Predicting S&P500 Monthly Direction with Informed Machine Learning

David Romain Djoumbissie[1,2](✉) and Philippe Langlais[1]

[1] Department Computer Science and Operational Research, University of Montreal, Montreal, Canada
david.romain.djoumbissie@umontreal.ca, felipe@IRO.UMontreal.CA
[2] Canada Mortgage and Housing Corporation, Montreal, Canada

Abstract. We propose a systematic framework based on a dynamic functional causal graph in order to capture complexity and uncertainty on the financial markets, and then to predict the monthly direction of the S&P500 index. Our results highlight the relevance of (i) using the hierarchical causal graph model instead of modelling directly the S&P500 with its causal drivers (ii) taking into account different types of contexts (short and medium term) through latent variables (iii) using unstructured forward looking data from the Beige Book. The small size of our training data is compensated by the a priori knowledge on financial market. We obtain accuracy and F1-score of 70.9% and 67% compared to 64.1% and 50% for the industry benchmark on a period of over 25 years. By introducing a hierarchical interaction between drivers through a latent context variable, we improve performance of two recent works on same inputs.

Keywords: Financial knowledge representation · Functional causal graph · Prediction & informed machine learning

1 Introduction

Analyzing and predicting the dynamics of financial markets for investment decision-making over a monthly/quarterly horizon is an old challenge both in academy and in the asset management industry. The environment is complex, uncertain and modeling must take into account many factors, including incomplete, noisy and heterogeneous information with almost 80% in unstructured form [1,2].

The crucial parameter in this type of study is the prediction horizon. Indeed, it is strongly linked to the investment objectives/horizon; the paradigm used for modelling and the size of the training data. For the short-term (month/quarter) prediction, the losses recorded during the financial crisis (2008), in addition to all the previous ones have led many people to question the dominant paradigm. The latter is based essentially on a rational assumption and a direct relationship between S&P500 and a few causal drivers. In the literature, the solution for the short-term prediction might be classified into three groups.

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1239, pp. 553–566, 2020.
https://doi.org/10.1007/978-3-030-50153-2_41

The first group of studies are those from [3–7]. The foundations of their approach is based on pure rational argument and passive decision process without prediction. They assume that markets are efficient and it is difficult to predict the S&P500 index or to do better than a random walk. This solution serves as a benchmark in dynamic or active management segments of the industry.

The second group [8–14] proposes a solution based on a direct relationship (supervised algorithm) between S&P500 and a set of features from causal analysis or data mining. In [10,11], the authors found a direct relationship between S&P500 and four causal variables. The features obtained serve as input for a SVM model with innovation on the Kernel function in order to predict the monthly direction of S&P500 over 2006 to 2014. The main weaknesses from this group are: i) The weak predictions in an unstable environment; (ii) an adhoc approach to select the causal variables, the omission of the context and hierarchical interaction between drivers; (iii) The mismatch between the drivers and the prediction horizon.

The third group [13–17] is the most active at the moment and proposes: (i) to use all potential numerical/textual drivers; (ii) a deep architecture for learning representations directly on data. iii) and a prediction through deep supervised algorithms. In [15,16], the authors use NLP and deep learning on daily financial news to predict monthly direction of S&P500 without a priori knowledge. They learn features and predict directly from the data. The findings of this group are encouraging. However, a review we conducted on nearly 60 recent papers, the prediction horizon was less or equal to one day and more than 80% were tested on a very short period (less than 2 years). This prediction horizon (minutes, hours,..) has the advantage of providing a large training sample[1] but resolves a specific type of problem (high frequency transactions on financial markets), which are totally different from the problematic of monthly prediction.

The difference in terms of objectives, investment horizons, as well as the lack of validated studies over longer periods which will reflect the multiple changes in market regimes, make the notion of the state of art somewhat confused. Although there are a few names in the industry known for their ability to do better than the benchmark, recent studies and statistics [18,19] show it is difficult to conclude that one approach dominates the others.

In this paper, we propose a solution for a dynamic decision-making process based on the monthly prediction of the S&P500 index. The investor has an investment horizon of less than one year and uses a dynamic framework which is updated on a monthly basis. This frequency is also that of the publication and update of economic and financial information.

In order to reach our objectives, our contributions are threefold.

– Firstly, we combine our expertise with those of many studies in order to create the structure of a functional causal graph with four levels of the dynamics of the S&P500. Thus, we avoid learning this structure on small size and unstable data. Instead, we learn the distributional representation of latent variables

[1] data collected every second or minute.

(short/medium term context) from an unsupervised method (auto-encoder, similarity, rules). Level 1 includes observable causal variables, then a priori causal functional relationships allow the link with other 3 levels. The latent variables at the last level serve as features for a classifier.

- Secondly, we use unstructured forward looking data (1970–2019) in order to characterize the state of the business cycle. [15,16,20] confirmed the relevance of using unstructured daily data or events on companies published in 8-k form[2]. But the tests are conducted over short periods (24 months) and the aim was not to propose an effective decision-making process.
- Lastly, we perform a systematic validation and comparison with industry's benchmark over 25 years, as well as four sub-periods known as unstable and difficult to predict. We also make some comparison with other studies in the literature, which we formulated as special cases of our solution.

The remainder can be seen into four points: the description of the functional causal graph, the methodology for learning the representation of latent variables, the experiments with empirical results, finally the conclusion and future work.

2 Stock Market Dynamic via a Causal Functional Graph

Predicting S&P500 direction on the monthly horizon is formulated as a binary supervised classification task:

$$y_{t+1}^{S\&P500} = f(V_t) \tag{1}$$

where $y_{t+1}^{S\&P500}$ is the monthly price direction to be predicted (Up/Down), V_t the vector of features characterising the period t, derived from a functional causal graph (Fig. 1 and 2) of the dynamics of the S&P500 and f represents a classifier.

We describe the causal process of the dynamics of S&P500 through a causal functional graph with two essential goals: i) representing causal interactions (direct or hierarchic, linear or non-linear, static or dynamic,..) between short, medium and long term drivers, ii) learning dynamic embedding from temporal interactions between drivers. The a priori graph structure lies on two main source of knowledge. More than 50 years of literature on the financial markets (financial economic theory, behavioral finance, fundamental analysis, market microstructure, technical analysis), and our 15 years of experience in the conception/implementation of solutions for dynamic and tactical asset allocation.

Figure 1 describes different theories and the hierarchical top down interaction between long, medium, short term drivers and the stock market index. Figure 2 is a specific case based on three important medium/short term context (Business cycle, Market regime, Risk aversion). This choice is supported essentially by various works of two nobel prices in economic (Eugene Fama on empirical

[2] broad form used to notify investors in United States public companies of specified events that may be important to shareholders or the United States Securities and Exchange Commission.

analysis of asset prices[3], Daniel Kahneman on behavioral finance[4]) and one of the best portfolio manager of the century, Ray Dalio[5].

At time t, our biggest challenge is to characterize the current market environment (between t−k ...t) with a set of feature derived from Fig. 2 and use it to predict the S&P500 direction for time t+1. We use X_t to denote the realization of variable X at t and $X_{1:t}$ to denote the history of X between the period 1 to t. At time t, we are able to identify where was the market regime between 1..t−k, but we can only estimate the current market regime and we use the k most recent realisations to do. We will sometimes use $X_{1:t-k}$ and $X_{t-k:t}$.

The functional graph of Fig. 2, describe the dynamics of the S&P500. They have four levels and three main component: i) A set of 130 observable causal variables (ex: daily price index of 10 economics sectors); ii) A set of 6 latent context variables (ex: Risk aversion regime of Investors); and iii) 8 functions or algorithms that reflect a direct causal link between the variables (observable or latent). We suggest [21,22] for more details on functional causal graph in finance.

2.1 Graph Level 1: Observable Variables

The level 1 of the graph represents basic inputs organised around 4 groups of observable causal numerical variables and one group of textual information.

Observable Causal Numerical Variables: The variables in light blue (rectangular shape) designate observable numerical variables. All are available on the St. Louis Federal Reserve and Kenneth R. French websites.

S&P500$_{1:t-k}$: A numerical daily variable on the main U.S. equity market. At time t, the history from 1 to t-k (k represents the recent observations for which the regime is not known) serves as an input for ex-post identification algorithm of market regime (f_2, described in Sect. 3). The output of f_2 is an intermediate latent variable that characterizes the regime (bear/bull/range bound) in which the market was in the past (Market_Regime$_{1:t-k}$).

32_Risk_Factors$_{1:t}$: A set of 32 daily numerical variables denoting financial indexes. At time t, the history from 1 to t serves as an input for an unsupervised learning algorithm (f_3). The output is a set of intermediate latent variables characterizing the risk aversion of investors (Risk_Aversion$_t$).

80_Risk_Factors$_{1:t}$: A set of 80 numerical variables designating indices covering all asset classes and sectors of the economy. At time t, the history from 1 to t combined with Num_Repr_Beige_Book$_{1:t}$ (distributional representation of each Beige Book from 1..t) serves as an input for an unsupervised learning algorithm (f_4). The output of this algorithm is a set of intermediate latent variables that characterize the phase of the business cycle (Economic_Cycle$_t$).

[3] https://www.nobelprize.org/prizes/economic-sciences/2013/fama/facts/.

[4] https://www.nobelprize.org/prizes/economic\discretionary{-}{}{}sciences/2002/kahneman/biographical/.

[5] https://en.wikipedia.org/wiki/Ray_Dalio.

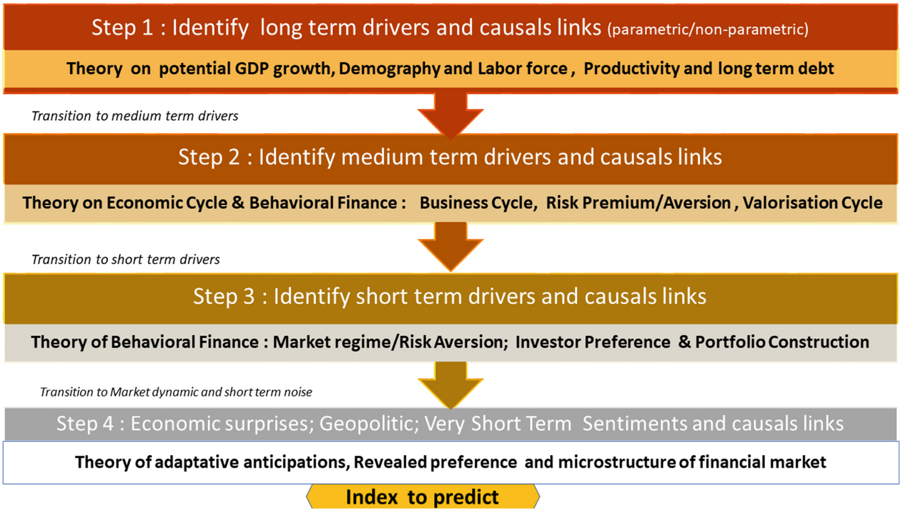
Step 1 : Identify long term drivers and causals links (parametric/non-parametric)
Theory on potential GDP growth, Demography and Labor force , Productivity and long term debt
Transition to medium term drivers
Step 2 : Identify medium term drivers and causals links
Theory on Economic Cycle & Behavioral Finance : Business Cycle, Risk Premium/Aversion , Valorisation Cycle
Transition to short term drivers
Step 3 : Identify short term drivers and causals links
Theory of Behavioral Finance : Market regime/Risk Aversion; Investor Preference & Portfolio Construction
Transition to Market dynamic and short term noise
Step 4 : Economic surprises; Geopolitic; Very Short Term Sentiments and causals links
Theory of adaptative anticipations, Revealed preference and microstructure of financial market
Index to predict

Fig. 1. Main component of the Causal hierachical top down dynamic of any stock index

S&P500_and_Rate$_{1:t}$: A set of the 3 numerical variables designating three of the most relevant indexes on US financial market. At time t, the history covering period 1 to t serves as an input with (Risk_Aversion$_{1:t}$, Economic_Cycle$_{1:t}$) for obtaining features via functions/algorithms or links (f_6, f_7) in the graph.

Observable Textual Causal Variables: It is shown in grey (rectangular shape with rounded side) on the graph (Textual_Data$_{1:t}$). It represent set of textual documents called the Beige Book, published 8 times a year by the U.S. Federal Reserve (≈2000 words for each edition of national summary) on highlights of economic activity, employment and prices. At time t, we use a function/algorithm ($f_1, Doc2vec$) to transform the most recent document into a set of p embedding denoting their distributional representation (Num_Repr_Beige_Book$_t$).

2.2 Graph Level 2: Latent Medium Term Context

Level 2 consists of three groups of light orange (lozenge shape) intermediate latent variables designating medium term context.

Market_Regime$_{1:t-k}$: it is a set of homogeneous cluster on historical price index S&P500$_{1:t-k}$. It summarizes ex-post the state or regime of the financial market for period 1..t-k via the function/algorithm or link f_2 in the graph.

Risk_Aversion$_{1:t}$: it summarizes other medium-term context. The risk aversion of investors on the markets for each period from 1 to t. It is obtained via the function/algorithm or link f_3 on the inputs 32_Risk_Factors$_{1:t}$.

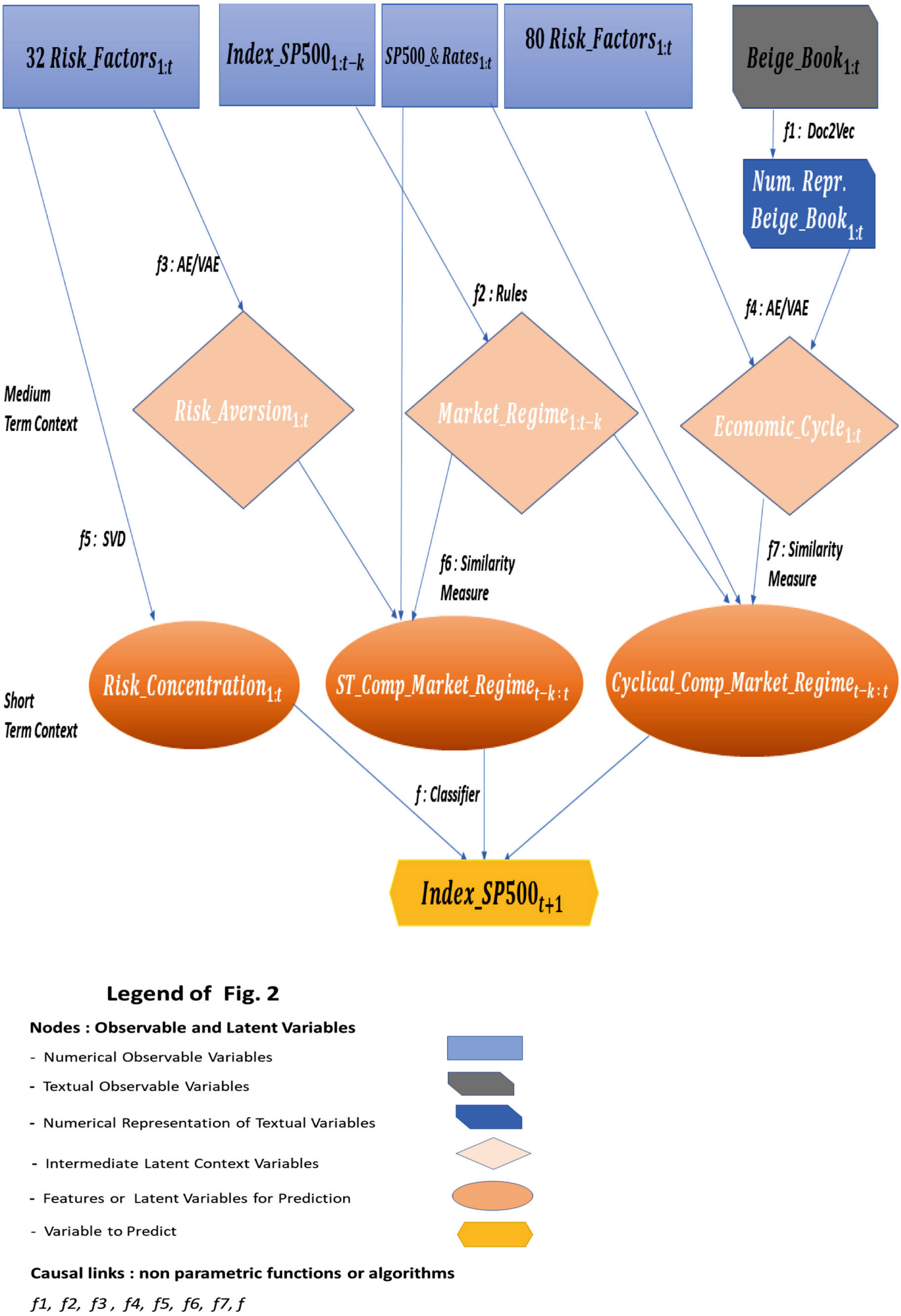
32 Risk_Factors$_{1:t}$
Index_SP500$_{1:t-k}$
SP500_& Rates$_{1:t}$
80 Risk_Factors$_{1:t}$
Beige_Book$_{1:t}$
f1 : Doc2Vec
Num. Repr. Beige_Book$_{1:t}$
f3 : AE/VAE
f2 : Rules
f4 : AE/VAE
Medium Term Context
Risk_Aversion$_{1:t}$
Market_Regime$_{1:t-k}$
Economic_Cycle$_{1:t}$
f5 : SVD
f6 : Similarity Measure
f7 : Similarity Measure
Short Term Context
Risk_Concentration$_{1:t}$
ST_Comp_Market_Regime$_{t-k:t}$
Cyclical_Comp_Market_Regime$_{t-k:t}$
f : Classifier
Index_SP500$_{t+1}$
Legend of Fig. 2
Nodes : Observable and Latent Variables
- Numerical Observable Variables
- Textual Observable Variables
- Numerical Representation of Textual Variables
- Intermediate Latent Context Variables
- Features or Latent Variables for Prediction
- Variable to Predict
Causal links : non parametric functions or algorithms
f1, f2, f3 , f4, f5, f6, f7, f

Fig. 2. Specific case of functional causal graph of the S&P500 dynamic

Economic_Cycle$_{1:t}$: it summarizes the last medium-term context, the phases of the economic cycle for each period 1 to t. It is obtained via the function/algorithm or link f_4 on the inputs (80_Risk_Factors$_{1:t}$ Num. Repr. Beige_Book$_{1:t}$).

2.3 Graph Level 3: Latent Short Term Drivers

Level 3 of the graph includes three other groups of latent variables, dark orange (Oval shape) used as features for a classifier on the S&P500.

ST_Comp_Market_Regime$_{t-k:t}$: A set of 8 latent variables designating the short-term component of the current market regime. For each of the recent periods between t-k to t, we obtain a statistical summary (median and asymmetry coefficient) of k measures of similarity with a function or link (f_6) in the graph. We measure the similarity between the recent k observations (t-k to t) and the historical observations (1 to t-k) of (Risk_Aversion$_{1:t}$, S&P500_and_Rate$_{1:t}$) of each of the homogeneous group obtained with Market_Regime$_{1:t-k}$.

MT_Comp_Market_Regime$_{t-k:t}$: A set of 8 latent variables designating the medium-term component of the current market regime. It is obtained in the same way, replacing the variable (Risk_Aversion$_{1:t}$) with (Economic_Cycle$_{1:t}$).

Risk_Concentration$_{1:t}$: A set of 4 latent variables denoting the concentration of sources of uncertainty in the markets. For each period from 1 to t, the percentage of the explained variance of the first 4 factors is obtained via a singular value decomposition (f_5) on the input 32_Risk_Factors$_{1:t}$.

3 Functions/Algorithms for the Latent Variables

We use a priori knowledge to going through the graph, learn separately the representation of each node and extract 20 business features (level 3) as inputs for a classifier. We validate this process on the reduced graph of Fig. 2 and the task of monthly prediction on S&P500. The generalization with a global graph in a unified embedding learning framework will be for the next step.

3.1 Algorithm f_2 for Intermediate Latent Variables Market_Regime

Market regime is identifiable ex-post. f_2 is a set of rules to separate the history of S&P500 into regimes (3 homogeneous groups). If $SP500_{t_0}...SP500_{t_n}$ is the sequence observed between $t_0, ...t_n$, we define 3 market regimes :
Bull Market: Ex-post, the market was in a bullish mode between period $t_0, ...t_h$ if starting to t_0, the S&P500 rises gradually to cross a certain threshold without returning below the initial price at t_0. Meaning the set of points
$\{t_0.., t_i, ..t_h;\ 0 \leq i \leq h and SP500_{t_i} \geq SP500_{t_0} \,\&\, SP500_{t_h} \geq (1+\lambda) SP500_{t_0}\}$
λ : Hyper-parameter based on empirical studies on risk premium.
Bear Market: Opposite of Bull Market (decrease trend)
Range Bound Market: Neither Bull, neither Bear

3.2 Algorithms f_3, f_4 for 2 Others Intermediate Latent Variables

$f_3\, and\, f_4$ are two auto-encoders to learn the distributional representations of 2 others intermediate latent variables. At time t, a simple/variational auto-encoder (Fig. 3.) takes (32_Risk_Factors$_{1:t}$) as inputs and produces a representation of dimension q (Risk_Aversion$_t$), then takes (80_Risk_Factors$_{1:t}$; Num. Repr. Beige_Book$_{1:t}$) as inputs for other representation of dimension q (Economic_Cycle$_{1:t}$).

3.3 Algorithms f_5, f_6, f_7 and Features

We consider 3 others categories of latent variables to describe the current state of market and use distributional representations as features for a classifier.

8 Statistics Summarizing the Short-Term Component of the Current Market Regime. The current market regime is characterized by the similarity between the recent realisations of (Risk_Aversion$_{1:t}$; S&P500_and_Rate$_{1:t}$) and the historical observations organised in homogeneous groups.

Ex: Consider F_t, the similarity measure (Mahalanobis distance) in date t between recent (last month) observations (V_t^R) and Average/Variance of historical observations in the bullish regime (μ_t^H, S_t^H). We define F_t by :

$$F_t : R^n \times R^n \times R^{n\times n} \longrightarrow R^+$$
$$(V_t^R, \mu_t^H, \tfrac{1}{S_t^H}) = \quad \sqrt{(V_t^R - \mu_t^H)^t \times \tfrac{1}{S_t^H} \times (V_t^R - \mu_t^H)}$$

Therefore, on a monthly horizon (20 days), we obtain a vector of 20 similarity measures that we aggregate by calculating a statistic like the median.

8 Statistics Summarizing the Cyclical Component of the Current Market Regime. They are also obtained by similarity measures as previously but replacing Risk_Aversion$_{1:t}$ by Economic_Cycle$_{1:t}$.

4 Factors Designating the Percentage of Explained Variance, obtained by singular values decomposition (f_5) and explaining more than 90% of the variance of key market risk factors (32_Risk_Factors$_{1:t}$).

These 20 features characterize the current market regime and constitute the main input for a classifier to predict the direction of the S&P500 index.

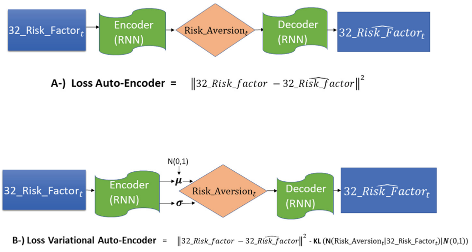
32_Risk_Factor$_t$
Encoder (RNN)
Risk_Aversion$_t$
Decoder (RNN)
32_$\widehat{Risk_Factor}_t$
A-) Loss Auto-Encoder = $\|32\_Risk\_factor - 32\_\widehat{Risk\_factor}\|^2$
N(0,1)
μ
σ
32_Risk_Factor$_t$
Encoder (RNN)
Risk_Aversion$_t$
Decoder (RNN)
32_$\widehat{Risk_Factor}_t$
B-) Loss Variational Auto-Encoder = $\|32\_Risk\_factor - 32\_\widehat{Risk\_factor}\|^2$ - KL (N(Risk_Aversion$_t$|32_Risk_Factor$_t$)|N(0,1))

Fig. 3. Simple & Variational Auto-encodeur

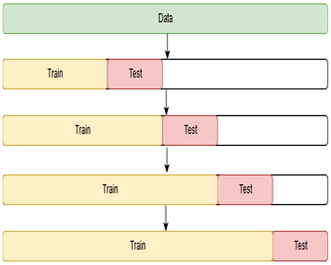

Fig. 4. Walk Forward Validation

4 Experiments and Empirical Results

4.1 Training, Validation and Testing Protocol

Time series have a certain dependence and the chronological order is a crucial element in the training and validation process. Walk Forward Validation (Fig. 4), based on an out-of-time dynamic validation that respects the chronological order, is widely used in finance and [27,28] provides additional details. The first 21 years (19 for training and 2 for validation) are used for initial training and to fix all hyper-parameters. Afterwards, at each date t, we obtain the features and predict the direction $\widehat{y_{t+1}}$ for the month t+1 on training sample $(y^{S\&P500}_{1...t}, V_{1...t})$. At t+1, we compare the prediction with realized $y^{S\&P500}_{t+1}$ and update two models. i) We update the parameters of the auto-encoder every ten years, which is enough to cover various market cycles and, ii) We update the parameters of the classifier every month with all training data until t+1 $(y^{S\&P500}_{1...t+1}, V_{1...t+1})$. We then iterate on the sample from 1992 to 2018 (Fig. 4).

We use three metrics detailed in [13] and adapted for Classification problems. The first is the accuracy (ACC) which is the total percentage of good predictions up and down. The second is the F1-score[6] and the last is the Matthews Correlation Coefficient[7] (MCC). The last two metrics allow a relevant analysis of the cost of errors. Indeed, the cost of bad decisions is high in the markets and the challenge is to have models with good *accuracy*, but especially an ability to limit false positives and false negatives. We make a comparison with the industry benchmark over the entire period, then in sub-periods known to be very unstable and difficult to predict. Our experiments are articulated into four points:

- We compare with the industry benchmark over the test period (1992 to 2018), then four unstable sub-periods (2000-02, 2007-08, 2011-12, 2015-16). This last comparison is typically not conducted in recent works.
- We Analyze the impact of features and latent variables by comparing 3 models of increasing complexity:
  i) **Model 1:** the link is direct between S&P500 and only the observable numerical variables of level 1 (no features, no latent variables). ii) **Model 2:** we consider the features, but they are obtained without two main medium-term context (Risk_Aversion and Cycle_Economic). iii) **Model 3:** we use all component of the graph (Fig. 2) and compare simple/variational auto-encoder to get the latent intermediate variables.
- The model with the textual data is compared to the model obtained only on numerical variables. The textual data from the Beige Book is replaced by traditional numerical Business cycle (Inflation, Industrial production).
- A comparison over the same test period and the same inputs deriving the work from [15] and [10,11] as specific cases of our solution. We identified five recent studies on monthly prediction of the direction of S&P50, then we selected the 2 best recent with available input. We transform the input to introduce a hierarchical interaction via a short term latent context.

[6] Harmonic average Precision and Recall.

[7] correlation coefficient between the observed and predicted binary classification.

4.2 API and Hyper-parameters Selection

We used the gensim implementation of Doc2vec to represent the Beige Book documents into vector of dimension k(hyper-parameter) and TensorFlow/Keras for training auto-encoders. For training the classifier, we used scikit-learn on python 3.6 (SVM, RandomForestClassifier and Ensemble.GradientBoostingClassifier).

We have two categories of Hyper-parameter: i) the first category (number of phases in the business cycle, number of market regimes, dimension of latent variables) are choosing based on our experience and some relative consensus on empirical studies on financial market [23–26]. ii) The second category for auto-encoder and classifier (dimension vector for Beige Book, learning rate, number of estimates, maximum depth of the trees, size of the sub-samples) are chosen to maximize output over the training (1970-89) and validation (1990–1991) period.

4.3 Performance and Comparison with the Industry Benchmark

The Table 1 shows our model consistently outperforms the benchmark over the test period (1992 to 2018) on all metrics. It highlights the limits of accuracy in predicting stock market. Indeed, the ACC of the long-term benchmark is around 64 and more when the markets are stable, but the cost of errors are better represented in the F1-score and MCC. We observe the absolute gain on all metrics with our model. The ACC, F1-Score and MCC are respectively 70.9%, 67%, 0.3 compare to 64.1%, 50% and 0 for the industry benchmark.

During the most unstable periods (2000-02, 2007-08), our model has an ACC of 72.2% and 70% versus 38.8% and 41.6% for the benchmark. The spread is more higher on f1-score (72% and 70% versus 22% and 25%) and MCC (0.52 and 0.54 vs 0 and 0). On the other relative unstable period, our model still outperforms the benchmark in (2011-12) but performs similarly in (2015–2016).

Globally, the benchmark has a good ACC on the long term, but masks the cost of error with poor f1-score and MCC and poor output during the unstable sub-periods.

4.4 Analysis of the Impact of the Short/Medium-Term Latent Context

We compared various auto-encoders, simple versus variational auto-encoder and feed-forward versus recurrent. The recurrent VAE gave us the best output. It was not possible to improve prediction with a convolutional auto-encoder. In order of importance, the 3 main points that Table 2 brings are :

- Overall, using the graph (Model 3) with all observable variables and latent context helps, and clearly outperforms model 1 in all test period and unstable sub-periods.
- Reccurent variational auto-encoder seems overall preferable to AE
- For the unstable periods, we don't have absolute conclusion and we need more investigation.

4.5 Impact of the Unstructured Data

The use of backward looking numerical data is considered as one limitation when analysing the financial markets. We explore and confirm the potential of the Beige Book to contain forward looking information for prediction.

The Table 3 shows the comparison with numerical data traditionally used to analyse the business cycle. On the test period (1992–2018), the ACC, f1-score and MCC of the final model are respectively 70.9%, 67%, 0.3 versus 68.5%, 64% and 0.25 for the model without unstructured data. This trend is the same on two highly unstable sub-periods, (2000-02, 2007-08) and one of the relative unstable sub-periods (2011-12). But it underperforms over the relative unstable sub-period of 2015-16 with statistic of (62.5%, 52%, 0.25) vs (54.1%, 46%, −0.04).

Table 1. Monthly prediction of S&P500 on different test periods

All test Period 1992-2018			
	ACC	**F1-S.**	**MCC**
Bench. Industry	64.1	50	0
This work	**70.9**	**67**	**0.3**
Sub Period 2000-2002			
Bench. Industry	38.8	22	0
This work	**72.2**	**72**	**0.52**
Sub Period 2007-2008			
Bench. Industry	41.6	25	0
This work	**70.8**	**70**	**0.54**
Sub Period 2011-2012			
Bench. Industry	58.3	43	0
This work	**70.8**	**70**	**0.39**
Sub Period 2015-2016			
Bench. Industry	**58.3**	43	0
This work	54.1	**46**	-0.04

Table 2. Impact of latent context

All test Period 1992-2018			
	ACC	**F1-S.**	**MCC**
Model 1	62	52	0
Model 2	65.7	60	0.16
Model 3 AE	68.8	65	0.26
Model 3 VAE	**70.9**	**67**	**0.3**
Sub Period 2000-2002			
Model 1	47.2	43	0.1
Model 2	55.6	56	0.14
Model 3 with AE	69.4	69	0.35
Model 3 with VAE	**72.2**	**72**	**0.52**
Sub Period 2007-2008			
Model 1	41.7	25	0
Model 2_F	54.1	54	0.07
Model 3 with AE	54.1	54	0.13
Model 3 with VAE	**70.8**	**70**	**0.54**
Sub Period 2011-2012			
Model 1	54.2	41	-0.18
Model 2	58.3	43	0
Model 3 with AE	**75**	**74**	**0.48**
Model 3 with VAE	70.8	70	0.39
Sub Period 2015-2016			
Model 1	54.2	41	-0.18
Model 2	**66.7**	**59**	**0.36**
Model 3 with AE	58.3	49	0.05
Model 3 with VAE	54.1	46	-0.04

Table 3. Impact of Textual Data

All test Period 1992-2018			
	ACC	**F1-S.**	**MCC**
Without Textual Data.	68.5	64	0.25
With Textual Data.	**70.9**	**67**	**0.3**
Sub Period 2000-2002			
Without Textual Data.	66.7	67	0.37
With Textual Data.	**72.2**	**72**	**0.52**
Sub Period 2007-2008			
Without Textual Data.	58.3	58	0.20
With Textual Data.	**70.8**	**70**	**0.54**
Sub Period 2011-2012			
Without Textual Data.	66.7	65	0.29
With Textual Data.	**70.8**	**70**	**0.39**
Sub Period 2015-2016			
Without Textual Data.	**62.5**	**52**	**0.25**
With Textual Data.	54.1	46	-0.04

Table 4. Comparison with related work

Ding Model : 2013	
	ACC
Benchmark Industry	75
Ding Model	55.9
This work	**80**
Pena Model : 2006-2014	
	ACC
Benchmark Industry	63.1
Pena Model	69.4
This work	**72**

4.6 Comparison with Two Works on the Same Inputs and Test Period

The metric available for comparison with two recent studies on monthly S&P500 prediction is the accuracy. We use the same inputs by formulating as specific cases of our result. [15] use neural networks on textual data and get a 55.9% accuracy over 12-month test period (Table 4). Although the test period is very short, the industry benchmark is 75% and the special case obtained from our solution is 80%. [10,11] use prior knowledge to select causal variables and innovate in the kernel function of an SVM algorithm. The accuracy is 69.4% for the period 2006–2014 (Table 4). We also obtain a specific case of our solution on the same causal variables with an accuracy of 72%. We show the importance of introducing a hierarchical interaction through a latent variable characterizing the short-term context.

5 Conclusion and Future Work

In this work, we tested several intuitions which should serve as a basis for generalizing of an integrated process with a small training sample of predicting financial markets on a monthly and quarterly basis. This is based on a framework of informed machine learning with an a priori functional causal graph of the S&P500 dynamics as the main input for predictive algorithms.

By combining our market experience, domain literature, we propose an a priori functional causal graph of the market dynamics. We learn separately the representation of each node, and then treat two similar work as special cases of our solution.

The proposed solution reconciles the theory, the selection of causal and context variables with great predictive powers, the domain knowledge features for monthly prediction on the small size of training data. The prediction are better than those of 5 similar works (including 2 studied here) and dominate the industry benchmark in all environments (stable and unstable).

The next step is to generalize using a global, dynamic functional causal graph with multiple unstructured data sources as the main input, then to automatically learn in a unified framework the embedding of all nodes and finally use it to predict the direction of various financial index and horizons (month, quarter,...).

References

1. Vuppala, K.: BlackRock, Text Analytics for Quant investing (2015)
2. Squirro: Use of unstructured data in financial services, White Paper (2014)
3. Fama, E.: Efficient capital market: a review of theory and empirical work. J. Finan. **25**, 383–417 (1970)
4. Sharpe, W.F.: Capital asset prices: a theory of market equilibrium under conditions of risk. J. Finan. **19**, 425–442 (1964)
5. Fama, E., French, K.: Efficient capital markets: II. J. Finan. **46**, 1575–1617 (1991)
6. Fama, E.F., French, K.R.: Common risk factors in the returns on stocks and bonds. J. Finan. Econ. **33**(1), 3–56 (1993)
7. Campbell, J.Y.: Empirical asset pricing: Eugene Fama, Lars Peter Hansen, and Robert Shiller. Working Paper, Department of Economics, Harvard University (2014)
8. Khaidem, L., Saha, S., Dey, S.R.: Predicting the direction of stock market prices using random forest (2016). arXiv:1605.00003
9. Harri, P.: Predicting the direction of US stock markets using industry returns. Empir.Finan. **52**, 1451–1480 (2017). https://doi.org/10.1007/s00181-016-1098-0
10. Peña, M., Arratia, A., Belanche, L.A.: Multivariate dynamic kernels for financial time series forecasting. In: Villa, A.E.P., Masulli, P., Pons Rivero, A.J. (eds.) ICANN 2016. LNCS, vol. 9887, pp. 336–344. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44781-0_40
11. Pena, M., Arratia, A., Belanche, L.: Forecasting financial time series with multivariate dynamic kernels. In: IWANN (2017)
12. Weng, B., Ahmed, M., Megahed, F.: Stock market one-day ahead movement prediction using disparate data sources. Expert Syst. Appl. **79**, 153–163 (2017)
13. Ican, O., Celik, T.: Stock market prediction performance of neural networks: a literature review. Int. J. Econ. Finan. **9**, 100–108 (2017)
14. Gu, S., Kelly, B.T., Xiu, D.: Empirical asset pricing via machine learning. Chicago Booth Research Paper No. 18–0 (2018)
15. Ding, X., Yue, Z., Liu, T., Duan, J.: Using structured events to predict stock price movement: an empirical investigation. Association for Computational Linguistics (2014)

16. Ding, X., Yue, Z., Liu, T., Duan, J.: Deep learning for event-driven stock prediction. In: Intelligence (IJCAI) (2015)
17. Chong, E., Han, C., Park, C.: Deep learning networks for stock market analysis and prediction: methodology, data representations, and case studies. Expert Syst. Appl. **83**, 187–2015 (2017)
18. Kenechukwu, A., Kruttli, M., McCabe, P., Osambela, E., Shin, C.: The shift from active to passive investing: potential risks to financial stability? Working Paper RPA 18–04 (2018)
19. Johnson, B.: Actively vs. Passively Managed Funds Performance. Morningstar Research Services LLC (2019)
20. Lee, H., Surdeanu, M., MacCartney, B., Jurafsky, D.: On the importance of text analysis for stock price prediction. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC (2014)
21. Denev, A.: Probabilistic Graphical Models : A New Way of Thinking in Financial Modelling. Risk Books, London (2015)
22. Denev, A., Papaioannou, A., Angelini, O.: A probabilistic graphical models approach to model interconnectedness. SSRN (2017)
23. Gonzalez, L., Powell, J.G., Shi, J., Wilson, A.: Two centuries of bull and bear market cycles. Int. Rev. Econ. Finan. **14**(4), 469–486 (2005)
24. Andrew, L.: Adaptive Markets, Financial Evolution at the Speed of Thought, Editions (2017)
25. Bry, G., Boschan, C.: Programmed selection of cyclical turning points. In: Cyclical Analysis of Time Series: Selected Procedures and Computer Programs, pp. 7–63 (1971)
26. Cotis, J.-P., Coppel, J.: Business cycle dynamics in OECD countries: evidence, causes and policy implications. In: RBA Annual Conference 2005: The Changing Nature of the Business Cycle Reserve Bank of Australia (2005)
27. Davide, F., Narayana, A.L., Turhan, B.: Preserving order of data when validation defect prediction model. ArXiv (2018)
28. Yao, M., et al.: Understanding hidden memories of recurrent neural networks. In: Conference on Visual Analytics Science and Technology (2017)