Finance Research Letters 22 (2017) 211–226

Contents lists available at ScienceDirect

Finance Research Letters

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

Can tree-structured classifiers add value to the investor?[☆]

Ricardo Laborda[a,*], Juan Laborda[b,c]

[a] Centro Universitario de la Defensa de Zaragoza, Academia General Militar, Ctra. Huesca s/n, 50090 Zaragoza, Spain
[b] Syracuse University, C/Miguel Ángel 8, 28010 Madrid, Spain
[c] Universidad Carlos III Madrid, C/ Madrid, 126, 28903 Getafe (Madrid), Spain

ARTICLE INFO

Article history:
Received 23 December 2016
Revised 17 May 2017
Accepted 3 June 2017
Available online 6 June 2017

JEL classification:
G11
G19

Keywords:
Market timing
Tree-structured classifiers
State variables
Performance evaluation

ABSTRACT

We analyse the investor welfare gain of including tree-structured classifiers' predictions about the relative performance of stock vs. cash. The CART, bagging, and random forest methods select the VIX level and momentum, the earning bond yield level and momentum, and the detrended risk-free rate as the most important state variables to predict the outperformance of the S&P 500 vs. cash out-of-sample. These tree-structured classifiers' predictions are used as a binary state variable to estimate optimal investor portfolios that also deliver out-of-sample higher Sharpe ratios and certainty equivalent return gains than competing portfolio strategies that exclude them.

1. Introduction

The upsurge of passive investing is a clear trend in financial markets that reinforces the concerns about active fund managers. Fund managers may not charge fees according to their added value (Gil-Bazo and Verdú, 2009), and there is mixed evidence of the persistence of funds' performance (see Carhart, 1997; Ferreira et al., 2013; Hendricks et al., 1993; Kacperczyk et al., 2014). Therefore, there is no clear evidence of the existence of skilled or informed mutual fund managers.

The classical approach to investors' asset allocation problem requires the modelling of the underlying assets' return distribution and then the plugging of the estimated parameter values into an analytical or numerical solution to the problem defined over an objective function. There is an enormous literature on return predictability that tries to determine the factors that anticipate the dynamics of the underlying assets' return distribution,[1] which can be used to allocate investors' wealth optimally to different assets depending on the perceived signal of their relative attractiveness. However, the empirical evidence also shows a clear difficulty in delivering consistently superior out-of-sample forecasts of the associated models relative to a simple forecast based on the historical average (Goyal and Welch, 2008); consequently, these models may not help investors to time the market. The existence of non-linearities and structural breaks in the relationship between the

☆ The financial support from Gobierno de Aragón and Fondo Europeo de Desarrollo Regional (CREVALOR) is gratefully acknowledged.

* Corresponding author.
E-mail address: rlaborda@unizar.es (R. Laborda).

[1] See, among others, Ang and Bekaert (2007), Baker and Wurgler (2006), Campbell (1991), Cochrane and Piazzesi (2005), Fama and French (1988, 1989), Fama and Schwert (1977), Ferreira and Santa Clara (2011), Lettau and Ludvigson (2001), Ludvigson and Ng (2009), and Maio (2013).

http://dx.doi.org/10.1016/j.frl.2017.06.002
1544-6123/

state variables and the asset returns (Fransens and Dijk, 2000) as well as the stylized behaviour of financial asset returns, such as negative skewness, fat tails, and volatility clustering (Cont, 2001), make the activity of investing a delicate task.

In this paper we study the ability of tree-structured classifiers and regression tree classifiers (CART hereafter) to help investors to engage a profitable investment strategy in-sample and especially out-of-sample. The high level of complexity of financial markets makes the use of these classifiers particularly suitable, as they can handle problems characterized by high dimensionality, non-normality, and non-homogeneity, that is, different relationships between variables in different parts of the measurement space (Breiman et al., 1984). Thus, the tree-structured classification not only is used to produce an accurate classifier but also can be used to identify the most important variables and their interactions or to uncover the predictive structure of any problem, especially in a non-linear context. Therefore, we can summarize the advantages of the CART over other existing methodologies for forecasting, especially the linear parametric ones, as follows: (1) it is a non-parametric technique that does not require specification of the underlying distribution of the variables being modelled, (2) it does not require variables to be selected in advance, as the CART algorithm will identify the most significant variables and eliminate the non-significant ones itself,[2] (3) the CART can easily handle outliers, noisy data, and especially non-normality that is related to the negative skewness and fat tails of financial asset return distribution, and 4) CART results are intuitive and very easily interpreted, especially compared with other plausible non-parametric techniques such as the neural networks considered as 'black-box' methods.[3] The main disadvantages of the CART methodology are that it may have unstable decision trees and that insignificant modifications of the learning sample could lead to radical changes in splitting variables and values. To address these problems, we explore more complex alternatives to the CART approach by growing an ensemble of trees and letting them vote for the most popular class. We consider the bagging (Breiman, 1996a, 1996b) and random-forest (Breiman, 2001) methodologies to try to increase the accuracy of the resulting predictions and to handle the instability issue. Due to the increasing complexity of the proposed methods, we try to assess their relevance to our asset allocation context particularly compared with simpler strategies, such as the linear parametric or the passive out-of-sample strategy.

Tree-structured classifiers are used especially in the credit risk literature to analyse different problems, such as the classification of financially distressed firms (Frydman et al., 1985) or corporate credit rating changes (Jones et al., 2015), and to a lesser extent in the asset allocation context (see Sorensen et al., 2000).

An important contribution of our paper is to show how to integrate the CART prediction easily into an asset allocation problem by considering a new binary state variable that takes the value one if the S&P 500 outperforms the cash and zero otherwise. Our methodology involves two steps. First, we use the classification and regression tree classifier (CART hereafter) to establish the relationship between our binary dependent variable, which takes the value one if the S&P 500 outperforms the cash and zero otherwise, and a list of potential predictors of the excess stock return. Second, we use the CART prediction in a portfolio context that consists of an investor who optimally invests her/his wealth among stocks and the Treasury bill rate. We consider the CART prediction as a new binary state variable that tries to capture the attractiveness of the stock market vs. cash. We try to uncover the usefulness of the information embedded in the CART predictions to add value to the investor. For comparison we compare the ability of the CART predictions to add value to the optimal portfolio formed by investors who only consider the same state variables selected by the tree-structured classifier. To obtain the set of optimal weights, we operate in an optimal asset allocation setting that follows the literature (see Ait-Sahalia and Brandt, 2001; Brandt, 1999; Brandt et al., 2009; and more recently Barroso and Santa-Clara, 2015). We directly estimate the linear relationship between the state variables and the optimal portfolio allocation to the stock market and consider a CRRA utility function. We focus on the added value to the investors produced by the tree-structured classifiers, using different metrics to measure the portfolios' economic performance, such as the main descriptive statistics of the optimal excess portfolio return distribution, the annualized Sharpe ratio, and the certainty equivalent return gain.

From a practical standpoint, our paper provides several useful insights. Firstly, the proposed methodology based on the new binary state variable provided by the CART adds value out-of-sample to the investor, who could operate alternatively using the passive strategy or the linear parametric portfolio rules based on the isolated state variables. Therefore, our binary state variable would subsume the information embedded in the relevant state variables considered in a linear portfolio rule, as it seems to capture adequately the non-linear relationship between them. Secondly, the consideration of bagging and the random forest does not necessarily lead to an increase in the investor welfare. Our out-of-sample results show that the accuracy of the predictions related to the outperformance of the stock market vs. cash based on the binary classifiers is about a remarkable 70%. In our out-of-sample period, neither the bagging nor the random forest method can increase significantly the accuracy attained by the CART predictions, despite their greater complexity. We interpret this result as a signal of the CART's outcome stability. Obviously, this result depends on our sample and is confirmed by the fact that the CART algorithm mainly chooses the same state variables to predict the excess stock return out-of-sample. The bagging and random-forest variable importance measures' analysis out-of-sample also consistently leads to a ranking of predictors that is similar to the CART algorithm, which we interpret as an indicator of stability. Therefore, our empirical analysis reveals that the consideration of more sophisticated models does not always lead to investor welfare gains. Thirdly, all the tree-structured classifiers emphasize the importance of the VIX level and momentum, the earning yield bond yield (EYBY

[2] Alternatively, it could be considered a shortcoming not to be able to force the use of variables in the model.

[3] Zhang et al. (1998) outline different limitations of artificial neural networks: they are black-box methods with no explicit relationship form to explain and analyse the relationship between inputs and outputs, prone to overfitting problems, and time consuming.