Expert Systems With Applications 115 (2019) 256–263

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Algorithmic sign prediction and covariate selection across eleven international stock markets

Markku Karhunen[1]

Discipline of Economics, Department of Political and Economic Studies, University of Helsinki, Arkadiankatu 7, 00014 Finland

ARTICLE INFO

Article history:
Received 4 April 2018
Revised 12 June 2018
Accepted 29 July 2018
Available online 30 July 2018

Keywords:
Stock market indices
S&P 500
Sign prediction
Efficient-market hypothesis
Regularized regression
Similarity-based classification

ABSTRACT

I investigate whether an expert system can be used for profitable long-term asset management. The trading strategy of the expert system needs to be based on market predictions. To this end, I generate binary predictions of the market returns by using statistical and machine-learning algorithms. The methods used include logistic regressions, regularized logistic regressions and similarity-based classification. I test the methods in a contemporary data set involving data from eleven developed markets. Both statistical and economic significance of the results are considered. As an ensemble, the results seem to indicate that there is some degree of mild predictability in the stock markets. Some of the results obtained are highly significant in the economic sense, featuring annualized excess returns of 3.1% (France), 2.9% (Netherlands) and 0.8% (United States). However, statistically significant results are seldom found. Consequently, the results do not completely invalidate the efficient-market hypothesis.

1. Introduction

This paper examines whether it is possible to use expert systems for long-term asset management in the stock markets. The decision rule of such expert systems is trivially simple: Invest in stocks if the stock market is likely to rise and invest in the money market if the stock market is likely to decline. However, to this end, one needs predictions of the market movements. According to mainstream opinion in economics, it is impossible to predict the stock markets, as that would generate an arbitrage opportunity. This view is known as the efficient-market hypothesis (EMH; e.g. Fama, 1991). However, there are other schools of thought. For example, the adaptive-markets hypothesis of Lo (2004) states that individuals use simple heuristics to trade in the stock markets, and consequently, they are not completely rational. This seems to contradict EMH. Moreover, there are theoretical constructions within the discipline of neoclassical economics (e.g. Singleton, 2006, Chapter 9) which show that there can be some degree of predictability in the stock markets, even if the assumptions of EMH are in force. Thus, it is a question of obvious empirical interest if the markets can be predicted or not.

The empirical evidence regarding stock market predictability is mixed. In an influential paper, Welch and Goyal (2008) refuted previous reports of market predictability. The argument was that most authors hitherto had investigated in-sample correlations and the models had no out-of-sample predictive power. Even the in-sample correlations were often lost when the models were updated by new data. Thus, the results could be refuted as statistical artefacts. However, others have challenged the findings of Welch and Goyal (2008). For example, Chevapatrakul (2013) has produced significant out-of-sample predictions for the UK stock market. Similarly, Skabar (2013) and Fiévet and Sornette (2018) have published significant results regarding daily data from the US market. Thus, the debate is ongoing.

In this paper, I use contemporary statistical and machine-learning methods to generate out-of-sample predictions in 11 developed stock markets. The methods considered involve ordinary least squares, logistic regressions, regularized regressions (e.g. Tibshirani, 1996; Zou, 2006) and similarity-based classification (Skabar, 2013). Some authors have reported it to be easier to give a binary prediction of profit or loss than to give an estimate of the expected return (e.g. Leung, Daouk, & Chen, 2000; Nyberg, 2011; Nyberg & Pönkä, 2016). At any rate, it is such sign predictions that the expert system ultimately needs to manage the investment. Thus, I have chosen sign prediction as the objective of this study. I use a combination of statistical tests and trading simulations to assess the potential of the expert system to perform profitable asset management. The rest of this paper is organized as follows. Chapter 2 surveys related work. (The lessons learned from previous work to a large degree guide the modelling choices made in this paper.) Chapter 3 introduces the material and methods. The results are presented in Chapter 4. These are divided in

E-mail address: markku.karhunen@gmail.com
[1] Present address: Medical Bioinformatics Centre, Turku Centre for Biotechnology, Tykistökatu 6A, 20520 Turku, Finland

https://doi.org/10.1016/j.eswa.2018.07.061
0957-4174/

two main categories: Main results (Chapter 4.1) and results obtained from sensitivity analyses (Chapter 4.2). Chapter 5 concludes and presents directions for future work.

2. Related work

There is a large body of literature regarding stock market prediction, both in-sample and out-of-sample. Consequently, it is possible to give only a few references to recent work in this paper. As noted in the Introduction, Welch and Goyal (2008) refuted predictability in the stock markets. Some authors (Fiévet & Sornette, 2018; Lanne, Meitz, & Saikkonen, 2013) have pointed out that predictability in the stock markets is rather non-linear than linear, and consequently, most previous authors have been using methods ill-suited for the problem at hand. On the other hand, many authors have also found linear predictability in the stock markets (e.g. Chevapatrakul, 2013; Nyberg & Pönkä, 2016; Pönkä, 2016). Consequently, I use a combination of linear and non-linear methods in this study. (Linearity here is to be understood in context of the effects of covariates. If the effects of all covariates on the probability of profit are monotone, a model is said to be 'linear'.)

A number of factors are believed to affect the stock markets, thus allowing for profitable prediction. These involve market volatility (Chevapatrakul, 2013), oil price (Gupta & Wohar, 2017; Liu, Ma, & Wang, 2015; Pönkä, 2016) and the lags of the US stock market return (Narayan, Phan, & Narayan, 2018; Nyberg & Pönkä, 2016). Dividend yield, interest rate, industrial production growth and exchange rate growth have also been suggested as potential predictors (Ang & Bekaert, 2007; Rapach, Strauss, & Zhou, 2013), among others. The factors affecting the stock market are believed to be country-specific (Hadhri & Ftiti, 2017). Thus, one would like to have a data set sufficiently rich in covariates to attempt stock market prediction. One such data set is offered by the monthly data of Rapach et al (2013), also analyzed by Nyberg and Pönkä (2016) and Pönkä (2016). These data are also adopted for this paper.

Regarding sign prediction in the stock markets, any method suited for binary classification may be used. A natural starting point are the logit and probit regressions (e.g. Leung et al., 2000). In the machine-learning literature, decision trees are well known, and they have also been applied to stock market prediction (e.g. Fiévet & Sornette, 2018). Same applies to artificial neural networks (Zhong & Enke, 2017a,b) and linear discriminant analysis (Leung et al., 2000). More exotic methods used in this domain include fuzzy robust principal component analysis (Zhong & Enke, 2017a), copulas (Anatolyev & Gospodinov, 2010), empirical mode decomposition (Pan & Hu, 2016) and similarity-based classification (Skabar, 2013) which is also used in this paper.

Previous literature offers a number of suggestions regarding the methods used for testing the expert system. Firstly, Welch and Goyal (2008) stress the importance of out-of-sample testing. Secondly, trading simulations carried out by using historical data are central to the credibility of the algorithm (Nyberg, 2011). Ideally, the trading strategy obtained from a predictive model is robust towards substantial trading costs. It may also be desirable to test the sensitivity of the results towards other assumptions (cf. Narayan et al., 2018). However, strategies may often appear economically profitable, even if the predictions are not statistically significant (Nyberg, 2011). Consequently, it is advisable to test the binary predictions by contrasting them against the reality, e.g. by using the Pesaran–Timmermann test (2009). As a result of these considerations, this paper features trading simulations, statistical significance tests and excessive sensitivity analyses. All of these are performed out of sample, by using rolling windows. Additionally, the predictive accuracy of the methods is compared and a ranking of the methods is attempted. This comparison is based on the model-confidence set algorithm of Hansen, Lunde, and Nason (2011) which tests all the models against each other.

3. Material and methods

Let us consider a binary variable y_t and a vector of continuous and binary covariates $\boldsymbol{x}_{t-1}$. The subscript $t-1$ refers to the fact that the covariates are observed in the previous period. The problem is to predict y_t from $\boldsymbol{x}_{t-1}$. In this paper, y_t is a monthly indicator of profit or loss and $\boldsymbol{x}_{t-1}$ involves macroeconomic variables. I discuss two types of predictions, binary ($\hat{y}_{mt} = 0, 1$) and continuous ($p_{mt} \in \mathbf{R}$) where $m = 1, \ldots, 9$ denotes the predictive model. The continuous predictions are called probability scores and are model-based estimates of $E(y_t|\boldsymbol{x}_{t-1})$. The binary predictions are calculated by truncating the probability scores.

Perhaps the most basic predictive method is ordinary least squares (OLS), also known as the linear probability model in this type of setting (Cameron & Trivedi, 2005, Chapter 14). In certain cases, it is possible that OLS predicts $p_{mt} < 0$ or $p_{mt} > 1$, but even if this occurs, OLS may give better binary predictions than other, more sophisticated models. Consequently, OLS is used as a baseline model in this study.

3.1. Logistic regression

Logistic regression is a type of generalized linear model (McCullagh & Nelder, 1989). It is used extensively in many fields of science to model binary data. It is based on the logistic link function

$$\Lambda(x) = \frac{e^x}{e^x + 1}, \quad x \in \mathbf{R}. \tag{1}$$

Using this link function, logistic regression can be defined as

$$(y_t|\boldsymbol{x}_{t-1}) \sim B\left(\Lambda\left(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_{t-1}\right)\right) \tag{2}$$

where B denotes a Bernoulli distribution. This implies the following log-likelihood function

$$\ell\left(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}\right) = \sum_{t=1}^{T} y_t \log \Lambda\left(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_{t-1}\right) + \sum_{t=1}^{T} (1 - y_t) \log\left(1 - \Lambda\left(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_{t-1}\right)\right) \tag{3}$$

where $\mathbf{X}$ is the matrix of covariates, $\boldsymbol{\beta}$ is the vector of their regression coefficients and β_0 is an intercept. Another popular choice is to use the probit link function in place of Λ. This yields the probit regression model. In this paper, logit regression is taken as a starting point, in line with Anatolyev and Gospodinov (2010) and Chevapatrakul (2013). With a suitable scaling of regression coefficients, the two link functions are virtually indistinguishable (Cameron & Trivedi, 2005, Chapter 14).

Logistic regression if fitted by maximizing (3) over β_0 and $\boldsymbol{\beta}$. If all covariates are included in $\mathbf{X}$, this yields a full model which is prone to overfitting. To this end, one usually performs some sort of model choice. In this paper, I use logistic regression in combination with Akaike's information criterion (AIC; e.g. Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978). I perform stepwise search to minimize AIC and BIC. The idea of this algorithm is that it starts with the full model and then removes and adds covariates one by one, until it reaches a local minimum of AIC or BIC. To summarize, there are three variants of logistic regression in this paper: full model (henceforth, FM) and models chosen by AIC and BIC.