume and a high bid-ask spread provide evidence of private information, we find that they play a limited role. The second indicator that plays an important role to predict future returns and to build a successful trading strategy is provided by the auto-correlation structure of the return-volume time series. We can conclude that an extra return coupled with a structural break in the volume/return correlation structure provide a signal that something is happening in the market.

The paper is organized as follows. In Section 2 we provide theoretical insights that drive the selection of the algorithm. In Section 3 we describe the market indicators employed in our analysis. In Section 4 we describe the trading algorithm. In Section 5 we provide an empirical analysis applying the methodology to two portfolios built using stocks belonging to the EUROSTOXX50 and the DOW JONES index.

2. Literature insights

The design of the algorithm comes from the private information/insider trading literature which identifies a series of regularities of financial time series that are associated with trading activity due to private/asymmetric information.

We refer to two strands of literature: models with homogeneous information, models with heterogeneous-asymmetric information, see also Barucci, Bianchi, Casciari, and Squillantini (2006) for the literature discussion.

The literature on financial markets with homogeneous information has shown that under the risk neutral probability measure (assuming no arbitrage opportunities in the market) or under the historical probability measure with risk neutral agents, the discounted asset price is a martingale and therefore the market is a fair game: the conditional expected excess return (asset return minus the risk free return) is equal to zero and excess returns are serially uncorrelated. This framework

rationalizes the so called market efficiency hypothesis, see Fama (1970): according to the weak market efficient hypothesis, future excess returns cannot be predicted on the basis of past returns, e.g., they follow a random walk.

In presence of insider trading, return serial correlation is expected. As a matter of fact the insider trader tends to trade a limited amount in the direction of his information and therefore we expect positive (negative) daily or weekly returns to follow positive (negative) returns because private information is incorporated gradually in asset prices. A model that rationalizes this type of behavior of insider traders is provided by Kyle (1985).

Insight 1. In presence of private information dissemination, we observe positive serial correlation in security returns (trend).

The random walk hypothesis holds true in case agents are risk neutral. If agents are risk averse then asset demand depends on its riskiness. Financial markets theory has proposed several models that explain asset risk premia on the basis of no arbitrage/equilibrium arguments. The benchmark is provided by the Capital Asset Pricing Model (CAPM): if agents' preferences are represented by a quadratic utility function or the two mutual funds separation theorem holds true and markets are in equilibrium, then the asset risk premium is positively and linearly related to its beta. According to the CAPM we can establish the equilibrium risk premium of an asset and then we can detect anomalies with respect to it: we can take the market model derived from the CAPM as a benchmark to evaluate abnormal co-movements of the asset return with the market return.

Insight 2. In absence of private information dissemination, security returns should be in line with the CAPM: excess returns (asset return minus the risk free return) should not be different in a statistical sense from the value estimated by the market model.

Classical financial markets theory with homogeneous information is unable to explain the large trading volume observed in the markets. Trading volume in financial markets is due to two main motivations: risk sharing among agents and speculative trading. If information is homogeneous then the second motivation is absent and agents only trade to exploit Pareto improvements associated with differences in agents' risk expositions. In particular, if markets are complete, then trading is rather limited and occurs only in case of a preference/technology shock.

Under general assumptions, it can be shown that in a perfectly competitive market with heterogeneous private information (all agents observe a private signal on the asset value) and no noise (e.g., liquidity traders are absent) prices fully transmit private information, they instantaneously reveal private information and coincide with those of an economy where all private signals are public (prices are fully revealing), see Grossman (1989). If noise is added, then prices are not fully revealing and the trade size is increasing in the precision of information, on this point see for example Kim and Verrecchia (1991). Therefore, precise private information (insider trading) is associated with large trades, for a discussion on the relationship between trading size and information content see Chakravarty (2001).

Insight 3. In absence of private information dissemination, trading volume is limited compared to the free float, private information is associated with large trades.

Speculative trading, and therefore large trading volume, can originate from public or private information. In the first case we have a news for example on company profitability (e.g. investment decisions or mergers) and agents trade because they revise company's growth opportunities (time varying investment opportunities). If this is the case, then large trading volume is mainly concentrated around the announcement date and does not last for a long period. Instead, in case of private information we have that insiders trade until the asset price incorporates the new information (leakage of information), i.e., there is a public announcement or other agents detect private information. Notice that a large trading volume in a day with no serial correlation can also be observed in case of trades by institutional investors for liquidity reasons with no information content. As a consequence, serial correlation of trading volume is an interesting indicator to discern between pure risk sharing/public information based trading and private information trading. A model that disentangles the type of information arriving in the market according to trading volume serial correlation is provided by He and Wang (1995).

Insight 4. When public information arrives on the market, trading volume is not serially correlated. Trading volume serial correlation is associated with the dissemination of private information.

The presence of heterogeneous information also affects the relation between trading volume and asset returns. If large trading volume is due to uninformative motives (liquidity/preference shocks), then market pressure lasts for a short period and it is likely that we observe price reversal or mean reversion, i.e., negative return-volume correlation, see Campbell et al. (1993), Conrad et al. (1994) and Rosu (2019); instead, if trading volume is due to private information then the relation can have a different sign, i.e., positive return-volume correlation, see Blume, Easley, and O'Hara (1994), Llorente et al. (2001), McGorty, Gwilym, and Thomas (2009) and Wang (1994).

Insight 5. In presence of private information dissemination, large trading volume is associated with a price trend (positive return-volume correlation) and high volatility, if trades are due to liquidity motives then negative return correlation is more likely.

In a dealer market, dealers defend theirselves from trading with informed traders by setting a large bid-ask spread. As a matter of fact, there are two strands of literature for the bid-ask spread: inventory and adverse selection models, see Foucault, Pagano, and Ailsa (2013) and O'Hara (1995). In adverse selection models, see Copeland and Galai (1983) and Glosten and Milgrom (1985), it turns out that the bid-ask spread is increasing in the degree of asymmetric information in the market. Notice that bid-ask spread is positively associated with volatility, see Goodhart and O'Hara (1997).

Insight 6. In presence of private information dissemination, the bid-ask spread and the volatility are high.

The above insights have been empirically tested through two different exercises: considering illegal insider transactions and transactions by directors of companies. There are few papers on illegal insider trades. The literature provides little evidence in favor of the above theoretical insights. Meulbroek (1992) and Barucci et al. (2006) showed that days with trades by insiders are characterized by large trading volume and high excess returns (in absolute value) with respect to the market model (CAPM). Similar results have been obtained by Cornell and Sirri (1992). Weak evidence on price movements associated with insider trades has been detected in Chakravarty and McConnell (1999). On trading by directors and illiquidity the evidence is mixed: Bettis, Cole, and Lemmon (2000), Cao, Field, and Hanka (2004), Cheng, Firth, Leung, and Rui (2006) and Chung and Charoenwong (1998) provide evidence that spread widens and market depth falls on insider trading days as compared to non-insider trading days; Cornell and Sirri (1992) and Collin-Dufresne and Fos (2015) provide no evidence.

3. Market indicators

We consider the following time series for each security on a weekly basis[2]:

- r_t: **weekly return** which is defined as the total return of the security $\frac{P_t+D_t}{P_{t-1}}-1$, where D_t is the dividend at time t (during the week) and P_t is the end of the week closure price. In case the dividend is null at t, then r_t is the standard weekly return $\frac{P_t-P_{t-1}}{P_{t-1}}$.

[2] In what follows, writing "at time t" we refer to the weekly observation according to the specification of Thomson Reuters: as far as trading volume is concerned, we refer to the cumulative trading volume during the week; price information (closure price, bid, ask, high a low) refers to the day of observation.

- v_t: **rate of growth of trading volume** of the security at time t. Let V_t be the adjusted turnover volume during week t, then $v_t = \frac{V_t}{V_{t-1}}$. The adjusted turnover volume accounts for capital events that might affect the volume turnover.
- BA_t: **bid-ask spread** of the security, the spread is computed as the difference between the average bid price and the average ask price observed during the last day of week t.
- $\frac{P_t^H}{P_t^L}$: **highest/lowest price** observed during the last day of week t, where P_t^H and P_t^L are the highest and the lowest price during the day.

We opt for weekly observations as a week frequency allows to smooth the noise of daily observations. From the weekly time series we build four binary indicators $idx_i \in \{0, 1\},\ i = 1, 2, 3, 5$, that represent outliers in the time series signaling the possibility of the presence of asymmetric/private information in the market. The fourth indicator renders three different values: $idx_4 \in \{-1, 0, 1\}$. The indicators are defined as follows:

1. **Excess trading volume**
   At time t, we reconstruct the historical distribution of the latest $N-1$ growth rates of trading volume $\{v_{t-j}\}_{j=1,\dots,N}$ and define $\underline{v}$ as the upper 1-c% quantile where c takes values in the interval [0.025, 0.2] with equally spaced values of length 0.025.[3] Then, $idx_1^c = 1$ if and only if $v_t > \underline{v}$, that is if the observed growth rate of trading volume at time t is higher than the 1-c% quantile of the historical distribution of the last N observed values $\{v_{t-j}\}_{j=1,\dots,N}$, otherwise $idx_1^c = 0$.
2. **Excess bid-ask spread**
   At time t, we reconstruct the historical distribution of the security's bid-ask spread $\{BA_{t-j}\}_{j=1,\dots,N}$ and define $\underline{BA}$ as the upper 1-c% quantile. Then, $idx_2^c = 1$if and only if $BA_t > \underline{BA}$, that is, the observed bid-ask spread at time t is higher than the 1-c% quantile of the historical distribution of the last N observed values $\{BA_{t-j}\}_{j=1,\dots,N}$, otherwise $idx_2^c = 0$.
3. **Excess volatility**
   At time t, we estimate a $GARCH(1,1)$ model for the volatility of the stock return using data up to time $t-1$. We opt for this model for the volatility as there is evidence showing that it provides a parsimonious representation of the volatility dynamics, e.g., see Andersen and Bollerslev (1998). Therefore, for any $j = 1,\dots,N$, we consider the following model:

   $$r_{t-j} = \sigma_{t-j} z_{t-j}$$

   where

   $$\sigma_{t-j}^2 = \alpha_0 + \alpha_1 \sigma_{t-j-1}^2 + \alpha_2 r_{t-j-1}^2, \quad j = 1, \dots, N.$$

   z_t is a sequence of identically and independently distributed random variables with zero mean and variance equal to 1. We use the estimated parameters at time t $(\widehat{\alpha}_{0t}, \widehat{\alpha}_{1t}, \widehat{\alpha}_{2t})$ to obtain a forecast of the volatility at time t $(\widehat{\sigma}_t^2)$. This value is compared to the realized range volatility estimator, see Parkinson (1980):

   $$s_t^2 = \frac{1}{4\log(2)}\left[\log\frac{P_t^H}{P_t^L}\right]^2 \tag{1}$$

   Then, $idx_3^A = 1$ if and only if $A\widehat{\sigma}_t^2 < s_t^2$, where $A \in [0.4, 1.6]$ with equally spaced values of length 0.1, otherwise $idx_3^A = 0$. Notice that we vary significantly the benchmark on the volatility allowing the algorithm to select the indicators from a large set of variables.
4. **Excess return**
   At time t we estimate the market model for the security. We regress the security return $\{r_{t-j}\}_{j=1,\dots,N}$ on the total return of the stock index to which the security belongs $\{r_{t-j}^*\}_{j=1,\dots,N}$:

   $$r_{t-j} = \beta_0 + \beta_1 r_{t-j}^* + z_{t-j}, \quad j = 1, \dots, N.$$

   We use the parameters estimated at time t $(\widehat{\beta}_{0t}, \widehat{\beta}_{1t})$ and the realized stock index return at t (r_t^*) to estimate the return of the security at time t $(\widehat{r}_t)$:

   $$\widehat{r}_t = \widehat{\beta}_0 + \widehat{\beta}_1 r_t^*,$$

   we compare it to the observed return r_t. Set $r_t - \widehat{r}_t$ the excess return, we set $idx_4^c = 1$ if the excess return is above the 1-c% quantile of the distribution, $idx_4^{(c)} = -1$ if the excess return is below the c% quantile, and $idx_4^c = 0$ otherwise.
5. **Autoregressive structure**
   At time t, we estimate a vector autoregressive model of the form

   $$Y_{t-j} = A_0 + A_1 Y_{t-j-1} + E_{t-j}, \quad j = 1, \dots, N,$$

   where $Y_{t-j} = [r_{t-j}, v_{t-j}]^\top, E_{t-j}$ is a sequence of independent and identically distributed vectors of zero mean random variables. We test the single element significance of the autoregressive matrix by testing the null hypotheses: $H_0^{i,j} : A_1^{ij} = 0$ for $i, j = 1, 2$. We only consider the first three coefficients of A_1 omitting the coefficient on the serial correlation of the growth rate of trading volume because a preliminary investigation of the data set showed that the hypothesis is violated too frequently. We set $idx_5^c = 1$ if at least two of the nulls $H_0^{i,j}$ are rejected at significance level 1-c%, otherwise $idx_5^c = 0$.

These indicators can be associated to the literature discussion presented in Section 2: the excess trading volume indicator builds on Insight 3 and 4; the excess bid-ask spread and the excess volatility indicator are motivated by Insight 6; the excess return indicator builds on Insight 1 and 2 (private information induces excess return in t and this is likely to be observed in $t+1$); the indicator on the autoregressive structure is motivated by Insight 1 and 5.[4]

4. The trading algorithm

In this Section we present our selection/trading algorithm. We address this task through two sections: in Section 4.1 we define our building blocks of the trading algorithm while in Section 4.2 we provide a description of the engine of the algorithm and its implementation.

4.1. Building blocks

Our analysis is based on the following ingredients: *predictor, response variable, trading signal, model, trading model, sample.*

1. Predictor

Predictors are indicators as defined in Section 3 computed for a time window N and a confidence level $1-c$%. The universe of the indicators $(idx_1^c, idx_2^c, idx_3^A, idx_4^c, idx_5^c)$ is built varying the size of the estimation window N which means that at time t only the last N observations (weeks) are used for the computation of the indicator. N belongs to the set $W_0 = \{3,4,5,6,8,10,12,26,38,52,78,104\}$. Varying the size of the

[3] We use the same values of c for the other indicators presented in this section.

[4] In Barucci et al. (2006), a procedure to identify market abuse episodes was proposed based on the following indicators: trading volume, excess returns, autocorrelation of returns, autocorrelation of trading volume, correlation between trading volume and one step ahead return.

window used to estimate the indicator, we can group indicators tracking short, medium or long-term effects: estimating the indicators over a short time window we have a reactive indicator, considering a long window we have a much more stable/smoother indicator. Some indicators cannot be computed for all $N \in W_0$ as we need a large sample to get convergence of the estimate. In particular, the excess volatility indicator is computed only for $N = 104$ and the indicator on the autoregressive structure for $N \geqslant 6$.

2. Response variable

The response variable is a categorical variable reflecting the direction of the asset price movement. As response variable associated to the predictors computed at time t, we consider y_{t+1} which is based on market return r_{t+1}. In particular, given a threshold parameter θ and $q_{f(\theta)}$ the associated quantile of the distribution of the asset return computed from the observations in the training set, we set:

$$y_{t+1,\theta} = \begin{cases} 1 & if\ r_{t+1} > q_{0.5+\theta} \\ -1 & if\ r_{t+1} < q_{0.5-\theta} \\ 0 & otherwise \end{cases} \tag{2}$$

Note that for $\theta = 0$ we are back to a simple binary response variable. As θ increases, the response variable, and therefore the trading algorithm, becomes more selective as there is an interval of returns centered on zero with a neutral signal. A positive θ captures the possibility that there is uncertainty in the model, i.e., the regions with a BUY signal and with a SELL signal are not clearly identified and therefore we set a non-decision region.

The universe of the response variables is obtained varying the quantile threshold θ in the interval $[0, 0.05]$ with a step length of 0.01 yielding six specifications. So we allow for an interval $[0, 0.1]$ with a neutral signal.

3. Trading signal

The procedure is based on a classification algorithm which receives the predictors computed at time t as inputs and yields a trading signal: a positive signal (BUY) in case the classification algorithm yields $+1$, a negative signal (SELL) in case the classification algorithm yields -1 and a neutral signal (NEUTRAL) in case the classification algorithm yields 0.

In the training/validation set the algorithm exploits the information contained in the predictors to match the response variable. Then, out of sample the algorithm is used to define a trading signal: a positive signal leads to buy one unit of the stock and to hold it for the next week or to maintain the stock in the portfolio if it already belongs to the portfolio. A negative signal leads to sell one unit of the stock if it is already in the portfolio and not to buy it otherwise. A neutral signal yields no trading maintaining the actual position. Notice that we do not allow for short sales. All the transactions are deployed borrowing or lending the surplus at the risk free rate (set equal to zero).

At time 0 we suppose to have a capital equal to the amount of money required to buy one unit of each stock in the set of eligible securities.

4. Model

A model is made up of the response variable y_θ, for a specific θ, and the set of predictors each one computed for a specific $N \in W_0$ and confidence level c/parameter A. Therefore, a model is identified by the parameters N, c, A, θ for each variable and is associated to the corresponding data set obtained from the original observations. The data set provides the input for the algorithm.

5. Trading model

The selection procedure described in the next Section renders the *trading model* at each t, i.e., the best combination of response variable and subset of predictors for the generation of the trading signal.

6. Sample

The sample of weekly observations of the primitive variables (return, trading volume, highest/lowest price, bid-ask price) allows us to identify the *out of sample set* as the set of the most recent observations to be defined in Section 5. For each t in the out of sample data set the trading algorithm is estimated in the dataset which contains all the observations up to t. The observations are divided in two subsamples: the *training set* containing 80% observations and the *validation set* containing the most recent 20% observations. As we move to $t + 1$, the training set and the validation set include observations at time t and then again the sample is divided in two data sets according to the above fraction.

4.2. Selection procedure

The selection procedure at the heart of the trading algorithm builds on several steps. The trading signal depends on a classification algorithm where for each security we need to match the response variable in sample and to derive the trading signal out of sample. To this end, we employ the Naive-Bayes classification algorithm described in Appendix A, but the procedure can be adapted to other classification methods that allow for multiclass response variables. The choice of Naive-Bayes is strictly linked to the aim of having a non computationally expensive classification method as we have many stocks and indicators.

At time t we have to address three different tasks:

1. for each θ (and response variable $y_{t,\theta}$) estimate the parameters of the Naive-Bayes classifier for subsets of predictors varying $N,\ c, A$. This task is performed on the training set;
2. for each θ, given the parameters of the classifier, the optimal subset of predictors is chosen evaluating the performance of models in the validation set in terms of prediction accuracy;
3. choose the trading model among the models (obtained for varying θ) calibrated through the first two steps.

Therefore, the Naive-Bayes algorithm is calibrated on the training set and the definition of the subset of predictors and the (final) choice of the trading model at time t are addressed through the analysis of the performance of the models on the validation set. To this end, we have to define an accuracy measure to evaluate the performance of the models on the validation set and the procedure adopted to select the predictors.

1. Accuracy measures

There exist several accuracy measures based on the confusion matrix. We use the Mathew's Correlation Coefficient (MCC) originally developed in Matthews (1975) and recently proposed as a performance metric in machine learning applications. The MCC is a method of calculating the Pearson product moment correlation coefficient between actual and predicted values, i.e. values predicted by the trading algorithm and those observed for the response variable.

Referring to the confusion matrix that contains the following information based on predicted values TN:=true negative, TP:=true positive, FP:=false positive and FN:=false negative, the MCC is defined as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}.$$

MCC ranges in the interval $[-1, +1]$, where the values -1 and $+1$ are obtained in case of perfect misclassification and perfect classification, respectively. We choose the MCC measure as it can be easily extended to the case of multiclass response variable and it is well-suited for unbalanced data sets (see Boonamnuay, Nittaya, & Kittisak (2018) and Jurman, Riccadonna, & Furlanello (2012) for details).

The trading signal (SELL, NEUTRAL or BUY) is derived by the response variable that takes values in the set $\{-1, 0, 1\}$. As we are interested in the accuracy of the classification algorithm and also in avoiding huge losses and in exploiting potential future large upward/downward movements, we introduce a metric for model comparison defined as Return Weighted Accuracy (RWA). It is strictly linked to the Accuracy measure which is defined as the proportion of correct predictions among the total number of cases considered in the binary classification problem, i.e. we have that:

$$Accuracy = \frac{TN + TP}{TN + FN + FP + TP}.$$

As we want to adopt a strategy that allows us to correctly identify potential large movements of stock returns, we modify the Accuracy taking into account the absolute return and define RWA as:

$$RWA = \frac{\sum_{t=1}^{T} r_t 1_{\{r_t>0\}} 1_{\{y_t=1\}} + \sum_{t=1}^{T} |r_t| 1_{\{r_t<0\}} 1_{\{y_t=-1\}}}{\sum_{t=1}^{T} |r_t| 1_{\{y_t=-1 \vee y_t=+1\}}}.$$

The nice feature of the RWA is that it provides a high score to a model that is able to correctly define the response variable in case of a large market movement.

Given a θ for defining the classes in the response variable, the algorithm selects the predictors in a sequential way. A predictor is included in the Naive-Bayes classifier estimated on the training set if it yields an improvement on the validation set considering $MCC+RWA$ as performance indicator.

A crucial point is the order that is followed to consider and select the predictors.

2. Forward–Backward algorithm based on the Information Gain ranking of predictors

The searching algorithm is based on an iterative switch between *sequential forward selection* for the inclusion of new variables (predictors) and *backward selection* for variable elimination.

For each security, given a set of $\overline{n}$ variables, the procedure starts with $2^{\overline{n}}$ possible models. As the dimension can be quite large, we perform a *pre-selection* of predictors. Considering the training set, we compute the correlation matrix of predictors. For each couple of predictors showing a correlation higher than 90% we eliminate one of them.

We follow an heuristic approach that allows us to select a limited number of variables to be included in the trading algorithm. In the sequential forward search algorithm, that is a *wrapper method*, we start with an empty set of variables (predictors in our setting) and we sequentially test the inclusion of a new variable. The inclusion or not of a variable depends on the score $MCC+RWA$ computed in the validation set for the Naive-Bayes classifier calibrated on the training set. If the score increases we include the predictor, otherwise we look for another predictor.

We thus have to define a sequential order for the introduction of a predictor in order to reduce computational costs. To this end we first rank the predictors using the Information Gain (IG) criterium which is widely used for high dimensional data set to measure the effectiveness of variables in a classification exercise, see Kononenko (1994).

IG is derived from the Shannon entropy. In information theory the entropy of a random variable Y is defined as

$$H\left(Y\right) = -\sum_{y\in Y} p\left(y\right) \log p\left(y\right)$$

where $p(y)$ is the probability of observing a realization y of Y. It is possible to compute the conditional entropy of Y given X as follows:

$$H\left(Y|X\right) = -\sum_{x\in X,\ y\in Y} p\left(x,y\right) log\left(\frac{p(x,y)}{p(x)}\right),$$

where $p(x,y)$ is the probability of observing a realization x of X and y of Y. $H(Y|X)$ quantifies the amount of information needed to describe the outcome of Y given that the value of X is known. Notice that $H(Y|X) = H(Y)$ if the two variables are independent, instead $H(Y|X) < H(Y)$ in case there is a relationship between the two variables.

The IG measures the change in information entropy (Y) from a prior state to a state that takes some information as given (X):

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y).$$

The predictors are ordered according to the IG from the most informative to the less informative.

In a forward search algorithm we can only add variables and never remove a variable included in the set for the classification exercise. This feature increases computational complexity as the size of the set of variables can only increase. Notice that following this approach, one or more variables in the model may become redundant once we add a new variable. To address this issue, we also include a *backward selection* step. In practice, when we include a new variable we also check whether the objective function ($MCC + RWA$) increases by excluding one of the variables already included in the set in the previous step (we repeat this procedure for each variable). If we do not observe an improvement in the objective function ($MCC + RWA$) on the validation set, then we proceed with the forward step by testing the inclusion of the new variable. If the removal of at least one of the variables provides an improvement in the objective function, we exclude it and repeat the backward selection step by looking for a more parsimonious model. The backward selection step stops when the inclusion of a variable does not provide anymore an improvement in the objective function. Subsequently, we proceed with the inclusion of a new variable in the forward step. The discarded variables in the backward step enter in the set of variables that can be selected in the next forward step where the procedure is still defined by the IG criterium. Indeed, a variable that is redundant in a given set may become relevant in a new set (with a different variable mix). The procedure stops when all the variables have been tested at least once in the forward selection step with no improvement of $MCC+RWA$ on the validation set.

We repeat the procedure described above for each response variable $y_{t,\theta}$ identifying the best model for each θ. Then we choose (step 3) the trading model ($\overline{\theta}$, the corresponding $y_{t,\overline{\theta}}$ and the predictors selected as described above thanks to the MCC and the RWA metric). This choice renders the θ and the model yielding the largest $MCC+RWA$ in the validation set.

The procedure described above is named Information Gain Forward–Backward Model Selection (IGFBMS) algorithm. and is represented in Fig. 1.

5. Application to the EUROSTOXX50 and DOW JONES index

Our application concerns weekly observations of 30 and 29 stocks belonging to the EUROSTOXX50 and to the DOW JONES index, see Tables 1 and 2 respectively. The market capitalization of stocks included in the dataset is about 67% in the EUROSTOXX50 index and 98% for the DOW JONES index. The data set covers the period 8/01/2004 −14/05/2020, the window 8/01/2004 −31/12/2018 is used for the training/validation set of the trading algorithm for the first observation out of sample (the data set is split according to the 80 −20% ratio). The remaining part of the data set is used to perform the out of sample analysis. As described in Section 4.1, moving from the first week out of sample to the second one, the sample on which the algorithm is calibrated is augmented by one observation and the sample is split according to the above ratio. Few stocks of the two indices are not included in the analysis because they belonged to the Index for a smaller time window.

In order to test the performance of our trading algorithm under different market conditions, we consider the full out of sample data set (1/1/2018 −19/5/2020) and a truncated data set (1/1/2018 −31/12/2019). The first data set includes a stable period and then a bull market, the second data set also includes the period characterized by the COVID-19 pandemic with an abrupt crash and then recovery.

Our trading algorithm combines market indicators obtained from some time series for each stock such as return, volume and bid-ask spread. As a preliminary analysis we evaluate the information contained in these series. We check the relevance of each source and its non redundancy. To this end in Tables 3 and 4 we compute the correlation between the three pairs of time series for the period 01/01/2017 – 19/05/2020. We observe that the correlation is very low with the only exception provided by the correlation between asset return and bid-ask

```
Input: V: Set of variables selected by correlation threshold
       Y: Set of 11 response variables built from returns using the vector θ
Output: Best model for each Y
Initialize RWA
Initialize MCC
Model=∅
for k=1:length(Y) do
   IG: Information Gain based Ranking of the set "V"
   Model(k): one variable Set // The first variable in IG
   NT= IG\ Model(k) //Initialize set of never tested features with IG based ranking
   NNF= IG\ Model(k) //Initialize set of not selected features with IG based ranking
   while NT≠ ∅ do
      for i=1:length(NNF) do
         // Forward step
         S= Model(k) ∪ NNF(i)
         if (NNF(i) ∈ NT) then
            NT=NT \ NNF(i) // Remove the variable NNF(i) from NT
            Train "Naïve-Bayes" classifier on S and Y(k)
            Produce prediction in the validation set and compute MCC(S) and RWA(S)
            if not (MCC(S)>MCC & RWA(S)>RWA) then
               next// Stop and pass to test the next variable: Forward step
            else:
               Model(k)=S; MCC=MCC(S); RWA=RWA(S); NNF= IG\ Model(k)
               //Backward step
               while length(Model(k))>1) do
                  for j=1: (length(Model(k)) - 1) do
                     H(j)=Model(k)\ Model(k, j) // Exclude the j-th variable in Model(k)
                     Train "Naïve-Bayes" classifier on H(j) and Y(k)
                     Produce prediction in the validation set and compute MCC(H(j)) and RWA(H(j))
                  end for
                  T= best(H(j)) // Best model based on the sum of MCC(H(j)) and RWA(H(j))
                  if not (MCC(T)>MCC & RWA(T)>RWA) then
                     break // Exit from the backward step
                  else:
                     Model(k)=T; MCC=MCC(T); RWA=RWA(T); NNF= IG\Model(k)
                  end if
               end while
               break //Exit from the FOR loop and restart it with new NNF
            end if
         end if
      end for
   end while
   Model=[Model; [Model(k) MCC RWA]]
end for
Return Model;// Best model based on the sum of MCC and RWA
```

Fig. 1. Pseudocode for the Information Gain Forward–Backward Model Selection (IGFBMS) algorithm.

spread which is high (above 30%) for four stocks belonging to the DOW JONES index. We also perform a Principal Component Analysis. In Table 5 we report the fraction of variance of the time series explained by the first k principal components. In emerges that several factors are necessary to explain the variance of these time series. The exception is provided by the bid-ask spread and (in part) by the return of securities belonging to the DOW JONES index, in these cases the first component explains more than 90% of the variance. In the other cases, more than 10 factors are needed to reach that threshold.

To save computational time, we select predictors monthly while the calibration of the classifier is performed weekly. Our methodology works as follows. We first run the method on the training/validation set for the first observation out of sample (8/01/2004 −31/12/2018). For each week of the data set we use the information provided by predictors to extract a trading signal which is matched to the response variable. The models/sets of predictors are selected training the Naive-Bayes classification algorithm on the training set and evaluating their performance on the validation set leading to the definition of the trading model for the first week out of sample. Then the parameters of the classifier are re-defined using all the information contained in the training/validation set and the trading model is employed to define a trading signal for the first week out of sample. As we move to the second week out of sample – and then to further observations in the data set – the selection of the predictors and of the trading model on the training set/validation set is performed every four weeks. The parameters of the selected trading

Table 1
Stocks included in the analysis of the EUROSTOXX50, main statistics are computed using weekly returns.

Stock	Weight	Mean	Std. dev	Skewness	Kurtosis
ANHEUSER-BUSCH INBEV	1.84%	-0.0062	0.0541	-0.7349	7.5343
KONINKLIJKE AHOLD DELHAIZE	1.19%	0.0024	0.0329	-1.5634	9.3932
ADIDAS	1.93%	0.0018	0.0499	-1.3612	11.6333
AIR LIQUIDE	2.72%	0.0023	0.0312	-3.3902	21.2102
ASML HOLDING	5.01%	0.0062	0.0506	0.3614	4.4741
AXA	1.55%	-0.0017	0.0542	0.2406	15.9672
BASF	2.05%	-0.0051	0.0404	-0.9292	6.3422
BAYER	2.56%	-0.0026	0.0509	-1.5217	5.1858
BMW	0.73%	-0.0041	0.0505	0.2099	6.5277
BNP PARIBAS	1.73%	-0.0056	0.0536	-0.9199	3.5837
CRH	0.82%	0.0009	0.0611	1.5034	18.5905
DAIMLER	1.04%	-0.0054	0.0626	0.4486	8.0578
DANONE	2.08%	6.8191E-06	0.0305	-0.3224	11.9893
DEUTSCHE TELEKOM	2.15%	0.0011	0.0307	-3.2363	22.7216
ENEL	2.50%	0.0031	0.0457	-4.6189	37.3469
ENI	1.01%	-0.0032	0.0551	-2.6117	29.1351
ESSILORLUXOTTICA	1.79%	0.0004	0.0385	-1.3366	6.7986
IBERDROLA	2.78%	0.0041	0.0367	-3.5055	26.3426
INTESA SANPAOLO	1.31%	-0.0045	0.0487	-1.7956	10.4551
LVMH	4.55%	0.0022	0.0465	0.1683	7.7953
ORANGE	1.25%	-0.0022	0.0340	-1.9399	17.9341
L'OREAL	3.01%	0.0025	0.0345	-0.9260	6.6713
BANCO SANTANDER	1.92%	-0.0077	0.0524	-0.8165	7.3403
SAP	5.44%	0.0025	0.0428	0.5993	4.2418
SANOFI	4.59%	0.0045	0.0317	-1.7632	9.7246
SCHNEIDER ELECTRIC	2.10%	0.0022	0.0482	-0.9941	11.5907
TELEFONICA	1.12%	-0.0044	0.0461	-0.7267	14.7854
VOLKSWAGEN	0.94%	-0.0006	0.0506	-0.0638	3.7187
ALLIANZ	3.08%	-0.0003	0.0476	0.1755	14.0075
SIEMENS	2.93%	-0.0005	0.0448	-0.0902	9.0762

Table 2
Stocks included in the analysis of the DOW JONES, main statistics are computed using weekly returns.

Stock	Weight	Mean	Std. dev	Skewness	Kurtosis
3 M	4.10%	-0.0020	0.0374	-0.6157	1.8742
AMERICAN EXPRESS	2.90%	0.0010	0.0328	-1.3955	5.7337
APPLE	6.46%	0.0064	0.0380	-0.1335	1.5649
BOEING	8.85%	-0.0028	0.0627	-3.2376	23.3027
CATERPILLAR	3.49%	-0.0010	0.0444	-0.1492	2.2808
CHEVRON	2.53%	-0.0008	0.0369	-0.5581	3.8797
CISCO SYSTEMS	1.09%	0.0025	0.0340	-0.4608	1.0765
COCA COLA	1.29%	0.0013	0.0303	-1.6111	10.0665
EXXON MOBIL	1.65%	-0.0027	0.0370	-0.9843	3.9654
GOLDMAN SACHS GP.	5.35%	-0.0010	0.0370	-0.2659	1.8583
HOME DEPOT	5.33%	0.0029	0.0341	-2.3288	15.4366
INTEL	1.40%	0.0039	0.0411	-0.4232	1.0386
INTERNATIONAL BUS. MCHS.	3.25%	0.0000	0.0371	-0.7384	2.6631
JP MORGAN CHASE & CO.	3.32%	0.0003	0.0334	-0.6655	2.1219
JOHNSON & JOHNSON	3.19%	0.0013	0.0283	-1.0686	5.5331
MCDONALDS	4.70%	0.0018	0.0339	-3.5954	29.3686
MERCK & COMPANY	2.11%	0.0033	0.0295	0.0329	2.3172
MICROSOFT	3.66%	0.0070	0.0284	-0.4342	1.9535
NIKE 'B'	2.26%	0.0039	0.0364	-1.4203	10.0375
PFIZER	0.93%	0.0012	0.0294	-0.2403	1.8999
PROCTER & GAMBLE	2.95%	0.0025	0.0256	-0.6316	5.7054
RAYTHEON TECHNOLOGIES	3.59%	0.0007	0.0398	-1.2525	8.2337
TRAVELERS COS.	3.31%	-0.0012	0.0327	-1.5021	10.1541
UNITEDHEALTH GROUP	6.70%	0.0036	0.0395	-0.4904	2.3060
VERIZON COMMUNICATIONS	1.46%	0.0013	0.0252	-0.1837	1.0156
VISA 'A'	4.46%	0.0050	0.0303	-1.3132	4.4255
WALGREENS BOOTS ALLIANCE	1.44%	-0.0023	0.0390	-0.2318	1.0654
WALMART	2.87%	0.0023	0.0241	-0.0789	2.1764
WALT DISNEY	3.66%	0.0017	0.0334	-0.6310	4.9514

model are calibrated using all the information contained in training/validation set. This procedure is implemented for each stock.[5]

In Table 6 we present the main features of the trading algorithm for the stocks belonging to the two indices. For the analysis performed on the stocks of the first index (EUROSTOXX50) we have 373 possible predictors varying $N,\ c,\ A$ as pointed out above: $N = 12$ and $c = 8$ for $idx_i,\ i = 1, 2, 4, N = 9$ and $c = 8$ for idx_5 and $N = 1, A = 13$ for idx_3. As we repeat the procedure of variable selection 31 times (every four weeks), the total number of models is $930 = 31 \times 30$ and the total number of variables is 11301. On average, for each model, we select 12 predictors while the average value for $\overline{\theta}$ is 0.020. For the second set of stocks (DOW JONES) the average number of selected predictors is lower while the average value for the selected $\overline{\theta}$ does not change.

In Table 7 we further investigate the distribution on the number of predictors employed by the algorithm and the extension of the neutral signal.

We notice that in 75% of the cases, the trading model selects between eight and sixteen predictors in both data sets. This result suggests that market indicators, set with different parameters (length of the estimation period, confidence parameter) provide useful information in defining the trading signal. We can conclude that the time series considered in our analysis are not redundant and that in most of the cases the same market indicator enters the model with different lengths of the estimation period and/or confidence level. This result confirms that time series contain complementary/non redundant pieces of information and that the same time series enters the trading model with different parameters defining the confidence level and the estimation window as in trading strategies based on short and long moving averages.

We notice that the model seems to be well designed. Allowing for a degree of uncertainty in the trading signal with up to a 10% of neutral signals in the response variable (weekly return) distribution, we end up with a 4% interval on average and actually in 1/4 of the cases (stock-observation) we end up with a sharp trading signal in both stock indexes ($\theta = 0$).

In Tables 8 and 9 we report the performance measures of the trading strategy provided by the trading model. The performance is computed on the full out of sample data set and on the subset terminating by the end of 2019. We consider the trading strategy defined by the trading model selected as above and we compare it to the performance of the Buy&Hold strategy, where we assume to buy at time $t = 0$ one unit of each of the 30 stocks of the EUROSTOXX50 (29 for the DOW JONES index). As recent literature has shown that the performance of technical rules is likely to vanish if transaction costs are considered, we evaluate the performance including 0, 5, 10, 15, 20 basis points as transaction costs. In the penultimate column (Break even) we also report the level of

[5] The procedure described in the paper can be extended by using other classifiers that deal with response variables with more than two classes. As a robustness check, we repeated the analysis using the Multinomial Logistic regression for the classification problem. It is important to highlight the fact that the algorithm based on the Multinomial Logistic regression is more time consuming compared to that based on the Naive-Bayes classifier. Results are similar. For example, if we consider the period end of May 2019-end of May 2020 we observe that for the EUROSTOXX 50 index the algorithm that uses the Naive Bayes classification approach performs better (performance of the Naive-Bayes −3.09% versus −8.29% for the Multinomial Logistic regression) but the opposite holds true for the DOW JONES index (performance of the Naive-Bayes −1.79% versus −0.61% for the Multinomial Logistic regression). However, we can observe that the resulting portfolios have similar dynamics suggesting that the procedure is robust but the choice of the classification algorithm may provide some differences in the performance of the strategy. Results are available upon request.

Table 3
Correlation coefficients for the stocks of the EUROSTOXX50 index computed on return, bid-ask spread and growth rate of volume for the period 01/01/2017–18/05/2020.

	$\rho_{ret,Bid-ask}$	$\rho_{ret,vol}$	$\rho_{Bid-ask,vol}$
ANHEUSER-BUSCH INBEV	−0.045	0.183	−0.012
KONINKLIJKE AHOLD DELHAIZE	−0.003	0.123	−0.017
ADIDAS	0.067	−0.115	0.079
AIR LIQUIDE	0.046	0.143	−0.011
ASML HOLDING	−0.187	0.059	0.051
AXA	0.029	0.151	−0.032
BASF	0.115	0.033	0.051
BAYER	0.176	0.256	0.005
BMW	−0.064	0.084	−0.060
BNP PARIBAS	0.055	−0.058	0.101
CRH	0.286	0.003	0.062
DAIMLER	−0.146	0.049	−0.040
DANONE	−0.025	0.104	−0.042
DEUTSCHE TELEKOM	0.257	−0.026	−0.005
ENEL	0.016	0.113	−0.045
ENI	0.004	0.102	−0.022
ESSILORLUXOTTICA	−0.076	0.013	0.035
IBERDROLA	−0.142	0.116	0.027
INTESA SANPAOLO	−0.142	−0.158	−0.011
LVMH	−0.019	0.095	−0.060
ORANGE	0.068	0.040	−0.130
L'OREAL	−0.027	0.159	−0.031
BANCO SANTANDER	−0.139	0.011	−0.009
SAP	0.178	0.014	0.022
SANOFI	−0.022	−0.005	−0.036
SCHNEIDER ELECTRIC	0.074	0.045	−0.013
TELEFONICA	−0.050	0.038	0.007
VOLKSWAGEN	0.152	−0.109	0.036
ALLIANZ	−0.019	0.042	0.035
SIEMENS	0.090	−0.040	−0.006

Table 4
Correlation coefficients for the stocks of the DOW JONES index computed on return, bid-ask spread and growth rate of volume for the period 01/01/2017–18/05/2020.

	$\rho_{ret,Bid-ask}$	$\rho_{ret,vol}$	$\rho_{Bid-ask,vol}$
3M	−0.224	−0.018	0.018
AMERICAN EXPRESS	0.244	0.009	0.032
APPLE	0.376	0.008	0.082
BOEING	0.254	−0.027	−0.098
CATERPILLAR	−0.072	0.008	−0.060
CHEVRON	−0.006	−0.016	−0.059
CISCO SYSTEMS	−0.149	−0.019	−0.133
COCA COLA	−0.020	−0.003	0.036
EXXON MOBIL	−0.497	0.002	0.071
GOLDMAN SACHS GP.	0.179	0.009	−0.005
HOME DEPOT	0.306	0.015	−0.011
INTEL	0.003	−0.027	−0.071
INTERNATIONAL BUS.MCHS.	−0.019	0.052	0.092
JP MORGAN CHASE & CO.	0.248	−0.001	0.154
JOHNSON & JOHNSON	0.194	0.006	−0.050
MCDONALDS	−0.048	−0.007	0.113
MERCK & COMPANY	−0.145	−0.015	0.068
MICROSOFT	0.250	0.001	0.046
NIKE 'B'	−0.067	0.001	0.036
PFIZER	0.040	−0.005	−0.025
PROCTER & GAMBLE	0.366	−0.007	−0.029
RAYTHEON TECHNOLOGIES	0.176	−0.002	0.153
TRAVELERS COS.	−0.021	−0.027	−0.050
UNITEDHEALTH GROUP	0.252	−0.018	0.097
VERIZON COMMUNICATIONS	−0.124	−0.027	0.015
VISA	0.268	0.027	0.009
WALGREENS BOOTS ALLIANCE	0.021	−0.023	0.022
WALMART	0.538	0.012	0.071
WALT DISNEY	0.209	−0.015	0.037

Table 5
Fraction of variance explained by the first k principal components for three series (return, growth rate of volume and bid-ask spread) for the period 01/01/2017–19/05/2020 for the stocks of the EUROSTOXX50 (left) and of the DOW JONES (right) indices.

	EUROSTOXX 50			DOW JONES		
k	Returns	Volume	Bid-ask spread	Returns	Volume	Bid-ask spread
1	57.14	39.51	55.86	91.80	50.22	96.03
2	11.32	10.54	31.39	3.85	5.29	3.66
3	5.79	7.78	5.81	2.09	4.42	0.09
4	5.58	7.00	2.25	0.872	4.08	0.05
5	3.18	6.10	2.06	0.61	3.59	0.04
6	2.77	5.10	0.56	0.37	3.34	0.03
7	2.53	3.73	0.41	0.23	2.72	0.02
8	2.28	3.06	0.39	0.09	2.54	0.02
9	1.87	2.89	0.27	0.03	2.46	0.01
10	1.64	2.48	0.27	0.03	1.99	0.01

Table 6
Trading model features.

Description of the Database	EUROSTOXX50	DOW JONES
Number of Predictors	373	373
Number of Stocks	30	29
Periods of selection	31	31
Number of Models	930	899
Number of Variables	11301	9251
Average number of predictors for each model	12.15	10.29
Average $\bar{\theta}$	0.020	0.021

Table 7
Number of predictors employed by the algorithm and selected θ.

	EUROSTOXX50	DOW JONES
Number of Predictors (NrP)		
$NrP < 4$	0.9%	0.1%
$4 \leqslant NrP < 8$	17.5%	8.5%
$8 \leqslant NrP < 12$	47.8%	34.5%
$12 \leqslant NrP < 16$	28.5%	40.2%
$16 \leqslant NrP < 20$	4.6%	14.5%
$20 \leqslant NrP$	0.8%	2.2%
$\bar{\theta}$		
0%	26.7%	24.5%
1%	19.6%	21.5%
2%	13.8%	19.%
3%	12.3%	11.8%
4%	13.0%	10.6%
5%	14.6%	12.%

transaction costs that renders the performance of the trading strategy (evaluated according to the Sharpe ratio) equivalent to the performance of the Buy&Hold strategy. In Figs. 2 and 3 we report the performance of the trading strategy for the two applications.

Notice that the trading strategy renders a Sharpe ratio higher than the Buy&Hold strategy in all the four data sets. In three out of four cases, the historical return is lower than the Buy&Hold strategy but also the standard deviation is smaller. The trading strategy is less volatile and less risky. Considering the shortest data set (the one with a bull market that excludes the COVID crisis) the performance is slightly better and vanishes when transaction costs accounting for five/ten basis points are included. Instead, when also the COVID crisis is included in the data set, the performance is significantly better than that of the Buy&Hold trading strategy and transaction costs accounting for ten/fifteen basis points should be inserted to allow for the Buy&Hold strategy to

Table 8
Statistics and performance measures of the trading strategy for the EUROSTOXX50 set of stocks computed using returns on an annual basis. The risk-free rate is $r_f = 0$ while break even refers to the transaction cost that implies the same Sharpe ratio for the trading strategy and the Buy&Hold strategy. The last column refers to the results of the trading strategy based on moving averages.

Period		Buy&Hold	Trading Strategy					Break even	MA
From 01/01/2018 to 19/05/2020	Transaction costs (bps)	-	0	5	10	15	20	11.07	0
	Final perf.	−3.90%	2.74%	0.53%	−1.68%	−3.88%	−6.09%	−2.15%	−6.95%
	Mean exc.ret	−1.65%	1.16%	0.22%	−0.71%	−1.64%	−2.57%	−0.91%	-2.94%
	Std. dev	22.49%	12.38%	12.39%	12.38%	12.19%	12.40%	12.39%	13.88%
	Sharpe Ratio	−7.30%	9.31%	1.80%	−5.70%	−13.19%	−20.68%	−7.30%	-21.07%
From 01/01/2018 to 31/12/2019	Transaction costs (bps)	-	0	5	10	15	20	5.94	0
	Final perf.	18.81%	13.07%	11.27%	9.47%	7.66%	5.86%	10.93%	8.52%
	Mean exc.ret	9.49%	6.60%	5.69%	4.78%	3.87%	2.96%	5.52%	4.30%
	Std. dev	10.25%	5.95%	5.96%	5.96%	5.97%	5.97%	5.96%	5.11%
	Sharpe Ratio	92.64%	102.92%	95.55%	80.19%	64.86%	49.55%	92.64%	84.13%

Table 9
Statistics and performance measures of the trading strategies for the DOW JONES set of stocks computed using returns on an annual basis. The risk-free rate is $r_f = 0$ while break even refers to the transaction cost that implies the same Sharpe ratio for the trading strategy and the Buy&Hold strategy. The last column refers to the results of the trading strategy based on moving averages.

Period		Buy&Hold	Trading Strategy					Break even	MA
From 01/01/2018 to 19/05/2020	Transaction costs (bps)	-	0	5	10	15	20	12.48	0
	Final perf.	13.46%	12.66%	10.42%	8.17%	5.92%	3.67%	7.05%	−0.67%
	Mean exc.ret	5.69%	5.35%	4.40%	3.45%	2.50%	1.55%	2.98%	−0.29%
	Std. dev	16.46%	8.63%	8.63%	8.63%	8.63%	8.63%	8.63%	10.20%
	Sharpe Ratio	34.43%	61.78%	50.82%	39.86%	28.89%	17.92%	34.43%	-2.75%
From 01/01/2018 to 31/12/2019	Transaction costs (bps)	-	0	5	10	15	20	4.97	0
	Final perf.	26.60%	17.31%	15.44%	13.56%	11.69%	9.81%	15.44%	9.69%
	Mean exc.ret	13.43%	8.74%	7.79%	6.85%	5.90%	4.95%	7.80%	4.89%
	Std. dev	11.27%	6.55%	6.54%	6.54%	6.53%	6.53%	6.54%	7.68%
	Sharpe Ratio	119.17%	133.46%	119.10%	104.71%	90.29%	75.86%	119.17%	63.76%

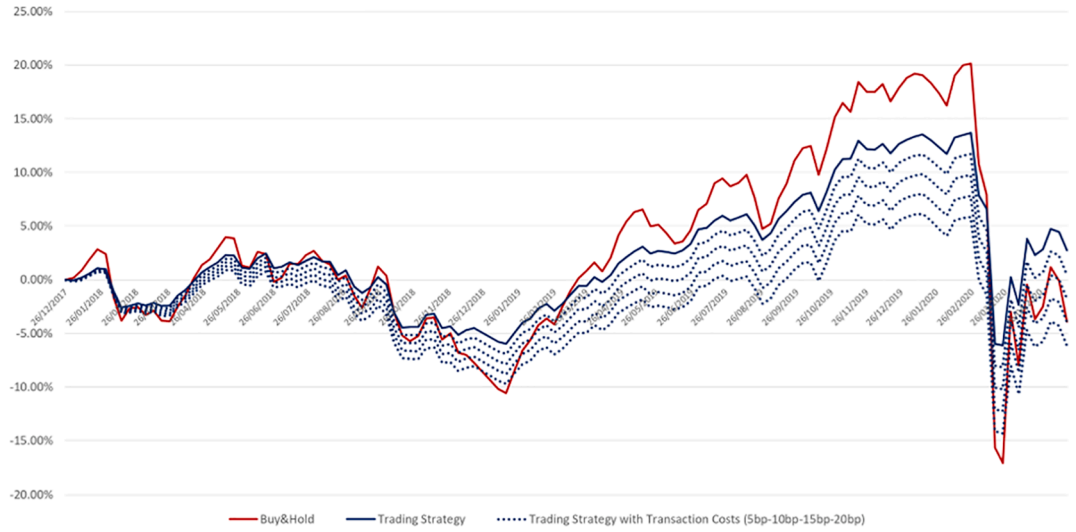

Fig. 2. Cumulative return of the trading strategy and of the Buy&Hold strategy where the set of stocks is provided by the EUROSTOXX50 Index.

outperform the trading strategy. This result shows that the trading strategy performs well in crisis periods as suggested in Kaucic (2010), Kim et al. (2011) and Taylor (2014). Notice that the strategy presented in this paper does not allow for short sales. We have developed the trading strategy allowing for short sales but omit to present the results for the sake of brevity. We notice that allowing for short sales, the trading strategy becomes smoother, the risk-adjusted performance compared to the Buy&Hold gets worse on the shorter subsample and improves over the longer sample.

As a benchmark, in the last column (MA) we also present results of a trading strategy based on two moving averages (MA) computed on prices. The strategy is defined as follows. At time t we define the l-weeks moving average based on closing prices up to time t as follows:

$$MA_{t,l} = \frac{1}{l}\sum_{i=t-l+1}^{t} P_i.$$

We consider a short term moving average ($l = 10$) and a long term moving average ($l = 40$) and consider the a trading rule based on the crossing of the two moving average patterns. In particular:

- (i) we buy the stock if $MA_{t,10} > MA_{t,40}$ and $MA_{t-1,10} < MA_{t-1,40}$
- (ii) we sell the stock if $MA_{t,10} < MA_{t,40}$ and $MA_{t-1,10} > MA_{t-1,40}$.

The two signals are known respectively as the golden cross (i) and the

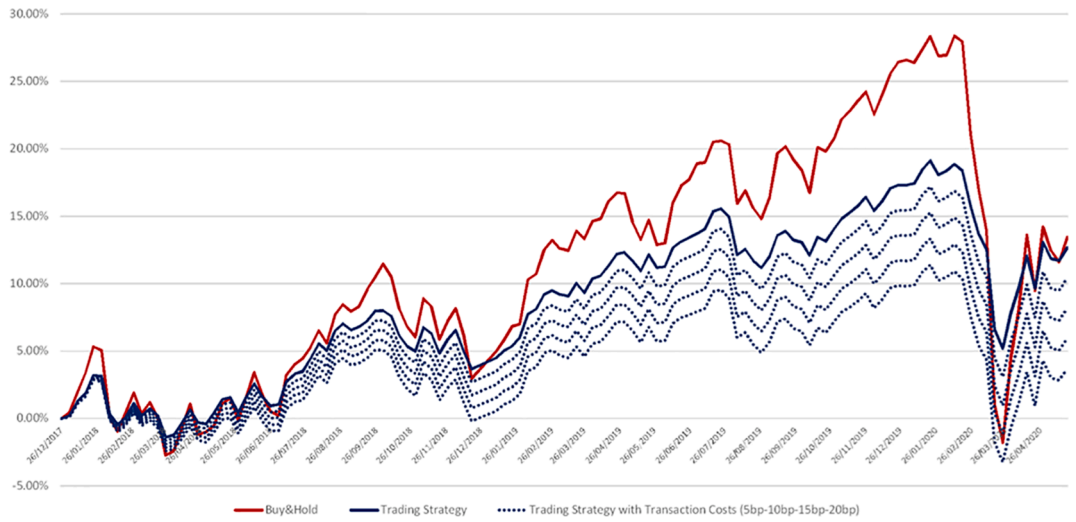

Fig. 3. Cumulative return of the trading strategy and of the Buy&Hold strategy where the set of stocks is provided by the DOW JONES Index.

dead cross (ii) and have been extensively used in literature to build a trading strategy, see e.g. Ni, Liao, and Huang (2015). For both data sets we observe that this strategy does not perform well compared to the Buy&Hold and to the strategy proposed by the algorithm even including transaction costs.

In Tables 10 and 11, we further investigate the outcome of the trading model analyzing number of weeks in which an asset is detained in the portfolio and the frequency of trading. Notice that there is a significant variance in the number of weeks in which the asset is detained (from 20% to more than 80%). Instead we notice that the number of weeks on which there is a trade is less variable. In Fig. 4 we show the

Table 10
Main statistics (mean, standard deviation, Sharpe ratio of weekly returns) and some information on the trading frequencies for stocks belonging to the EUROSTOXX 50 index in the out sample data set for the period 01/01/2018–17/05/2020.

Stock	Average return	Std. Dev. of returns	Sharpe ratio	N. holding weeks /N. weeks	N. Trades /N. weeks
ANHEUSER-BUSCH INBEV	−32.6%	36.9%	−88.2%	40.7%	47.2%
KONINKLIJKE AHOLD DELHAIZE	15.2%	22.6%	67.2%	61.0%	37.4%
ADIDAS	14.7%	34.7%	42.4%	81.3%	18.7%
AIR LIQUIDE	11.7%	21.8%	53.6%	44.7%	38.2%
ASML HOLDING	32.8%	36.0%	91.2%	52.0%	47.2%
AXA	−9.5%	37.5%	−25.5%	26.8%	43.1%
BASF	−25.8%	27.6%	−93.4%	52.8%	36.6%
BAYER	−13.6%	34.6%	−39.2%	35.0%	30.9%
BMW	−15.6%	34.2%	−45.8%	40.7%	35.0%
BNP PARIBAS	−24.2%	36.2%	−66.8%	25.2%	32.5%
CRH	4.8%	41.7%	11.6%	61.0%	39.0%
DAIMLER	−24.7%	42.0%	−58.7%	60.2%	37.4%
DANONE	−1.9%	21.1%	−9.0%	69.9%	39.8%
DEUTSCHE TELEKOM	3.2%	21.5%	14.7%	35.0%	35.0%
ENEL	15.0%	31.1%	48.2%	69.9%	39.0%
ENI	−7.9%	37.5%	−21.1%	63.4%	43.9%
ESSILORLUXOTTICA	2.0%	26.3%	7.6%	74.8%	39.8%
IBERDROLA	19.6%	25.5%	77.0%	82.9%	24.4%
INTESA SANPAOLO	−15.5%	33.0%	−47.1%	69.9%	26.8%
LVMH	17.7%	32.0%	55.2%	58.5%	36.6%
ORANGE	−7.9%	23.5%	−33.5%	33.3%	34.1%
L'OREAL	14.5%	23.7%	61.0%	79.7%	30.1%
BANCO SANTANDER	−34.9%	35.7%	−97.8%	32.5%	37.4%
SAP	10.8%	29.5%	36.5%	74.8%	25.2%
SANOFI	15.9%	22.1%	72.1%	35.0%	37.4%
SCHNEIDER ELECTRIC	12.6%	32.8%	38.5%	55.3%	32.5%
TELEFONICA	−18.5%	31.6%	−58.4%	49.6%	28.5%
VOLKSWAGEN	−2.1%	34.7%	−6.1%	41.5%	32.5%
ALLIANZ	−1.0%	32.1%	−3.1%	61.8%	28.5%
SIEMENS	−4.3%	30.6%	−14.2%	71.5%	46.3%

Table 11
Main statistics (mean, standard deviation, Sharpe ratio of weekly returns) and some information on the trading frequencies for stocks belonging to the DOW JONES index in the out sample data set for the period 01/01/2018–17/05/2020.

Stock	Average return	Std. Dev. of returns	Sharpe ratio	N. holding weeks /N. weeks	N. Trades /N. weeks
3M	−12.0%	27.0%	−44.6%	27.6%	35.8%
AMERICAN EXPRESS	1.7%	23.2%	7.2%	70.7%	36.6%
APPLE	33.2%	27.5%	120.7%	56.9%	33.3%
BOEING	−19.5%	44.8%	−43.6%	61.8%	35.8%
CATERPILLAR	−7.1%	32.0%	−22.3%	70.7%	39.8%
CHEVRON	−6.1%	26.6%	−22.9%	43.1%	53.7%
CISCO SYSTEMS	13.1%	24.6%	53.3%	53.7%	48.8%
COCA COLA	5.2%	21.8%	24.0%	32.5%	29.3%
EXXON MOBIL	−16.3%	26.6%	−61.2%	68.3%	29.3%
GOLDMAN SACHS GP.	−9.5%	25.9%	−36.6%	52.0%	38.2%
HOME DEPOT	14.6%	24.7%	59.2%	44.7%	39.0%
INTEL	19.0%	29.7%	64.2%	71.5%	28.5%
INTERNATIONAL BUS. MCHS.	−0.5%	26.9%	−1.9%	51.2%	39.8%
JP MORGAN CHASE & CO.	−1.8%	23.6%	−7.5%	71.5%	36.6%
JOHNSON & JOHNSON	7.8%	20.4%	38.0%	58.5%	37.4%
MCDONALDS	8.1%	24.5%	33.1%	40.7%	43.1%
MERCK & COMPANY	17.4%	21.4%	81.4%	38.2%	26.8%
MICROSOFT	37.0%	20.6%	180.1%	45.5%	35.8%
NIKE 'B'	18.2%	26.2%	69.5%	69.1%	41.5%
PFIZER	6.6%	21.3%	31.0%	83.7%	23.6%
PROCTER & GAMBLE	13.4%	18.6%	72.5%	48.8%	27.6%
RAYTHEON TECHNOLOGIES	−0.9%	27.9%	−3.2%	67.5%	41.5%
TRAVELERS COS.	−10.8%	22.5%	−48.0%	66.7%	35.0%
UNITEDHEALTH GROUP	17.8%	28.6%	62.3%	67.5%	32.5%
VERIZON COMMUNICATIONS	6.8%	18.2%	37.0%	30.1%	35.8%
VISA	25.2%	21.9%	115.1%	35.8%	35.8%
WALGREENS BOOTS ALLIANCE	−14.5%	28.0%	−52.0%	26.0%	39.0%
WALMART	12.6%	17.5%	71.9%	83.7%	23.6%
WALT DISNEY	6.4%	23.9%	26.6%	62.6%	43.9%

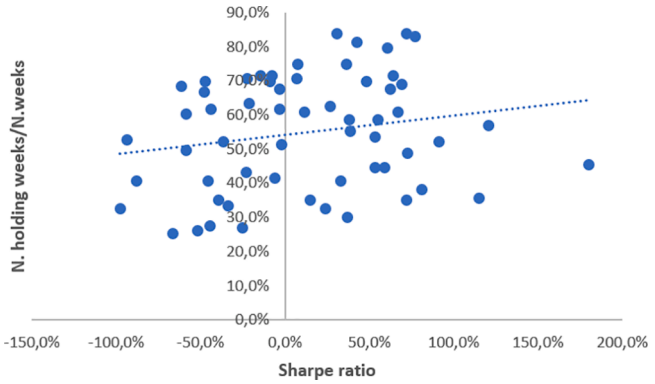

Fig. 4. Scatter plot of the pair (Sharpe ratio, N. holding weeks/N.weeks) for all stocks in the EUROSTOXX 50 and in the DOW JONES index.

relationship between the number of holding weeks over the number of weeks and Sharpe ratio for all the stocks considered in the two indexes. A positive relationship is observed highlighting that in this case the trading model is able to pick the best performing stocks.

Our machine learning methodology allows us to assess the informative content of the indicators and therefore the empirical relevance of the insights provided by the financial theory on the dissemination of private information in financial markets. In Tables 12 and 13 we provide statistics on the selection of predictors. The application to the two data sets provide similar results. Confirming the literature on momentum strategies that are built on a continuation of returns in the short run, the most relevant indicator turns out to be the anomaly of stock return, then the one on the autoregressive structure of return-volume turns out to provide significant information. Instead, excessive trading volume and large bid-ask spread are not informative as the theoretical literature would suggest. About the selectivity of the predictors we observe that a high confidence level (high c) is chosen most of the times and that a long enough estimation window (at least four months) is employed most of the times.

6. Conclusions

Exploiting machine learning tools, in this paper we have tried to answer a long standing question: Do financial time series reflect the dissemination of private information? We answer this question using a methodology that starts from a large set of indicators and aims to build a profitable trading strategy based on outliers of financial time series. We show that outliers in financial time series associated with the dissemination of private information contain some economic value as they allow to build a profitable trading strategy. The strategy is smoother than the Buy&Hold strategy and provides a better risk adjusted performance in particular in a bear period. However, excess performance disappears if transaction costs are included.

Among the indicators that are relevant to predict future returns we have three interesting results: first of all, the centrality of return to predict return in the short run is confirmed as the literature on momentum strategies suggests; contrary to the literature on asymmetric/heterogeneous information, the bid-ask spread and the trading volume time series do not contain interesting information; instead, a structural break in the autocorrelation of returns and in the lead-lag relation between return and trading volume turns out to have an economic value.

In future studies, the performance of the Information Gain Forward–Backward Model selection algorithm could be investigated with other high dimensional data sets. Also the combination of other feature selection methods and other machine learning algorithms could be explored.

Table 12
Percentage of selected predictors for stocks belonging to the EUROSTOXX50 index.

1-c%	Ex.BA	Ex.Ret	Ex. TrVo	Autoreg. Str	Ex. Vola	Overall
N.3 4 5	1.96%	14.76%	1.04%	–	–	17.77%
<90%	1.12%	3.73%	0.96%	–	–	5.81%
>=90%	0.84%	11.03%	0.09%	–	–	11.95%
N.6 8 10	1.51%	11.34%	2.11%	10.11%	–	25.06%
<90%	0.82%	3.11%	1.35%	2.56%	–	7.83%
>=90%	0.69%	8.23%	0.76%	7.55%	–	17.23%
N. 12 26 38	4.15%	9.49%	4.49%	11.41%	–	29.55%
<90%	1.06%	3.20%	0.78%	6.21%	–	11.26%
>=90%	3.09%	6.29%	3.71%	5.20%	–	18.29%
N.52 78 104	4.28%	7.92%	3.48%	7.98%	3.96%	27.63%
<90%	1.12%	2.71%	0.65%	4.44%	–	8.92%
>=90%	3.16%	5.21%	2.83%	3.54%	–	14.74%
A in [0.4;1[	–	–	–	–	2.91%	2.91%
A in [1; 1.6[	–	–	–	–	1.05%	1.05%
Overall	11.91%	43.51%	11.11%	29.50%	3.96%	100.00%

Table 13
Percentage of selected predictors for stocks belonging to the DOW JONES index.

1-c%	Ex.BA	Ex.Ret	Ex. TrVo	Autoreg. Str	Ex. Vola	Overall
N.3 4 5	1.51%	17.25%	1.56%	–	–	20.32%
<90%	0.76%	5.19%	1.38%	–	–	7.33%
>=90%	0.76%	12.06%	0.17%	–	–	12.99%
N.6 8 10	2.13%	10.13%	1.91%	9.84%	–	24.01%
<90%	1.06%	3.72%	1.47%	3.30%	–	9.54%
>=90%	1.07%	6.41%	0.44%	6.54%	–	14.46%
N.12 26 38	3.80%	8.98%	4.70%	11.43%	–	28.92%
<90%	0.94%	3.29%	1.30%	5.23%	–	10.76%
>=90%	2.86%	5.70%	3.41%	6.19%	–	18.16%
N.52 78 104	4.19%	6.77%	2.93%	7.82%	5.05%	26.75%
<90%	1.24%	2.32%	0.49%	5.42%	–	9.47%
>=90%	2.95%	4.44%	2.44%	2.40%	–	12.24%
A in [0.4;1[	–	–	–	–	3.19%	3.19%
A in [1; 1.6[	–	–	–	–	1.86%	1.86%
Overall	11.64%	43.13%	11.10%	29.08%	5.05%	100.00%

CRediT authorship contribution statement

Emilio Barucci: Methodology, Writing - original draft, Writing - review & editing, Supervision, Validation. **Michele Bonollo:** Methodology, Supervision, Writing - review & editing. **Federico Poli:** Methodology, Data curation, Writing - original draft, Writing - review & editing. **Edit Rroji:** Methodology, Data curation, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work has been partially supported from the European Union's Horizon 2020 training and innovation program "FIN-TECH", under the grant agreement No. 825215 (Topic ICT-35-2018, Type of actions: CSA)

Appendix A. Naive-Bayes classification algorithm

The goal of the classification algorithm is to predict a class label for a given set of input variables. Suppose we have K class labels $Y_1, Y_2, \ldots, Y_K$ and S input variables $X_1, X_2, \ldots, X_S$. If we compute the conditional probabilities

$$\mathbb{P}(Y_k|X_1, X_2, \ldots, X_S)$$

for each label $k = 1,\ldots,K$, then the class with the highest probability is considered to be the most likely outcome in the classification exercise. If we use the Bayes Theorem for the computation of the conditional probability we have

$$\mathbb{P}\left(Y_k|X_1, X_2, \ldots, X_S\right) = \frac{\mathbb{P}(X_1, X_2, \ldots, X_S|Y_k)\mathbb{P}(Y_k)}{\mathbb{P}(X_1, X_2, \ldots, X_S)}$$

where $\mathbb{P}(Y_k)$ is the prior (probability) of Y_k that can be computed from the data as the ratio between the number of observations yielding Y_k over the total number of observations in the sample. The computation of $\mathbb{P}(X_1, X_2, \ldots, X_S|Y_k)\mathbb{P}(Y_k)$ is more complex, especially as the number of input variables S increases.

The Naive-Bayes approach reduces the computation complexity by considering each input variable X_s as being independent from the others. Thanks to this assumption

$$\mathbb{P}(Y_k|X_1, X_2, \ldots, X_S) \propto \mathbb{P}(X_1, X_2, \ldots, X_S|Y_k)\mathbb{P}(Y_k) = \mathbb{P}(X_1|Y_k) \times \ldots \times \mathbb{P}(X_S|Y_k)\mathbb{P}(Y_k) \tag{3}$$

as $\mathbb{P}(X_1, X_2, \ldots, X_S)$ appears in the conditional probability of each class label and has a normalizing effect in the results.

This decision rule is refereed to as the Maximum a Posteriori rule for a classification exercise. The label $\overline{k}$ of the response variable Y with the largest probability computed as in (3) represents the classification outcome for the classification exercise, i.e.

$$\overline{k} = argmax_{k=1,\ldots,K}\mathbb{P}\left(Y_k\right)\prod_{i=1}^{S}\mathbb{P}\left(X_i|Y_k\right)$$

Local distributions $\mathbb{P}(X_i|Y_k)$ are specified by parameters $\Theta(X_i, Y_k)$. It is common to assume each local distribution to have parametric form, such as multinomial for discrete variables, or gaussian for continuous variables. Assuming a Dirichlet prior for $\Theta(X_i, Y_k)$ and the same hyperparameter α for all the local distributions, then the Bayesian estimator can be obtained as follows in closed form:

$$\Theta_{ijk} = \frac{N_{ikj} + \alpha}{N_{.k.} + r_i\alpha}$$

where N_{ikj} is the number of observations such that $X_i = j$ and $Y = k$. $N_{.k.}$ is the number of samples in which $Y = k, r_i$ is the number of possible values of X_i and $\alpha > 0$ is a prior hyper-parameter. Given different values of α the resulting estimate can vary between the empirical probability $\frac{N_{ijk}}{N_{j.}}$ given by relative frequency ($\alpha = 0$) and the uniform probability $\frac{1}{r_i}$ $(\alpha \gg 0)$. In this paper we use the `bnclassify` R package introduced in Mihaljevic, Bielza, and Larrañaga (2020) and assume $\alpha = 1$; this technique is called add-one smoothing (or additive smoothing in general). The goal is to increase the zero or near to zero probability values to a small positive number, imposing a uniform prior. For instance multiplying the probabilities during inference, a single zero value can bring down to zero the posterior probability.

References

Abellán, J., & Castellano, J. G. (2017). Improving the naive bayes classifier via a quick variable selection method using maximum of entropy. *Entropy, 19*(6), 247.

Allen, F., & Karjalainen, R. (1999). Using genetic algorithms to find technical trading rules. *Journal of Financial Economics, 51*, 245–271.

Andersen, T., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 885–905.

Arévalo, R., García, J., Guijarro, F., & Peris, A. (2017). A dynamic trading rule based on filtered flag pattern recognition for stock market price forecasting. *Expert Systems with Applications, 81*, 177–192.

Bajgrowicz, P., & Scaillet, O. (2012). Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Finance, 106*, 473–491.

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications, 42*(20), 7046–7056.

Barucci, E., Bianchi, C., Casciari, F., & Squillantini, E. (2006). Market abuse detection: a methodology based on financial time series. *Statistica Applicata*, 559–571.

Bettis, C., Cole, J., & Lemmon, M. (2000). Corporate policies restricting trading by insiders. *Journal of Financial Economics, 57*, 191–220.

Biais, B., Bossaerts, P., & Spatt, C. (2010). Equilibrium asset pricing and portfolio choice under asymmetric information. *The Review of Financial Studies, 23*(4), 1503–1543.

Bianchi, F., Mercuri, L., & Rroji, E. (2021). Portfolio selection with irregular time grids: an example using an ICA-COGARCH (1, 1) approach. *Financial Markets and Portfolio Management*, 1–29.

Blume, L., Easley, D., & O'Hara, M. (1994). Market statistics and technical analysis: the role of volume. *The Journal of Finance, 69*, 153–181.

Boonamnuay, S., Nittaya, K., & Kittisak, K. (2018). Classification and regression tree with resampling for classifying imbalanced data. *International Journal of Machine Learning and Computing, 8*(4), 336–340.

Brock, W., Lakonishok, J., & LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns. *The Journal of Finance, 47*, 1731–1764.

Campbell, J., Grossman, S., & Wang, J. (1993). Trading Volume and Serial Correlation in Stock Returns. *Quarterly Journal of Economics, 108*, 905–939.

Cao, C., Field, L., & Hanka, G. (2004). Does insider trading impair market liquidity? Evidence from IPO lockup expiration. *Journal of Financial and Quantitative Analysis*, 25–46.

Cervelló-Royo, R., & Guijarro, F. (2020). Forecasting stock market trend: a comparison of machine learning algorithms. *Finance, Markets and Valuation, 6*, 37–49.

Cervelló-Royo, R., Guijarro, F., & Michniuk, K. (2015). Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Systems with Applications, 42*(14), 5963–5975.

Chakravarty, S. (2001). Stealth-trading: Which traders' trades move stock prices? *Journal of Financial Economics, 61*(2), 289–307.

Chakravarty, S., & McConnell, J. (1999). Does insider trading really move stock prices? *Journal of Financial and Quantitative Analysis, 34*(2), 191–209.

Cheng, L., Firth, M., Leung, T., & Rui, O. (2006). The effects of insider trading on liquidity. *Pacific-Basin Finance Journal*.

Chen, Y., & Wang, X. (2015). A hybrid stock trading system using genetic network programming and mean conditional Value-at-Risk. *European Journal of Operational Research, 40*, 861–871.

Chong, E., Han, C., & Park, F. (2017). Deep learning networks for stock market analysis and prediction: methodology, data representations, and case studies. *Expert systems with applications, 83*, 187–205.

Chung, K., & Charoenwong, C. (1998). Insider trading and the bid-ask spread. *The Financial Review, 33*, 1–20.

Collin-Dufresne, P., & Fos, V. (2015). Do prices reveal the presence of informed trading? *Journal of Finance, 70*(4), 1555–1582.
Conrad, J., Hameed, A., & Niden, C. (1994). Volume and autocovariances in short-horizon individual security returns. *Journal of Finance, 49*, 1305–1329.
Copeland, T., & Galai, D. (1983). Information effects on the bid–ask spread. *Journal of Finance, 38*, 1457–1468.
Cornell, B., & Sirri, E. (1992). The reaction of investors and stock prices to insider trading. *Journal of Finance, 47*(3), 1031–1059.
Fama, E. (1970). Efficient capital markets: a review of theory and empirical work. *Journal of Finance, 25*, 383–417.
Fang, J., Jacobsen, B., & Qin, Y. (2014). Predictability of the simple technical trading rules: An out-of-sample test. *Review of Financial Economics, 23*, 30–45.
Fisher, T., & Krauss, C. (2018). Deep learning with long short-memory networks for financial market predictions. *European Journal of Operational Research, 270*, 654–669.
Foucault, T., Pagano, M., & Ailsa, R. (2013). *Market liquidity*. Oxford University Press.
Gerlein, E., McGinnity, M., Belatreche, A., & Coleman, S. (2016). Evaluating machine learning classification for financial trading: an empirical approach. *Expert Systems with Applications, 54*, 193–207.
Glosten, L., & Milgrom, P. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics, 14*, 71–100.
Goodhart, C., & O'Hara, M. (1997). High frequency data in financial markets: Issues and applications. *Journal of Empirical Finance, 4*, 73–114.
Grossman, S. (1989). *The informational role of prices* (p. 1). MIT Press Books.
He, H., & Wang, J. (1995). Differential information and dynamic behavior of stock trading volume. *Review of Financial Studies, 8*, 919–972.
Holopainen, M., & Sarlin, P. (2017). Toward robust early-warning models: A horse race, ensembles and model uncertainty. *Quantitative Finance, 17*(12), 1933–1963.
Hsu, P., Hsu, Y., & Kuan, C. (2010). Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance, 17*, 471–484.
Huang, C. (2012). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing, 12*, 807–818.
Hu, Y., Feng, B., Zhang, X., Ngai, E., & Liu, M. (2015). Stock trading rule discovery with an evolutionary trend following model. *Expert Systems with Applications, 42*(1), 212–222.
Jadhav, S., He, H., & Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing, 69*, 541–553.
Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency. *Journal of Finance, 56*, 699–720.
Jeng, L., Metrick, A., & Zeckhauser, R. (2003). Estimating the returns to insider trading: a performance-evaluation perspective. *The Review of Economics and Statistics, 85*(2), 453–471.
Jurman, G., Riccadonna, S., & Furlanello, C. (2012). A comparison of mcc and cen error measures in multi-class prediction. *PloS one, 7*(8), Article e41882.
Kaucic, M. (2010). Investment using evolutionary learning methods and technical rules. *European Journal of Operational Research, 207*, 1717–1727.
Kim, J., Lim, K., & Shamsuddin, A. (2011). Stock return predictability and adaptive markets hypothesis: evidence from century-long us data. *Journal of Empirical Finance, 18*, 868–879.
Kim, O., & Verrecchia, R. (1991). Market reaction to anticipated announcements. *Journal of Financial Economics, 30*, 273–309.
Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. In *European conference on machine learning* (pp. 171–182). Springer.
Kyle, A. (1985). Continuous auctions and insider trading. *Econometrica, 53*, 1315–1335.
Lee, M. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications, 36*, 10896–10904.
Leung, M., Daouk, H., & Chen, A. S. (2000). Forecasting stock indices: a comparison of classification and level estimation models. *International Journal of Forecasting, 16*, 173–190.
Lin, Q. (2018). Technical analysis and stock return predictability: An aligned approach. *Journal of Financial Markets, 38*, 103–123.
Llorente, G., Michaely, R., Saar, G., & Wang, J. (2001). Dynamic volume-return relation of individual stocks. NBER Working paper 8312.
Lo, A. (2004). The adaptive markets hypothesis: market efficiency from an evolutionary perspective. *Journal of Portfolio Management, 30*, 15–29.
Lo, A., Mamaysky, H., & Wang, J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The Journal of Finance, 55*, 1704–1765.
Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta year=BBA,-Protein Structure, 405*, 442–451.
McGorty, F., Gwilym, O., & Thomas, S. (2009). The role of private information in return volatility, bid-ask spreads and price levels in the foreign exchange market. *Journal of International Financial Markets, Institutions Money, 19*, 387–401.
Meulbroek, L. (1992). An empirical analysis of illegal insider trading. *Journal of Finance, 47*(5), 1661–1699.
Mihaljevic, B., Bielza, C., & Larrañaga, P. (2020). bnclassify: Learning bayesian network classifiers. *R package version*, (4), 5.
Muncharaz, J. O. (2020a). Comparing classic time series models and the LSTM recurrent neural network: An application to S&P 500 stocks. *Finance, Markets and Valuation, 6*, 137–148.
Muncharaz, J. O. (2020b). Leading research trends on trading strategies. *Finance, Markets and Valuation, 6*, 28–54.
Neely, C., Weller, P., & Dittmar, R. (1997). Is technical analysis in the foreign exchange market profitable? a genetic programming approach. *Journal of Financial and Quantitative Analysis, 32*, 405–426.
Neely, C., Weller, P., & Ulrich, J. (2009). The adaptive markets hypothesis: evidence from the foreign exchange market. *Journal of Financial and Quantitative Analysis, 44*, 467–488.
Neftci, S. (1991). Naive trading rules in financial markets and Wiener-Kolmogorov prediction theory: a study of technical analysis. *Journal of Business, 64*, 549–571.
Ni, Y., Liao, Y., & Huang, P. (2015). MA trading rules, herding behaviors, and stock market overreaction. *International Review of Economics & Finance, 39*, 253–265.
O'Hara, M. (1995). *Market microstructure theory*. Blackwell Press.
Paiva, F. D., Cardoso, R. T. N., Hanaoka, G. P., & Duarte, W. M. (2019). Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. *Expert Systems with Applications, 115*, 635–655.
Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *The Journal of Business, 53*, 61–65.
Rachev, A., Jasic, T., Stoyanov, S., & Fabozzi, F. (2007). Momentum strategies based on reward-risk stock selection criteria. *Journal of Banking and Finance, 31*, 2325–2346.
Rosu, I. (2019). Fast and slow informed trading. *Journal of Financial Markets, 43*, 1–30.
Seyhun, H. N. (1986). Insiders? profits, costs of trading and market efficiency. *Journal of Financial Economics, 16*, 189–212.
Seyhun, H. N. (1992). Why does aggregate insider trading predict future stock returns? *Quarterly Journal of Economics, 107*, 1303–1331.
Sullivan, R., Timmermann, A., & White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance, 54*, 1647–1691.
Taylor, N. (2014). The rise and fall of technical trading rule success. *Journal of Banking and Finance, 40*, 286–302.
Wang, J. (1994). A model of competitive stock trading volume. *Journal of Political Economy, 102*, 127–168.