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

Emilio Barucci,*Michele Bonollo, Federico Poli and Edit Rroji

Department of Mathematics, Politecnico di Milano

Abstract

We build an algorithm for stock selection based on indicators of time series of stocks (return, volume, volatility, bid-ask spread) that should be associated with the dissemination of private information in financial markets. We use a machine learning algorithm for the identification of the most relevant indicators for the prediction of stock returns and to define a trading strategy. The procedure combines a sequential inclusion of predictors with a classification algorithm for the trading signal. 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 5-10% transaction costs are inserted.

Keywords: Stock picking, Technical analysis, Transaction costs, Classification algorithm, Information Gain.

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)

*Corresponding author, e-mail: emilio.barucci@polimi.it.

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 (private-asymmetric) information in financial markets. Financial markets theory provides a large set of such indicators but the empirical evidence on their capability to reflect private information and on their usefulness to trade successfully is lacking. In what follows, we adopt an agnostic approach, we start from a wide set of indicators and we adopt a machine learning technique (wrapping procedure 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 inputs, the algorithm receives indicators of outliers of the financial time series that are associated by the financial theory to the dissemination of private-asymmetric information in financial markets. Such indicators are identified as outliers of the following variables: return with respect to the "market model", trading volume growth, bid-ask spread and volatility, serial correlation of returns and trading volume. The combination of 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: 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 [19, 21, 30, 44, 63] for theoretical analysis and [18, 49] for empirical analysis, see also [35, 59, 60] 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 [34, 58, 62], and to papers that evaluate the performance of technical analysis strategies, e.g., see [3, 9, 24, 31, 38, 46, 54, 61]. 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. 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 [2, 14, 16, 25, 26, 31, 32, 33, 37, 42, 52, 56].

The stock picking/trading algorithm works as follows. We start with a large set of predictors that are supposed to be useful for the definition of 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 [1, 40] for details. Then we build a hybrid approach (forward-backward) for the identification of most useful predictors for the stock return one step ahead. The procedure uses the Naive-Bayes classification algorithm that learns to predict the decision strategy once observed the set of market indicators. The trading model is selected evaluating the models on the validation set according to accuracy indicators.

We apply the methodology to two set 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 considering the Sharpe ratio as performance metric.

Performance results are mixed. Assuming no transaction costs, the trading strategy outperforms the Buy&Hold strategy in all the four out of sample subsets. Including transaction costs (5-10 basis points) outperformance disappears. This result is in line with the recent literature, e.g., see [53, 61, 62], showing that profitability of technical trading rules is weak in recent times because markets are becoming more efficient (traders are using them), e.g., adaptive market hypothesis see [45]. However the results are more positive than recent literature on the profitability of technical analysis, e.g., in [3] 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 [37, 38, 62] 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 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 marginal 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 of our methodology. 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. We refer to [4] for a reference on these topics.

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 [23]: 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. However, return serial correlation cannot be interpreted univocally as a signal of insider trading/private information. There is a large literature showing that asset returns with a holding period smaller than one year are positively serially correlated and that returns with a holding period greater than one year are negatively serially correlated, see [4, 10, 34]. The phenomenon is also observed on equity indexes and therefore cannot be attributed entirely to insider trading/market manipulation. As a matter of fact, the regularity may be due to time varying risk premia or to behavioral biases.

In the presence of insider trading and market abuse, return serial correlation is expected. While the insider trader always trades a limited amount in the direction of his information, the manipulator either releases information and trades in the opposite direction or trades intertemporally in different directions to gain from sequential trades (e.g. pump and dump strategies). Therefore, if insider trading occurs, then we expect positive (negative) daily or weekly returns to follow positive (negative) returns because private information is incorporated gradually in asset prices, in case of manipulation we expect a short term price reversal (mean reversion) due to the release of false information or to large trades in different directions. A model that rationalizes this type of behavior of insider traders is provided by [41].

Insight 1. In the presence of private information dissemination, we observe positive serial correlation in daily 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 a set of 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 (e.g. asset returns are distributed as a normal random variable) 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 the absence of private information dissemination, daily 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 provide an explanation to several stylized facts. In particular, the literature 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.

The literature on markets with heterogeneous information is quite large. 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, i.e., equilibrium prices are fully revealing, they instantaneously reveal private information and coincide with those of an economy where all private signals are public (they are observed by all agents), see [29]. 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 [39]. Therefore, precise private information (insider trading) is associated with large trades, for a discussion on the relationship between trading size and information content see [12].

Insight 3. In the 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, investment decision or mergers, 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 funds for liquidity reasons with no information content. As a consequence, serial correlation of trading volume is an interesting way 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 [30].

Insight 4. When public information arrives on the market, daily 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 [19, 21]; instead, if trading volume is due to private information then the relation can have a different sign, i.e., positive return-volume correlation, see [6, 44, 48, 63].

Insight 5. In the 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 [55]. In adverse selection models, see [22, 27], 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 [28].

Insight 6. In the 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. [49] 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 [18]. Weak evidence on price movements associated with insider trades has been detected in [13]. On trading by directors and illiquidity the evidence is mixed: [5, 11, 15, 17] provide evidence that spread widens and market depth falls on insider trading days as compared to non-insider trading days; [18, 20] provide no evidence.

3 Market indicators

We consider the following time series for each stock on a weekly basis[1]:

- 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 $\frac{P_t-P_{t-1}}{P_{t-1}}$.
- 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 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.

[1]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.

We opt for weekly observations as a week 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 signal outliers in the time series with respect to regularities that are identified according to the literature on asymmetric/private information. The fourth indicator renders three different values: $idx_4 \in \{-1,0,1\}$. The indicators are 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[2]. 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 [7]. 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 $(\hat{\alpha}_{0t}, \hat{\alpha}_{1t}, \hat{\alpha}_{2t})$ to obtain a forecast of the volatility at time t $(\hat{\sigma}_t^2)$. This value is compared to the realized range volatility estimator, see [57]:

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

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

Then, $idx_3^A = 1$ if and only if $A\hat{\sigma}_t^2 < s_t^2$, 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 $\{r^*_{t-j}\}_{j=1,\dots,N}$, computed as a weighted average of the total return of each security belonging to the index:

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

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

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

   we compare it to the observed return r_t. Then, we set $idx_4^c = 1$ if r_t is above the 1-c% quantile, $idx_4^{(c)} = -1$ if r_t 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 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 and 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 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 $N \geq 6$.

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.

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.

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 sum required to buy one unit of each stock in the set of eligible assets.

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 set 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.

The procedure is represented in Figure 1.

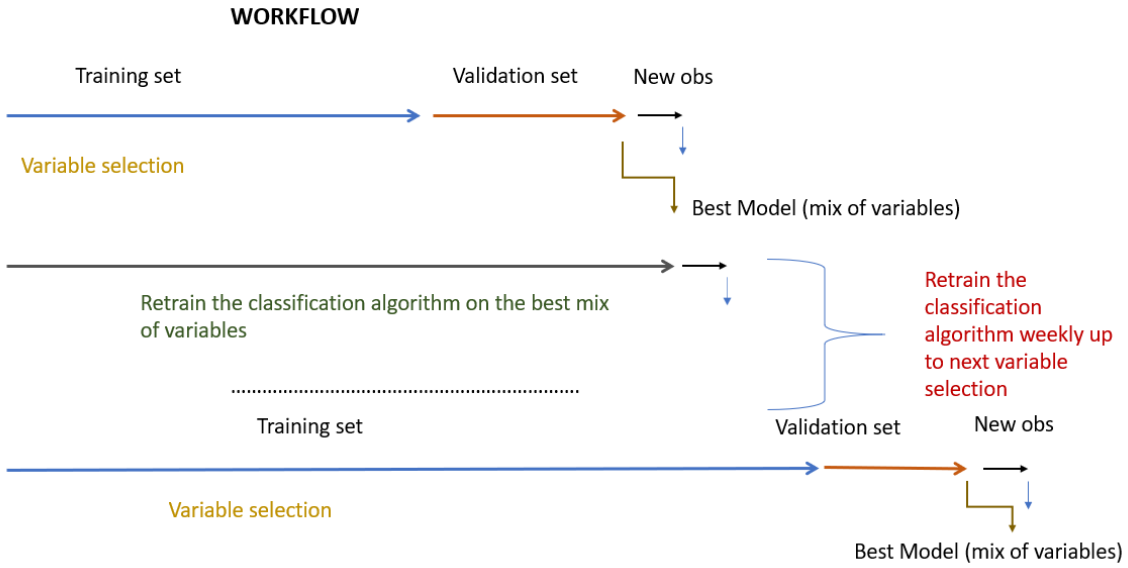

Figure 1: Scheme of the trading algorithm workflow

4.2 Selection procedure

The selection procedure at the heart of the trading algorithm builds on several steps. The trading signal is built through a classification algorithm where for each stock 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.

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;
3. choose the trading model among the models (eleven models obtained for different 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 performed 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 choose the Mathew's Correlation Coefficient (MCC) originally developed in [47] 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, 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 $[-1, +1]$, -1 and $+1$ are obtained in case of perfect misclassification and perfect classification, respectively. We choose the MCC ratio as it can be easily extended to the case of multiclass response variable and it is well-suited for unbalanced data sets (see [8, 36] for details).

The trading signal (BUY, NEUTRAL or SELL) 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 metrics for model comparison defined as Return Weighted Accuracy (RWA). The indicator builds on the Accuracy measure which is defined as the proportion of correct predictions among the total number of cases considered in the binary classification problem:

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

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, 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.

Given a set of $\bar{n}$ variables, the procedure starts with $2^{\bar{n}}$ 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 is driven by an increase of $MCC + RWA$ on the validation set for the Naive-Bayes classifier calibrated on the training set.

We have to define a sequential order for the introduction of a predictor. 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 [40].

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

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

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:

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

where $p(x,y)$ is the probability of observing a realization x of X and y of Y. This quantity 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 variables 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 problem, 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), 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 (eleven values). Then we choose (step 3) the trading model ($\bar{\theta}$, the corresponding $y_{t,\bar{\theta}}$ and the predictors selected as described above thanks to the MCC and the RWA metrics). This choice is driven by the θ and model yielding the largest $MCC + RWA$ in the validation set.

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 Table 1 and 2 respectively. The market capitalization of stocks included in both cases accounts for about 67% of the two indexes. 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 $80-20\%$ fraction). 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 fraction. Few stocks of the two indexes 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.

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 ($8/01/2004-31/12/2018$). For each week of the data set we use the information provided by predictors to extract a trading signal which is matched to a 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. The parameters of the classifier are defined using all the information contained in the training/validation set. Then 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, but the parameters of the trading model selected are calibrated using all the information contained in training/validation set .

In Table 3 we present the main features of the trading algorithm for the stocks belonging to the two indexes. 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: $N = 18$ 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. 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 4 and 5 we report the performance measures of the trading strategy provided by the trading algorithm. The performance is computed on the full out of sample data

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	0.04574	-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 1: Stocks included in the analysis of the EUROSTOXX50, main statistics are computed using weekly returns.

Stock	**Weight**	**Mean**	**Std. dev**	**Skewness**	**Kurtosis**
3M	1.84%	-0.0020	0.0374	-0.6157	1.8742
AMERICAN EXPRESS	1.19%	0.0010	0.0328	-1.3955	5.7337
APPLE	1.93%	0.0064	0.0380	-0.1335	1.5649
BOEING	2.72%	-0.0028	0.0627	-3.2376	23.3027
CATERPILLAR	5.01%	-0.0010	0.0444	-0.1492	2.2808
CHEVRON	1.55%	-0.0008	0.0369	-0.5581	3.8797
CISCO SYSTEMS	2.05%	0.0025	0.0340	-0.4608	1.0765
COCA COLA	2.56%	0.0013	0.0303	-1.6111	10.0665
EXXON MOBIL	1.73%	-0.0027	0.0370	-0.9843	3.9654
GOLDMAN SACHS GP.	0.82%	-0.0010	0.0370	-0.2659	1.8583
HOME DEPOT	1.04%	0.0029	0.0341	-2.3288	15.4366
INTEL	2.08%	0.0039	0.0411	-0.4232	1.0386
INTERNATIONAL BUS.MCHS.	2.15%	0.0000	0.0371	-0.7384	2.6631
JP MORGAN CHASE & CO.	2.50%	0.0003	0.0334	-0.6655	2.1219
JOHNSON & JOHNSON	1.01%	0.0013	0.0283	-1.0686	5.5331
MCDONALDS	1.79%	0.0018	0.0339	-3.5954	29.3686
MERCK & COMPANY	2.78%	0.0033	0.0295	0.0329	2.3172
MICROSOFT	1.31%	0.0070	0.0284	-0.4342	1.9535
NIKE 'B'	4.55%	0.0039	0.0364	-1.4203	10.0375
PFIZER	1.25%	0.0012	0.0294	-0.2403	1.8999
PROCTER & GAMBLE	3.01%	0.0025	0.0256	-0.6316	5.7054
RAYTHEON TECHNOLOGIES	1.92%	0.0007	0.0398	-1.2525	8.2337
TRAVELERS COS.	5.44%	-0.0012	0.0327	-1.5021	10.1541
UNITEDHEALTH GROUP	4.59%	0.0036	0.0395	-0.4904	2.3060
VERIZON COMMUNICATIONS	2.10%	0.0013	0.0252	-0.1837	1.0156
VISA 'A'	1.12%	0.0050	0.0303	-1.3132	4.4255
WALGREENS BOOTS ALLIANCE	0.94%	-0.0023	0.0390	-0.2318	1.0654
WALMART	3.08%	0.0023	0.0241	-0.0789	2.1764
WALT DISNEY	2.93%	0.0017	0.0334	-0.6310	4.9514

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

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

Table 3: Database description

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 are 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 last column we also report the level of 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 Figure 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 datasets. 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 outperform the trading strategy. This result shows that the trading strategy performs well in crisis periods as suggested in [37, 38, 62]. Notice that our strategy does not allow for short sales. We have developed the trading strategy allowing for short sales. We 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.

Our machine learning methodology allows us to assess the informative content of the indicators. In Table 6 and 7 we provide statistics on the selection of predictors. The two applications 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 window (at least four months) is employed most of the times.

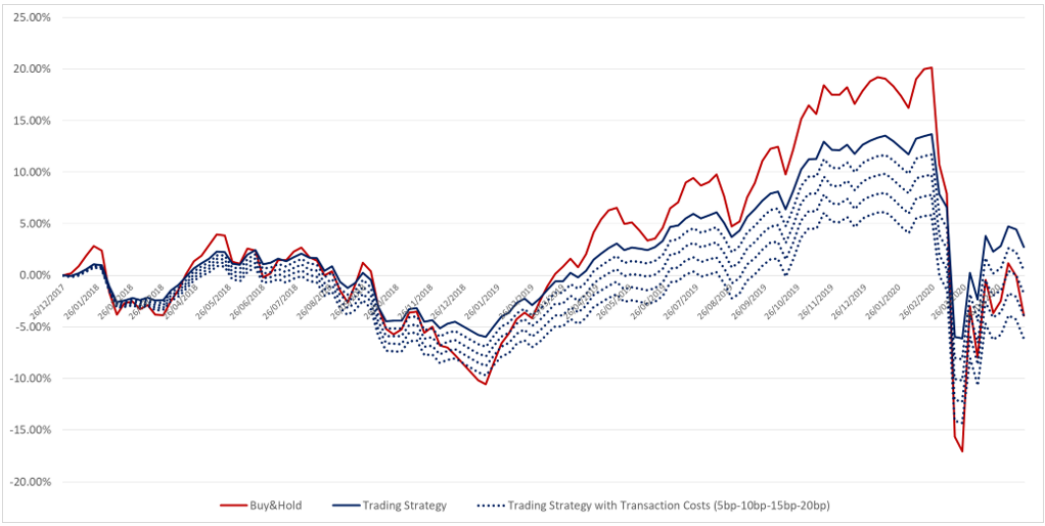

Figure 2: Cumulative return of the trading strategy and of the Buy&Hold strategy where the set of stocks is provided by the EUROSTOXX50 Index.

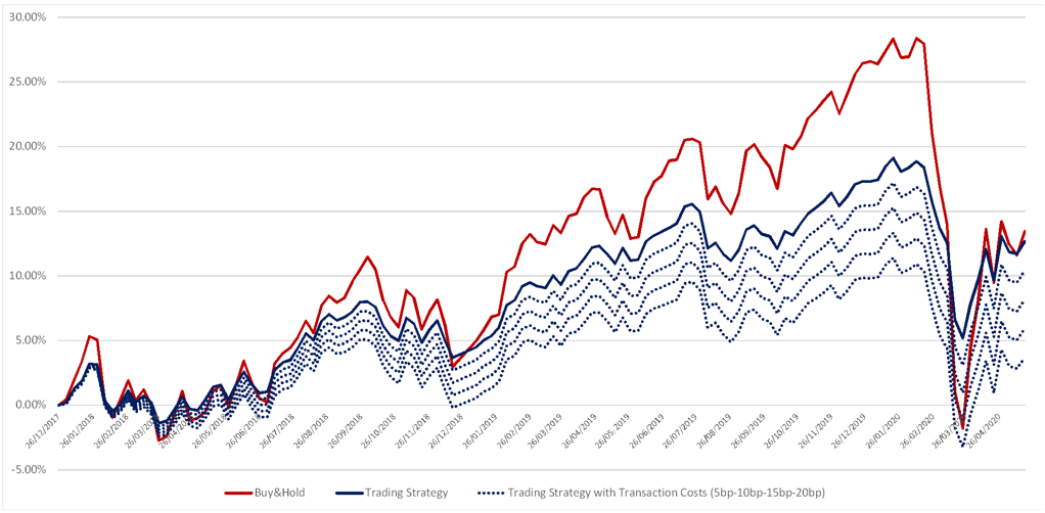

Figure 3: Cumulative return of the trading strategy and of the Buy&Hold strategy where the set of stocks is provided by the DOW JONES Index.

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

Table 4: 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 the last column (break even) refers to the transaction cost that implies the same Sharpe ratio for the trading strategy and the Buy&Hold strategy.

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

Table 5: 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 the last column (break even) refers to the transaction cost that implies the same Sharpe ratio for the trading strategy and the Buy&Hold strategy.

1-c%	**Ex.BA**	**Ex.Ret**	**Ex.TrVo**	**Autoreg.Str**	**Ex.Vola**	**Overall**
N.3 4 5	2.00%	15.06%	1.07%	-	-	18.13%
<90%	1.15%	3.81%	0.97%	-	-	5.93%
>=90%	0.86%	11.25%	0.09%	-	-	12.20%
N.6 8 10	1.54%	11.56%	2.15%	10.31%	-	25.57%
<90%	0.84%	3.17%	1.37%	2.61%	-	7.99%
>=90%	0.70%	8.40%	0.78%	7.70%	-	17.58%
N. 12 26 38	4.23%	9.69%	4.58%	11.65%	-	30.14%
<90%	1.08%	3.27%	0.79%	6.34%	-	11.48%
>=90%	3.15%	6.42%	3.78%	5.31%	-	18.66%
N.52 78 104	4.37%	8.08%	3.55%	8.14%	-	24.14%
<90%	1.15%	2.76%	0.66%	4.53%	-	9.10%
>=90%	3.22%	5.32%	2.89%	3.61%	-	15.04%
A in [0.4;1[	-	-	-	-	2.97%	2.97%
A in [1; 1.6[	-	-	-	-	1.07%	1.07%
Overall	12.15%	44.39%	11.34%	30.10%	4.04%	100.00%

Table 6: 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	1.51%	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%

Table 7: Percentage of selected predictors for stocks belonging to the DOW JONES index.

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 in a very agnostic way 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.

References

[1] Abellań. J. and Castellano. J. G. (2017), Improving the Naive Bayes classifier via a quick variable selection method using maximum of entropy, *Entropy*, 19(6), 247.

[2] Allen, F., Karjalainen, R. (1999) Using genetic algorithms to find technical trading rules, *Journal of Financial Economics*, 51: 245-271.

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

[4] Barucci, E. and Fontana, C. (2017) *Financial Markets Theory: Equilibrium, Efficiency and Information*, Springer.

[5] Bettis, C., Cole, J., Lemmon, M. (2000) Corporate policies restricting trading by insiders. *Journal of Financial Economics*, 57: 191-220.

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

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

[8] Boonamnuay, S., Nittaya, K. and Kittisak, K. (2018) Classification and Regression Tree with Resampling for Classifying Imbalanced Data, *International Journal of Machine Learning and Computing*, 8(4): 336-340.

[9] Brock, W., Lakonishok, J. and LeBaron, B. (1992) Simple Technical Trading Rules and the Stochastic Properties of Stock Returns, *The Journal of Finance*, 47: 1731-1764.

[10] Campbell, J., Lo, A. and MacKinaly, C. (1997) *The Econometrics of Financial Markets*, Princeton University Press.

[11] Cao, C., Field, L.C., Hanka, G. (2004) Does insider trading impair market liquidity? Evidence from IPO lockup expiration, *Journal of Financial and Quantitative Analysis* 39: 25-46.

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

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

[14] Chen, Y. and 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.

[15] Cheng, L., Firth, M., Leung, T. and Rui, O. (2006) The effects of insider trading on liquidity *Pacific-Basin Finance Journal*, 14: 467?483.

[16] 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.

[17] Chung, K.H., Charoenwong, C. (1998) Insider trading and the bid-ask spread, *The Financial Review* 33: 1?20.

[18] Cornell, B. and Sirri, E. (1992) The reaction of investors and stock prices to insider trading, *Journal of Finance*, 47, 3: 1031-1059.

[19] Campbell, J., Grossman, S. e Wang, J. (1993) Trading Volume and Serial Correlation in Stock Returns, *Quarterly Journal of Economics*, 108: 905-939.

[20] Collin-Dufresne, P. and Fos, V. (2015) Do Prices Reveal the Presence of Informed Trading?, *Journal of Finance*, 70, 4: 1555-1582.

[21] Conrad, J., Hameed, A. and Niden, C. (1994) Volume and Autocovariances in Short-Horizon Individual Security Returns, *Journal of Finance*, 49:1305-1329.

[22] Copeland, T. and Galai, D. (1983) Information effects on the bid–ask spread. *Journal of Finance*, 38:1457–1468.

[23] Fama, E. (1970) Efficient capital markets: a review of theory and empirical work, *Journal of Finance*, 25:383–417.

[24] Fang, J., Jacobsen, B. and Qin, Y. (2014) Predictability of the simple technical trading rules: An out-of-sample test, *Review of Financial Economics*, 23: 30-45.

[25] Fisher, T. and Krauss, C. (2018) Deep learning with long short-memory networks for financial market predictions, *European Journal of Operational Research*, 270: 654-669.

[26] 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.

[27] Glosten, L. and Milgrom, P. (1985) Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14:71–100.

[28] Goodhart, C. and O'Hara, M. (1997) High frequency data in financial markets: Issues and applications, *Journal of Empirical Finance*, 4: 73-114.

[29] Grossman, S. (1989) *The Informational Role of Prices*. MIT press, Massachusetts, Boston.

[30] He, H. and Wang, J. (1995) Differential Information and Dynamic Behavior of Stock Trading Volume, *Review of Financial Studies*, 8: 919-972.

[31] Hsu, P., Hsu, Y. and 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.

[32] Hu Y., Feng, B., Zhang, X. Ngai, E., and Liu, M. (2015) Stock trading rule discovery with an evolutionary trend following model, *Expert systems with applications*, 42: 212-222.

[33] Huang, C. (2012) A hybrid stock selection model using genetic algorithms and support vector regression, *Applied soft computing*, 12: 807-818.

[34] Jegadeesh, N., and Titman, S. (1993) Returns to buying winners and selling losers: implications for stock market efficiency, *Journal of Finance*, 56: 699-720.

[35] Jeng, L., Metrick, A. and Zeckhauser, R. (2003) Estimating the returns to insider trading: a performance-evaluation perspective, *The Review of Economics and Statistics*, 85(2): 453?471.

[36] Jurman, G., Riccadonna, S., and Furlanello, C. (2012) A comparison of MCC and CEN error measures in multi-class prediction, *PloS one*, 7(8), e41882.

[37] Kaucic, M. (2010) Investment using evolutionary learning methods and technical rules, *European Journal of Operational Research*, 207: 1717-1727.

[38] Kim, J., Lim, K. and Shamsuddin, A. (2011) Stock return predictability and adaptive markets hypothesis: evidence from century-long US data, *Journal of Empirical Finance*, 18: 868-879.

[39] Kim, O. and Verrecchia, R. (1991) Market Reaction to Anticipated Announcements, *Journal of Financial Economics*, 30: 273-309.

[40] Kononenko, Igor. (1994) Estimating attributes: analysis and extensions of RELIEF, In: *European conference on machine learning*. Springer, Berlin, Heidelberg. 171-182.

[41] Kyle, A. (1985) Continuous Auctions and Insider Trading, *Econometrica*, 53: 1315-1335.

[42] 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.

[43] Lempers, F. B. (1971) Posterior probabilities of alternative linear models. Rotterdam University Press

[44] Llorente, G., Michaely, R., Saar, G. e Wang, J. (2001) Dynamic volume-return relation of individual stocks. NBER Working paper 8312.

[45] Lo, A. (2004) The adaptive markets hypothesis: market efficiency from an evolutionary perspective, *Journal of Portfolio Management*, 30: 15-29.

[46] Lo, A., Mamaysky, H. and Wang, J. (2000) Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation, *The Journal of Finance*, 55: 1704-1765.

[47] Matthews BW. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta (BBA) Protein Struct.*, 405, 2: 44251.

[48] McGorty, F., Gwilym, O. and 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.

[49] Meulbroek, L. (1992) An empirical analysis of illegal insider trading, *Journal of Finance*, 47, 5: 1661-1699.

[50] Mihaljevi, B., Bielza, C., and Larraaga, P. (2020). bnclassify: Learning Bayesian Network Classifiers. *R package version 0.4.5*

[51] Mitchell, T. J., Beauchamp, J. J. (1988). Bayesian variable selection in linear regression (with discussion). *Journal of American Statistical Association* 83: 1023-1036.

[52] Neely, C., Weller, P. and 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.

[53] Neely, C., Weller, P. and Ulrich, J. (2009) The adaptive markets hypothesis: evidence from the foreign exchange market, *Journal of Financial and Quantitative Analysis*, 44: 467-488.

[54] 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.

[55] O'Hara, M. (1995) *Market Microstructure Theory.* Blackwell Press.

[56] 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.

[57] Parkinson, M. (1980) The Extreme Value Method for Estimating the Variance of the Rate of Return, *The Journal of Business*, 53: 61-65.

[58] 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.

[59] Seyhun, H. N. (1986) Insiders? Profits, Costs of Trading and Market Efficiency, *Journal of Financial Economics* 16: 189-212.

[60] Seyhun, H. N. (1992) Why Does Aggregate Insider Trading Predict Future Stock Returns?, *Quarterly Journal of Economics* 107: 1303-1331.

[61] Sullivan, R., Timmermann, A. and White, H. (1999) Data-Snooping, Technical Trading Rule Performance, and the Bootstrap, *The Journal of Finance*, 54: 1647-1691.

[62] Taylor, N. (2014) The rise and fall of technical trading rule success, *Journal of Banking and Finance*, 40: 286-302.

[63] Wang, J. (1994) A Model of Competitive Stock Trading Volume, *Journal of Political Economy*, 102: 127-168.

[64] Wang, W., Li, W., Zhang, N., Liu, K. (2020) Portfolio formation with preselection using deep learning from long-term financial data, *Expert Systems with Applications*, 143: 113042.

A Naive-Bayes classification algorithm

The goal of the classifier 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

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

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

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

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

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

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

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

$$\bar{k} = argmax_{k=1,\ldots,K}\mathcal{P}(Y_k)\prod_{i=1}^{S}\mathcal{P}(X_i|Y_k)$$

Local distributions $\mathcal{P}(X_i|Y_k)$ are specified by parameters $\Theta(X_i, Y_k)$. It is common to assume each local distribution has a 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. $\alpha > 0$ 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 $\frac{1}{r_i}$ ($\alpha >> 0$). In this paper we use the `bnclassify` R package introduced in [50] 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.