ISSN: 2366-1313

DOW JONES WEEKLY RETURNS USING MACHINE LEARNING WITH DATA ANALYSIS

B SUNIL KUMAR [1], S UDAYA LAKSHMI [2]
[1]Assistant professor, [2] Assistant professor
Department of Computer Science Engineering
CSE Department, Sri *Mittapalli College* of *Engineering*, *Guntur*, *Andhra Pradesh*-522233

Abstract- We employ a semi-parametric method known as Boosted Regression Trees (BRT) to forecast stock returns and volatility at the monthly frequency. BRT is a statistical method that gen- erates forecasts on the basis of large sets of conditioning information without imposing strong parametric assumptions such as linearity or monotonicity. It applies soft weighting functions to the predictor variables and performs a type of model averaging that increases the stability of the forecasts and therefore protects it against overfitting. Our results indicate that expanding the conditioning information set results in greater out-of-sample predictive accuracy compared to the standard models proposed in the literature and that the forecasts generate profitable portfolio allocations even when market frictions are considered. By working directly with the mean-variance investor's conditional Euler equation we also characterize semi-parametrically the relation between the various covariates constituting the conditioning information set and the investor's optimal portfolio weights. Our results suggest that the relation between predictor vari- ables and the optimal portfolio allocation to risky assets is highly non-linear.

Keywords: Equity Premium Prediction, Volatility Forecasting, GARCH, MIDAS, Boosted Regression Trees, Mean-Variance Investor, Portfolio Allocation.

1 Introduction

Information plays a central role in modern finance. Investors are exposed to an ever-increasing amount of new facts, data and statistics every minute of the day. Assessing the predictabil- ity of stock returns requires formulating equity premium forecasts on the basis of large sets of conditioning information, but conventional statistical methods fail in such circumstances. Non-parametric methods face the so-called "curse-of-dimensionality". Parametric methods are often unduly restrictive in terms of functional form specification and are subject to data overfitting concerns as the number of parameters estimated increases. The common practice is to use lin- ear models and reduce the dimensionality of the forecasting problem by way of model selection and/or data reduction techniques. But these methods exclude large portions of the conditioning information set and therefore potentially reduce the accuracy of the forecasts. To overcome these limitations we employ a novel semi-parametric statistical method known as Boosted Regression Trees (BRT). BRT generates forecasts on the basis of large sets of conditioning variables without imposing strong parametric assumptions such as linearity or monotonicity. It does not overfit because it performs a type of model combination that features elements such as shrinkage and subsampling. Our forecasts outperform those generated by established benchmark models in terms of both mean squared error and directional accuracy. They also generate profitable port-

ISSN: 2366-1313

folio allocations for mean-variance investors even when market frictions are accounted for. Our analysis also shows that the relation between the predictor variables constituting the conditioning information set and the investors' optimal portfolio allocation to risky assets is, in most cases, non-linear and non-monotonic.

Our paper contributes to the long-standing literature assessing the predictability of stock returns. Over the nineties and the beginning of the twenty-first century the combination of longer time-series and greater statistical sophistication have spurred a large number of attempts to add evidence for or against the predictability of asset returns and volatility. In-sample statistical tests show a high degree of predictability for a number of variables: Roze↵ (1984), Fama and French (1988), Campbell and Shiller (1988a,b), Kothari and Shanken (1997) and Ponti↵ and Schall (1998) find that valuation ratios predict stock returns, particularly so at long horizons; Fama and Schwert (1977), Keim and Stambaugh (1986), Campbell (1987), Fama and French (1989), Hodrick (1992) show that short and long-term treasury and corporate bonds explain variations in stock returns; Lamont (1998), Baker and Wurgler (2000) show that variables re- lated to aggregate corporate payout and financing activity are useful predictors as well. While these results are generally encouraging, there are a number of doubts regarding their accuracy as most of the regressors considered are very persistent, making statistical inference less than straightforward; see, for example, Nelson and Kim (1993), Stambaugh (1999), Campbell and Yogo (2006) and Lewellen, Nagel, and Shanken (2010). Furthermore, data snooping may be a source of concern if researchers are testing for many di↵erent model specifications and report only the statistically significant ones; see, for example, Lo and MacKinlay (1990), Bossaerts and
Hillion (1999) and Sullivan, Timmermann, and White (1999). While it is sometimes possible to correct for specific biases, no procedure can o↵er full resolution of the shortcomings that a↵ect the in-sample estimates.

Due to the limitations associated with in-sample analyses, a growing body of literature has argued that out-of-sample tests should be employed instead; see, for example, Pesaran and Timmermann (1995, 2000), Bossaerts and Hillion (1999), Marquering and Verbeek (2005), Campbell and Thompson (2008), Goyal and Welch (2003) and Welch and Goyal (2008). There are at least two reasons why out-of-sample results may be preferable to in-sample ones. The first is that even though data snooping biases can be present in out-of-sample tests, they are much less severe than their in-sample counterparts. The second is that out-of-sample tests facilitate the assessment of whether return predictability could be exploited by investors in real time, therefore providing a natural setup to assess the economic value of predictability.

The results arising from the out-of-sample studies are mixed and depend heavily on the model specification and the conditioning variables employed.[1] In particular, many of the studies conducted so far are characterized by one or more of these limitations. First, the forecasts are generally formulated using simple linear regressions. The choice is dictated by simplicity and the implicit belief that common functional relations can be approximated reasonably well by linear ones.[2] Most asset pricing theories underlying the empirical tests, however, do not imply linear relationships between the equity premium and the predictor variables, raising the issue whether the mis-specification implied by linear regressions is economically large. Second, linear mod- els overfit the training dataset and generalize poorly out-of-sample as the number of regressors

ISSN: 2366-1313

increases, so parsimonious models need to be employed at the risk of discarding valuable conditioning information. Approaching the forecasting exercise by way of standard non-parametric or semi-parametric methods is generally not a viable option because these methods encounter "curse-of-dimensionality" problems rather quickly as the size of the conditioning information set increases. Third, the models tested are generally constant: di↵erent model specifications are proposed and their performance is assessed ex-post. Although interesting from an econometric perspective, these findings are of little help for an investor interested in exploiting the condition-

[1]The data frequency also a↵ects the results. Stock returns are found to be more predictable at quarterly, annual or longer horizons, while returns at the monthly frequency are generally considered the most challenging to predict.

[2]Another reason underlying the use of linear frameworks is that those statistical techniques were known by investors since the beginning of the twentieth century. For this and other issues related to "real-time" forecasts, see Pesaran and Timmermann (2005). ing information in real time as he would not know what model to choose ex-ante.[3] Finally, apart from some important exceptions, much of the literature on financial markets prediction focuses on formulating return forecasts and little attention is dedicated to analyzing quantitatively the economic value associated with them for a representative investor

While conditional returns are a key element needed by risk-averse investors to formulate asset allocations, the conditional second moments of the return distribution are crucial as well. In fact, they are the only two pieces of information required by a mean-variance investor to formulate optimal portfolio allocations. It is widely known that stock market volatility is predictable and a number of studies attempts to identify which macroeconomic and financial time-series can improve volatility forecasts at the monthly or longer horizons.[5] But it is still unclear whether that conditioning information could have been incorporated in real-time and how much an investor would have benefitted from it.

In this paper we consider a representative mean-variance investor that exploits publicly avail- able information to formulate excess returns and volatility forecasts using Boosted Regression Trees (BRT). BRT finds its origin in the machine learning literature, it has been studied exten- sively in the statistical literature and has been employed in the field of financial economics by Rossi and Timmermann (2010) to study the relation between risk and return. The appeal of this method lies in its forecasting accuracy as well as its ability to handle high dimensional fore- casting problems without overfitting. These features are particularly desirable in this context, because they allow us to condition our forecasts on all the major conditioning variables that have been considered so far in the literature, guaranteeing that our analysis is virtually free of data-snooping biases. BRT also provide a natural framework to assess the relative importance of the various predictors at forecasting excess returns and volatility. Finally, the method allows for semi-parametric estimates of the functional form linking predictor and predicted variables, giving important insights on the limitations of linear regression.

Our analysis answers three questions. The first is whether macroeconomic and financial variables contain information about expected stock returns and volatility that can be exploited in real time by a mean-variance investor. For stock returns we use the major conditioning variables proposed so far in the literature and summarized by Welch and Goyal (2008). We propose two models of volatility forecasts. The first models volatility as a function of monthly macroeconomic and financial time-series as well as past volatility. The second is inspired by the family of MIDAS models proposed by Ghysels, Santa-Clara, and Valkanov (2006) and models monthly volatility as a function of lagged daily squared returns. We call this model "semi-parametric MIDAS" and show that its performance is superior to that of its parametric counterpart. Genuine out- of-sample forecasts require not only that the parameters are estimated recursively, but also that the conditioning information employed is selected in real-time. For this reason,

ISSN: 2366-1313

every predictive framework under consideration starts from the large set of predictor variables employed by Welch and Goyal (2008) and selects recursively the model specification. Our estimates show that BRT forecasts outperform the established benchmarks and possess significant market timing in both returns and volatility.

A related question we address is whether the conditioning information contained in macro and financial time-series can be exploited to select the optimal portfolio weights directly, as proposed by Ait-Sahalia and Brandt (2001). Rather than forecasting stock returns and volatility separately and computing optimal portfolio allocations in two separate steps, we model directly the optimal portfolio allocation as a target variable. Our approach can be interpreted as the semi-parametric counterpart of Ait-Sahalia and Brandt (2001),[6] because instead of reducing the dimensionality of the problem faced by the investor using a single index model, we employ a semi-parametric method that avoids the so-called "curse of dimensionality". Our analysis gives rise to two findings. First, formal tests of portfolio allocation predictability show that optimal portfolio weights are time-varying and forecastable; second, we show that the relation between the predictor variables constituting the conditioning information set and the mean-variance investor's optimal portfolio allocation to risky assets is highly non-linear. The third question we analyze is whether the generated forecasts are economically valuable in terms of the profitability of the portfolio allocations they imply. We assess this by com- puting excess returns, Sharpe ratios and Treynor-Mazuy market timing tests for the competing investment strategies. Our results highlight that BRT forecasts translate into profitable portfolio allocations. We also compute the realized utilities and the break-even monthly portfolio fees that a representative agent would be willing to pay to have his wealth invested through the strategies we propose, compared to the benchmark of placing 100% of his wealth in the market portfolio.

1.1 Mean-Variance Investor and Portfolio Allocation

We construct portfolio weights for the mean-variance investor described in Section 2 who uses boosted regression trees to formulate return and volatility predictions.[16]

We have a free parameter ц that determines the investor's degree of risk-aversion. We set it to 4 as that corresponds to a moderately risk-averse investor. We first present the results for unconstrained portfolio weights and evaluate the performance of two models. In the first, market returns and volatility are forecasted separately and the optimal mean-variance portfolio allocation is computed in a second stage (two-step BRT model). In the second, the optimal portfolio allocation is forecasted directly following the procedure described in Section 3 (one-step BRT model).

We compare our results to three benchmark strategies. The first uses stock returns predictions from a multivariate linear regression model selected recursively using the BIC on the full set of predictor variables listed in Section 2 and volatility predictions from a GARCH (1,1) model. The second model replaces the GARCH (1,1) model with a MIDAS (Beta) model (Ghysels et al (2005)). The third benchmark is a passive investment strategy that allocates 100% of wealth in the market portfolio at all times.

In Panel A of Table 3 we report the mean and standard deviation of returns for each investment strategy. We also report the Sharpe Ratio, the Jensen's Alpha measure of abnormal returns, i.e. the coefficient and t-statistic for the ↵ 0 coefficient in the regression

ISSN: 2366-1313

$$r_{p,t+1} = \alpha_0 + \alpha_1 r_{t+1} + \varepsilon_{t+1}, \quad (27)$$

where $r_{p,t+1}$ is the return on the investment portfolio and r_{t+1} is the return on the market portfolio. Finally, we report the Treynor-Mazuy (TM) test statistic, i.e. the coefficient and t-statistic for the α_2 coefficient in the regression rmance is evaluated over the sample 1969-2008. The best model is the BRT that forecasts the optimal portfolio allocation in two steps with a monthly Sharpe ratio of 15.81%. The second best is the one-step BRT model with a Sharpe ratio of 14.81%. The two-step BRT model has a higher and more significant Jensen's Alpha measure of market outperformance, but a lower and less significant Treynor-Mazuy (TM) measure of market timing compared to the one-step BRT benchmark strategies have a negative and insignificant Jensen's Alpha: -0.072% for the model that uses a GARCH (1,1) and -0.069% for the model that uses a MIDAS (Beta) specification. The Treynor-Mazuy measure of market timing is positive but insignificant for both models. The Sharpe ratios, 2.70% and 3.18% respectively, are smaller than that of the passive strategy (Panel D) that invests 100% of wealth in the market portfolio (7.71%).

In Panel B of Table 3 we repeat the same exercise imposing short-selling and borrowing constraints. As expected, the constraints reduce the profitability of both BRT-based investment strategies. The Sharpe ratio of the two-step BRT model drops by approximately 1.8% to 14.01%, while the Sharpe ratio of the one-step BRT model drops by approximately 2.2% to 12.59%. The lower Sharpe ratios are due to both lower mean returns and volatility of the constrained portfolio allocations compared to the unconstrained ones. The magnitudes of both Jensen's Alpha's and TM market timing measures are lower than their unconstrained counterparts, but they maintain a strong level of statistical significance. The opposite holds true for the two active benchmark strategies. The Sharpe ratio for the investment strategy that exploits GARCH (1,1) volatility forecasts increases by 2.8% to 5.52% and the one for the MIDAS (Beta) increases by 3.5% to 6.69%. There are a number of reasons for this, the most important being that there is a fair degree of estimation error in the estimated optimal portfolio allocations, particularly so for the linear regression, GARCH (1,1) and MIDAS (Beta) models that minimize L^2 criterion functions. We show in section 4.7 that accounting for estimation uncertainty improves the performance of the investment strategies.

Table 3 Panel C reports results for models that exploit only the predictability of stock returns: i.e. the investment strategy entails investing 100% of wealth in the risky asset if expected excess returns are greater than zero and 0% otherwise. For the one-step BRT model the decision is based on the sign of the expected Sharpe ratio rather than that of expected returns. For BRT, the portfolio allocations that do not exploit volatility predictions are less profitable than those that do. The Sharpe ratio of the two-step BRT model drops by 1.2% to 12.83% and the one for the one-step BRT model drops by 1.4% to 11.18%. The results for the BRT models indicate that the investment strategies that explicitly model volatility lead to more profitable portfolio allocations. The opposite holds true the two active benchmark strategies as the Sharpe ratio of the linear model that does not exploit volatility forecasts is 8.09%, higher than the 5.52% and 6.69% obtained by the models that use GARCH (1,1) and MIDAS (Beta) volatility forecasts, respectively.

Overall, this section highlights that BRT models outperform the established benchmarks in

ISSN: 2366-1313

term of constrained and unconstrained portfolio allocation performance. They also show that, for BRT models, portfolio allocations based on both conditional returns and volatility estimates are

ISSN: 2366-1313

more profitable than those that exploit return predictability only. Next, we focus our attention on BRT models with transaction costs and show that the outperformance of our framework is robust to the inclusion of such frictions.

2 Conclusions

We present new evidence on the predictability of stock returns and volatility at the monthly frequency using Boosted Regression Trees (BRT). BRT is a novel semi-parametric statistical method that generates forecasts on the basis of large sets of conditioning information without imposing strong parametric assumptions such as linearity or monotonicity. Our forecasts outperform those generated by benchmark models in terms of both mean squared error and directional accuracy. They also generate profitable portfolio allocations for mean-variance investors even when market frictions are accounted for. Finally, our analysis shows that the relation between the predictor variables constituting the conditioning information set and the investors' optimal portfolio allocation to risky assets is, in most cases, nonlinear and nonmonotonic.

References

Aït-Sahalia, Y., and M. Brandt (2001): "Variable Selection for Portfolio Choice," *The Journal of Finance*, 56(4), 1297–1351.

Bai, J., and S. Ng (2009): "Boosting di↵ usion indices," *Journal of Applied Econometrics*, 24(4), 607–629.

Baker, M., and J. Wurgler (2000): "The Equity Share in New Issues and Aggregate Stock Returns," *The Journal of Finance*, 55(5), 2219–2257.

Bossaerts, P., and P. Hillion (1999): "Implementing statistical criteria to select return forecasting models: what do we learn?," *Review of Financial Studies*, 12(2), 405–428.

Breen, W., L. Glosten, and R. Jagannathan (1989): "Economic significance of predictable variations in stock index returns," *Journal of Finance*, 44(5), 1177–89.

Breiman, L. (1984): *Classification and regression trees Regression trees The Wadsworth statistics/probability series*. Wadsworth International Group.

(1996): "Bagging predictors," *Machine learning*, 24(2), 123–140.

Campbell, J. (1988): "Stock Returns and the Term Structure," *NBER Working Paper*.

ISSN: 2366-1313

Campbell, J., and R. Shiller (1988a): "The dividend-price ratio and expectations of future dividends and discount factors," *Review of Financial Studies*, 1(3), 195–228. Campbell, J., and R. Shiller (1988b): "Stock prices, earnings, and expected dividends,"

Journal of Finance, 43(3), 661–676.

Campbell, J., and M. Yogo (2006): "Efficient tests of stock return predictability," *Journal of Financial Economics*, 81(1), 27–60.

Campbell, J. Y. (1987): "Stock returns and the term structure," *Journal of Financial Eco- nomics*, 18(2), 373–399.

Campbell, J. Y., and S. Thompson (2008): "Predicting excess stock returns out of sample: Can anything beat the historical average?," *Review of Financial Studies*, 21(4), 1509–1531.

Cumby, R., and D. Modest (1987): "Testing for Market Timing Ability: A Framework for Forecast Evaluation," *Journal of Financial Economics*, 19(1), 169–189.

Dangl, T., and M. Halling (2008): "Predictive regressions with time-varying coefficients," *Working Paper, Vienna University of Technology.*

Engle, R., E. Ghysels, and B. Sohn (2006): "On the Economic Sources of Stock Market Volatility," *Manuscript, New York University*.

Engle, R., and J. Rangel (2005): *The Spline GARCH Model for Unconditional Volatility and Its Global Macroeconomic Causes*. Czech National Bank.

Fama, E., and K. French (1988): "Dividend yields and expected stock returns," *Journal of Financial Economics*, 22(1), 3–25.

Fama, E., and K. French (1989): "Business conditions and expected returns on stocks and bonds," *Journal of Financial Economics*, 25(1), 23–49.

Fama, E., and G. Schwert (1977): "Asset Returns and Inflation," *Journal of Financial Economics*, 5(2), 115–146.

Fleming, J., C. Kirby, and B. Ostdiek (2001): "The Economic Value of Volatility Timing," *Journal of Finance*, 56, 329–352.

Friedman, J. H. (2001): "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, 29, 1189–1232.

(2002): "Stochastic gradient boosting," *Computational Statistics and Data Analysis*, 38(4), 367–378.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2005): "There is a Risk-Return Tradeo↵ After All," *Journal of Financial Economics*, 76(3), 509–548.

ISSN: 2366-1313

Ghysels, E., P. Santa-Clara, and R. Valkanov (2006): "Predicting volatility: getting the most out of return data sampled at di↵erent frequencies," *Journal of Econometrics*, 131(1-2), 59–95.

Goyal, A., and I. Welch, I. (2003): "Predicting the Equity Premium with Dividend Ratios," *Management Science*.

Hansen, P., and A. Lunde (2005): "A forecast comparison of volatility models: Does anything beat a GARCH (1, 1)?," *Journal of Applied Econometrics*, 20(7), 873–889.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

ISSN: 2366-1313

Henriksson, R., and R. Merton (1981): "On Market Timing and Investment Performance.

II. Statistical Procedures for Evaluating Forecasting Skills," *Journal of Business*, 54(4), 513.

Hodrick, R. (1992): "Dividend yields and expected stock returns: alternative procedures for inference and measurement," *Review of Financial Studies*, 5(3), 357–386.

Johannes, M., A. Korteweg, and N. Polson (2009): "Sequential learning, predictive re- gressions, and optimal portfolio returns," *Mimeo, Columbia University*.

Keim, D., and R. Stambaugh (1986): "Predicting returns in the stock and bond markets," *Journal of Financial Economics*, 17(2), 357–390.

Kothari, S., and J. Shanken (1997): "Book-to-market, dividend yield, and expected market returns: A time-series analysis," *Journal of Financial Economics*, 44(2), 169–203.

Lamont, O. (1998): "Earnings and Expected Returns," *The Journal of Finance*, 53(5), 1563– 1587.

Leitch, G., and J. Tanner (1991): "Economic Forecast Evaluation: Profits Versus the Con- ventional Error Measures," *American Economic Review*, 81(3), 580–590.

Lettau, M., and S. Ludvigson (2009): "Measuring and Modeling Variation in the Risk- Return Tradeo↵ ," *Handbook of Financial Econometrics, edited by Y. Ait-Sahalia and L.P. Hansen. North Holland.*

Lewellen, J., S. Nagel, and J. Shanken (2010): "A skeptical appraisal of asset pricing tests," *Journal of Financial Economics*, 96(2), 175–194.

Lo, A., and A. MacKinlay (1990): "Data-snooping biases in tests of financial asset pricing models," *Review of Financial Studies*, 3(3), 431–467.

Ludvigson, S. C., and S. Ng (2007): "The empirical risk-return relation: A factor analysis approach," *Journal of Financial Economics*, 83(1), 171–222.

Maenhout, P. (2004): "Robust Portfolio Rules and Asset Pricing," *Review of Financial Studies*, 17(4), 951–983.

Marquering, W., and M. Verbeek (2005): "The Economic Value of Predicting Stock Index Returns and Volatility," *Journal of financial and Quantitative Analysis*.

Nelson, C., and M. Kim (1993): "Predictable stock returns: The role of small sample bias," *Journal of Finance*, 48(2), 641–661.

ISSN: 2366-1313

Paye, B. (2010): "Do Macroeconomic Variables Predict Aggregate Stock Market Volatility?," *Working Paper Series*.

Pesaran, H., and A. Timmermann (2005): "Real-time econometrics," *Econometric Theory*, 21(01), 212–231.

Pesaran, M., and A. Timmermann (1995): "Predictability of Stock Returns: Robustness and Economic Significance," *Journal of Finance*, 50(4), 1201–1228.

Pesaran, M., and A. Timmermann (2000): "A Recursive Modelling Approach to Predicting UK Stock Returns 'I. Working; Paner No. 9625. Department of Applied Economics, Universitv of Cambridge," *Economic Journal*, 110, Issue 460.

Pontiff, J., and L. Schall (1998): "Book-to-market ratios as predictors of market returns," *Journal of Financial Economics*, 49(2), 141–160.

Rapach, D., J. Strauss, and G. Zhou (2010): "Out-of-sample equity premium prediction: Combination forecasts and links to the real economy," *Review of Financial Studies*, 23(2), 821.

Rossi, A., and A. Timmermann (2010): "What is the Shape of the Risk-Return Relation?," *UCSD Discussion Paper*.

Rozeff, M. (1984): "Dividend yields are equity risk premiums," *Journal of Portfolio Manage- ment*, 11(1), 68–75.

Stambaugh, R. (1999): "Predictive Regressions," *Journal of Financial Economics*, 54(3), 375– 421.

Sullivan, R., A. Timmermann, and H. White (1999): "Data-Snooping, Technical Trading Rule Performance, and the Bootstrap," *The Journal of Finance*, 54(5), 1647–1691.

Ter Horst, J., F. De Roon, and B. Werker (2000): "Incorporating Estimation Risk in Portfolio Choice," *CentER Working Papers Series*.

Welch, I., and A. Goyal (2008): "A Comprehensive Look at The Empirical Performance of Equity

4