Applied Economics and Finance; Vol. 10, No. 4; 2023
ISSN 2332-7294 E-ISSN 2332-7308
Published by Redfame Publishing

Measuring the Relevance of Factors on Cross-Sectional Returns with Decision Trees

Paul Felix Reiter[1]

[1] Technische Universität Dresden, Germany

Correspondence: Paul Felix Reiter, Faculty of Business and Economics, Technische Universität Dresden, Helmholtzstr. 10, 01069 Dresden, Germany. E-mail: paul_felix.reiter1@tu-dresden.de

Received: August 1, 2023 Accepted: October 31, 2023 Available online: November 20, 2023

doi:10.11114/aef.v10i4.6285 URL: https://doi.org/10.11114/aef.v10i4.6285

Abstract

This study is concerned with new ways to identify and analyse the factors on cross-sectional returns in financial markets with respect to their time-variability. Therefore, classification and regression trees and conventional regression models are applied. This study uses data on the S&P 500 from 1999 to 2019. Empirical findings show high time variability of factors on cross-sectional returns. The high level of time-variability is not dependent on the applied model. It is also shown that CARTs and conventional regression models have low power when it comes to identifying the factors on cross-sectional returns or predicting the returns themself.

Keywords: finance, econometrics, decision trees, machine learning, forecasting, variable importance

1. Introduction

Factors influencing the cross-sectional returns of stocks cover a large part of the financial econometrics literature. In 2016, an article criticizing the research methods utilized in this field was published by Harvey, Liu and Zhu. They highlighted that because the sample used to identify factors related to cross-sectional returns is almost always the U.S. stock market, the same sample was employed to test their hypothesis regarding possible factors. This approach leads to the problem of multiple testing, resulting in many false positive factors. This article identified 314 factors affecting cross-sectional returns and sparked a new discussion on the methods used in financial econometrics (Harvey, Liu, & Zhu, 2016).

Another discussion concerning the use of p-values predominant statistics and its adjacent fields, such as biology or psychology. In recent years, this discussion came to the domain of econometrics (Harvey, 2017).

A proposed solution was to adjust the p-values to correct the shift of thresholds caused by multiple testing such as Bonferroni-Holm or Benjamini-Hochberg corrections. These methods require the overall number of performed tests to correctly adjust the p-values. As outlined in the above mentioned 2016 article this number can only be estimated. By assuming that the number of significant and published factors is equal to the number of tested factors, they calculated a lower bound for the t-statistic. The correct critical value for a correct specified t-test at a 5 percent level of significance is to be 3.78 for a new factor to be published prior to 2012 (Harvey et al., 2016).

Assuming that the number of tested factors follows a linear growth path, it is expected for a statistically significant factor to have a t-statistic of at least 4.00 in 2032. As there is a strong publication bias for significant values it is estimated that 71 percent of all possible factors are not published. Hence, the true corrected critical values are higher than the above-stated numbers (Harvey et al., 2016).

Another caveat applying statistical tests with corrected values is that multiple test corrections are too conservative and have low statistical power (Emmert-Streib & Dehmer, 2019).

This study applies decision trees, a non-parametric approach, to the problem of identifying relevant factors affecting cross-sectional returns. With the rise of machine learning techniques utilized in almost all areas of science, the decision tree approach has come into focus of (Mullainathan & Spiess, 2017; Varian, 2014).

Decision trees have a hierarchical structure and therefore possess the ability to assess the importance and relevance of a variable, or in this case, a factor on cross-sectional returns. The algorithm employed to build the decision trees is the classification and regression tree algorithm (CART), published in 1984. This method stems from the field of machine learning and is widely employed to predict discrete (classification) and continuous outcomes (regression). Unlike other

machine learning algorithms it is not a black box and thus can be used for structural analysis of the underlying sample (Breiman, Friedman, Olsen, & Stone, 1984).

Decision trees have seen limited use in financial econometrics. When utilized as a multivariate sort, they are can display interaction effects (Moritz & Zimmermann, 2016). A similar approach using regression trees to identify the factors influencing cross-sectional returns was performed by (Coqueret & Guida, 2018), but they focused solely on regression trees. Both of the above articles find the dominance of momentum based factors and the possibility of mapping complex relationships with decision trees. A more recent approach has focused on the implementation of more economic guidance into regression trees (He, Cong, Feng, & He, 2021). As regression trees only provide a piecewise approximation of the range of possible values, we will instead focus on returns above average. Hence, classification trees can be used. This study adds to the existing literature by comparing the performance of regression models to the performance of decision trees when used for structural analysis.

The application of CARTs is reviewed with a focus on the possibility of time-variable coefficients. The problems of dynamic coefficients examining cross-sectional returns have been studied for some time (Jagannathan & Wang, 1996). This study is primarily concerned with the time variation of the importance of variables over time. A similar study has been published but with a strong focus on the slopes of the coefficients (Lewellen, 2015).

The use of machine learning in finance garnered much attention in recent years, with numerous publications concerned with the predictive power of machine learning models (Bryzgalova, Pelger, & Zhu, 2019; Gu, Kelly, & Xiu, 2020). The task of identifying factors impacting cross-sectional returns has not been in the focus of recent research in this field. Only one article titled ”Shrinking the cross-section” by (Kozak, Nagel, & Santosh, 2017) has been published but its authors are very skeptical the possibility of identifying such factors. These results can be confirmed. Another recent approach is the implementation of news-based and sentiment-based features (Zhu, Wu, & Wells, 2023).

In this article, CARTs will be employed in a classification role to determine the factors on above-average returns and in a regression role to determine the factors that influence returns. First, we provide a literature review on training and interpretation of CARTs and a comparison with the standard regression models such as ordinary least squares (OLS)-”like” and logit-regressions. This step is followed by an application to determine how CARTs behave in financial econometrics and how the results fit the financial theory.

2. Applied Methods and Properties

2.1 Classification and Regression Trees

CARTs are hierarchical and non-parametric models stemming from machine-learning. As trees do not estimate the parameters of an assumed data-generating process but derive a set of rules for splitting a sample, the machine-learning term “training” is more appropriate than the term “estimating” used in econometrics as the input space represented by the sample is divided into local regions. Dividing the sample is done by a function with the following form:

$$y_i = \theta(x_i, \alpha) \tag{1}$$

The form of θ and the process of finding the relationship between the endogenous variable y_i and the vector of exogenous variables x_i is described in the following. The necessity and choice of the stopping criterion α is explained in chapter 2.2. This function can be used for classification, θ is denoted as θ_{Cl}. If used for a regression, θ is denoted as θ_{Reg}. In a classification setting, with a sample containing $c = 1, \dots, C$ classes, the sample is divided into subsamples with only one class. In a regression setting, the sample is divided into subsamples with observations that deviate less from the arithmetic mean of the subsample. The sample is organized into a dependent variable y and independent variables x_i with $i = 1, \dots, l$. This sample contains $j = 1, .., n$ observations. Exogenous variables are used to divide the heterogeneous set of endogenous variables into homogeneous subsets. The partition is achieved by a set of test functions $f_m(x_i)$ housed in the nodes w_m with $m = 1, \dots, M$ and M being the total number of nodes. The test function is $f_m(x_{ij} > s)$, where s is the split point and $s \in S$. S is the set of all possible splits. The first node contains the first test function $f_1(x_i)$ and is called the root. Here a binary split is performed, and the sample is split into two sub-samples by checking each observation of x_i. At any node w_m a part of the sample is sent to the right branch leading to the next node on the right with proportion p_R and to the left branch with proportion p_L. The next two nodes contain the test functions $f_2(x_k)$ and $f_3(x_h)$. This process is repeated until the stopping criterion is satisfied. The last nodes are called endnodes or leaves. The training process is divided into two parts. The first part is concerned finding the optimal split point s^* at the nodes, a node w with its optimal split point s^* assigned is denoted w^*. The second part involves setting the correct stopping criterion. In particular the second part is crucial and will be discussed in depth later. To locate the optimal split, the CART-algorithm utilizes different metrics depending on whether it is a classification or a regression setting. The goal of the splitting process is to maximize the decrease in the impurity of the observations in the next two following nodes (Breiman et al., 1984).

$$max_{s \in S} \Delta I(s, w) = I(t) - p_L I(w_L) - p_R I(w_R) \quad (2)$$

In a classification setting, s is chosen to maximise the decrease of the Gini impurity $\Delta I(s,w)$ of the following sub-samples, with the Gini impurity defined as $I(w) = 1 - \sum_{c=1}^{C} p_c^2$ of the following sub-samples. Notably, p_i is the share of class c in node w, and the next two nodes are called w_L and w_R. In a regression setting s is chosen to minimize the sum of squared residuals in the next two following nodes. In both cases an exhaustive search is performed over all variables and possible split points (Breiman et al., 1984).

2.2 Stopping Criterion

The need for a stopping criterion arises in both applications of trees (i.e., structural analysis and prediction) from the complexity of a fully grown tree. In a prediction application the bias-variance trade-off leads to a large out-of-sample error. When utilizing a tree for structural analysis, the problem is that the algorithm splits the sample until only one element in each leave remains. This mechanism leads to as many paths as there are observations and makes aggregate results or a generalized conclusion impossible. A fully grown tree has a small in-sample prediction bias because the tree is very fine and perfectly splits the training sample. However, the variance component of the error is very high, as the tree cannot account for variance in out-of-sample observations (Hastie, Friedman, & Tisbshirani, 2017).

This issue is alleviated by employing a stopping criterion to limit tree growth, reducing tree growth (i.e., pruning). Pruning can be divided into pre- and post-pruning. The former involves choosing a certain threshold to limit the tree growth (e.g., a minimum sub-sample size). The latter allows the tree to grow to its full size and is then pruned. The CART algorithm uses cost-complexity-pruning. Here, the following cost function is used:

$$C_\alpha(M) = E(M) - \alpha * M \quad (3)$$

$E(M)$ Is an error measurement of a tree with M nodes in total, and α is the complexity parameter, which is used to determine the optimal trade-off. Here, the out-of-sample misclassification rate or mean squared error is used as error measure, and complexity is defined as the product of the complexity parameter and the number of nodes M. The higher α, the less complex the tree, and vice versa. Notably, the optimal α is determined by k-fold-cross-validation (Hastie et al., 2017).

When employing k-fold-cross-validation, the sample is split into k sub-samples, where $k-1$ sub-samples are employed to train the algorithm and the remaining sub-sample is used to test it. This process is repeated k-times, and the results are aggregated. As this process involves randomness, it is problematic to utilize it in small samples. Here, it can lead to non-optimal stopping criteria and non-replicability (Isaksson, Wallman, Göransson, & Gustafsson, 2008).

Cost-complexity pruning provides an advantage as it uses a non-arbitrary criterion as opposed to an arbitrary threshold set by the researcher. As our approach is more interested in the structure of the sample than the development of prediction rules, both approaches will be employed.

2.3 Interpretation of CARTs

As the arithmetic mean is an unbiased estimator for the expectation of a random variable
it is also an unbiased estimator of the success probability of a binary random variable. In the root node of a CART, this arithmetic mean is an unconditional expectation or probability. In the next node, it becomes conditional a value, as the full sample from the root is split into sub-samples. The means of the endogenous variable in these sub-samples are conditioned on the values of the split in the node above. Further down the decision tree the sub-samples become finer as more information is processed. Hence, the conditional probabilities and expectations have become more complex. When performing a regression, this leads to a piecewise approximation of the range of possible values, where the pieces can be examined by their conditions. When performing a classification the leaves are unbiased estimators of the respective conditional probabilities. In both cases a set of IF-THEN rules can be derived from the leaves (Alpaydin, 2014).

As the variables in the nodes are chosen to minimize the sum of squared residuals or to maximize the Gini impurity, the earlier a variable is chosen the more important it is. Variable importance is quantified by using surrogate splits. A surrogate split s^{SU} is defined by its ability to act in the same manner as the optimal split s^*. Therefore, two conditions must be met (Breiman et al., 1984). The first condition is:

$$p(s_i^{SU}, s^*) = max_{s_i} \left(p_L(s_i^{SU}, s^*) - p_R(s_i^{SU}, s^*) \right) \quad (4)$$

Here $p_L(s_i^{SU}, s^*)$ is the probability that an observation is assigned to the left node by the optimal split s^* and a split at any s_i. The behavior of $p_R(s_i^{SU}, s^*)$ corresponds to the right node. The surrogate split is the split of variable x_i which

acts most similarly to s^* (Breiman et al., 1984). The second condition is:

$$min(p_L, p_R) - \left(1-p(s^{SU}, s^*)\right) > 0 \tag{5}$$

This condition ensures that the surrogate split is of any value in the prediction of the outcome variable (Breiman et al., 1984). With these two conditions satisfied, the variable importance measure can be calculated as follows:

$$VI(x_i) = \sum_{m \in M} \Delta I(s^{SU}, s^*) \tag{6}$$

The variable importance $VI(x_i)$ is the sum of its possibilities to reduce the Gini impurity, respectively the sum of squared residuals. The possibility of a variable for reduction is the difference between optimal split and surrogate split $\Delta I(s^{SU}, s^*)$. If no split satisfies the above conditions, then this possibility becomes zero (Breiman et al., 1984).

2.4 Comparison to Parametric Regression Methods

The differences between CARTs and parametric regression techniques can be categorized into two parts. The first part is the performance and ability to capture the sample's structure. The second part is the interpretation of the resulting model.

Regarding the comparison of performance, a large-scale study revealed that CARTs outperformed logistic regressions in larger data sets. This notion holds, particularly when the signal separability is high (Perlich, Provost, & Simonoff, 2003). Compared to other classification algorithms, such as artificial neural nets or the random-forest-algorithm, CARTs and logit regressions offer a lower performance as a trade-off for interpretability (Caruana & Niculescu-Mizil, 2006).

In general, it is expected that regression trees perform worse than linear regression models. This expectation is valid because regression trees only offer a piecewise approximation of the range of possible values. Additionally linear regression models are well understood and sophisticated with respect to their ability to model individual and time effects. Regression trees are only expected to perform better when the advantages are offset by their ability to model nonlinear effects. This notion prove to be relevant case in econometrics, as shown by a study on the productivity of workers (Markham, 2011).

Another domain wherein CARTs can fit the data very well is in the presence of interaction effects. These are incorporated into the IF-THEN structure inherent to decision trees. In a parametric regression, it is assumed that all variables have the same influence. Therefore interaction effects have to be explicitly specified (James, Witten, Hastie, & Tibshirani, 2021; Long, Griffith, Selker, & D'Agostino, 1993).

Parametric regression models and tree-based models also differ in their ability to handle missing data. If the training sample contains missing data, the observations with missing values for a specific variable are not used to calculate the splits for this variable. These observations can still be used to calculate splits on other variables. This procedure alleviates the problems associated with missing data. When the decision tree is used for prediction, new observations with missing values can still be used. Surrogate splits can be used at splits where a missing value is needed (Breiman et al., 1984). In addition to the variable importance mentioned in the above chapter, tree-like models are easy to interpret and can be employed to formulate an easily applicable guideline for decision-making because of to their graphical representation (James et al., 2021). An example is the reaction to a possible heart attack in an emergency room. A quick assessment of a logistic regression is not possible by readers not skilled in statistics and/or econometrics. A tree model can be put into an easy-understandable flow chart, as described by (Tsien, L., Christine, Fraser, F.S., Hamisch, & Long, J., William, 1998). With a logit regression the marginal effects of a variable depend on the level of other variables, leading to wrong interpretations of scientific studies when presented to the general public (Hoetker, 2007). This problem is alleviated in an OLS regression unless there are interaction terms or the variables have been scaled to allow for the coefficients to be in the same range.

2.5 Parametric Regression Models

In most publications on cross-sectional returns a Fama-MacBeth regression is used. This approach can result in biased estimates of standard errors due to disregard for firm effects (Canitz et al., 2017; Hoechle, Schmid, & Zimmermann, 2018). In this study a panel regression with fixed effects is used to account for time and firm effects.

$$y_{it} = x'_{it}\beta + u_i + \varepsilon_{it} \tag{7}$$

The effects of each firm are considered by u_i, a firm specific and constant term. y_{it} is the endogenous variable of observation i in period t. The vector x_{it} contains the exogenous variables for observation i in period t and ε_{it} is the

associated shock. The vector β contains the parameters to be estimated. As this article also employs OLS-regressions for the single period case, both methods will be summed up as OLS-derivatives.

In the classification setting a binary choice model for panel data with fixed effects is used, namely the conditional logit estimator (Chamberlain, 1980).

$$P(y_{it}^{*} = 1) = g(x'_{it}\beta^{*} + u_i^{*} + \varepsilon_{it}^{*}) \tag{8}$$

Here the probability of a binary variable y_{it}^{*} being equal to one is the cumulated distribution function of the logistic distribution *g()*, whose arguments are the features of observation *i* in period *t*. The parameters denoted with a * have an analogue meaning to the ones in equation (7).

In the single-period case a standard logit-model is utilized. This article will use the general term "logit-regressions" for both types.

3. Empirical Analysis

The setting to test the change in capital markets over time stems from the article "Dissecting Anomalies" by (Fama & French, 2008). In this article seven factors on cross-sectional returns and their characteristics across different levels of market capitalization were studied. These seven potential factors concerning cross-sectional returns were chosen here because they are well known and understood in the literature. In this article the objects of the investigation will not be different levels of market capitalization but the time dimension and its influence on factors on cross-sectional returns. Thus, only stocks with a high market capitalization are considered (i.e., the constituents of the S&P 500 from 1999 to 2019). Overall, two periods are needed to calculate the variables, and one period is set aside for prediction, leaving 17 periods for training. Financial institutions were not considered. To determine the factors impacting cross-sectional returns, a fixed-effects regression and a regression tree are used. To make use of the possibility to fit a classification tree to data with binary outcomes, the factors on above-average returns will also be examined. For this purpose, a new dependent variable y_{it}^{*} it is defined as:

$$y_{it}^{*} = \begin{cases} 1, \text{ if } y_{it} > \bar{y}_t \\ 0, \text{ else} \end{cases} \tag{9}$$

Here, $\bar{y}_t$ is the average return of all stocks in the sample in month j of year t. The variables used stem from (Fama & French, 2008), together with their calculation and scaling. For a better comparison between parametric regressions and CARTs, the variables will be scaled to allow for "nice"-coefficients (i.e., in the same range). This step is not necessary for CARTs.

The data source was Thomson Reuters Datastream. The returns are calculated from month $j-1$ to month j. Momentum is updated each month. All other variables are updated in June of each year. The following exogenous variables were employed:

- momentum (mom), calculated from month $j-12$ to $j-2$
- market capitalization (mc), stocks outstanding at the beginning of year t the times stock price at the beginning of year t, logarithmized
- book-to-market ratio (btmr), book value at the beginning of year t is total assets minus total liabilities, divided by market capitalization, logarithmized
- net stock issues (ns), common stock outstanding at the beginning of year t minus common stock outstanding at the beginning of year $t-1$, logarithmized
- accruals (acc), difference in working capital between years t and $t-1$, working capital is calculated as current assets minus current liabilities, logarithmized
- change in assets (chA), total assets minus total liabilities at the beginning of year t minus total assets minus total liabilities in year $t-1$, logarithmized
- profitability (ptb), income in year $t-1$ divided by the total assets in the respective year

Summary statistics for all variables can be found in Table I in the Appendix. These seven variables were used to train the following models.

Classification trees:

$$y_{it} = \theta_{Cl}((mom_{it}, mc_{it}, btmr_{it}, ns_{it}, acc_{it}, chA_{it}, ptb_{it},), \alpha) \tag{10}$$

Regression trees:

$$y_{it} = \theta_{Reg}((mom_{it}, mc_{it}, btmr_{it}, ns_{it}, acc_{it}, chA_{it}, ptb_{it},), \alpha) \quad (12)$$

In every tree model trained, the stopping criterion α was determined by 10-fold-cross-validation.
OLS-derivatives:

$$y_{it} = \beta_0 + \beta_1 mom_{it} + \beta_2 mc_{it} + \beta_3 btmr_{it} + \beta_4 ns_{it} + \beta_5 acc_{it} \quad (13)$$
$$+\beta_6 chA_{it} + \beta_7 ptb_{it} + u_i + \varepsilon_{it}$$

Logit-regressions:

$$P(y_{it}^* = 1) = g(\beta_0 + \beta_1 mom_{it} + \beta_2 mc_{it} + \beta_3 btmr_{it} + \beta_4 ns_{it} + \beta_5 acc_{it} \quad (14)$$
$$+\beta_6 chA_{it} + \beta_7 ptb_{it} + u_i^* + \varepsilon_{it}^*)$$

Notably, two scenarios were utilized to determine the time variability. First, the data from $t-1$ was used to train the aforementioned models. Thereafter, this model will be used to predict the returns in the next period t. To compare parametric regression models and decision trees in their ability to fit the data, the out-of-sample error rate is calculated. The error measure in the classification setting is the misclassification rate (MCR), and in the regression part, the mean squared error (MSE) is used. The forecast horizon is always one month; here, we follow the forecasting framework proposed by (Dong, Li, Rapach, & Zhou, 2022). Additionally, the absolute values of the t-statistics and variable importance scores are calculated. Now the training sample is expanded with the data from $t-2$ and the process is repeated. In this backtesting approach further sub samples are added until all periods from $t-1$ to $t-17$ are used. If t-values and variable importance scores converge, larger datasets are more useful, and there are no changes in the factors impacting returns over time. The same holds for the development of error measures with larger samples sizes. If the forecasting error grows larger with more periods added, different periods distort the estimation owing to shifts in the data-generating process. Second, a rolling window approach where each period from $t-1$ till $t-17$ was used for training and forecasting of the next period. During stable conditions in financial markets, the factors on returns would be constant or slowly changing. In both settings, t-statistics, variable importance scores, and error measurements are displayed as time series.

4. Results

The absolute of the test statistic of the t-test $|z|$ for the estimated effects in the setup with backtesting is plotted in figure 1. The left side denotes the plot for the fixed-effect regressions and the right side for the fixed effect logit regressions. In both cases there is only the t-statistic for momentum available in t = 1, because all other factors influencing cross sectional returns are time-invariant by design. The sample size grows from 5988 observations in period $t-1$ to 70795 observations in period $t-17$. The number of observations does not grow at a constant pace because the number of companies leaving and entering the S&P 500 and their respective fields changes every year.

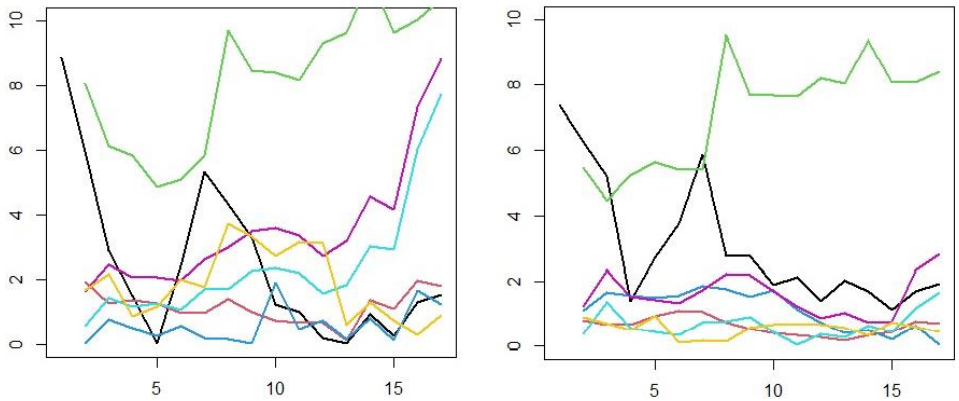

Figure 1. Z-score of parametric regressions with backtesting

Description: Left side: fixed effects regressions, right side: fixed effects logit regressions, y-axis = |z|, x-axis = number of periods included, black = momentum, red = accruals, green =market capitalization, blue = book-to-market ratio, turquoise = net stock issues, purple = change in assets, yellow = profitability

Looking at $|z|$ with a 5% significance level it is evident that only market capitalization is the factor that always has an effect significantly different from zero. This notion holds for the regression and classification settings. The momentum also shows a similar development in both cases. Unlike mc the significance of its effect is conditional on the periods

included. Change in assets and net stock issues have similar trajectories. Both have a declining $|z|$-value in the regression setting. In the classification setting the trend is inconclusive. Here chA sometimes has a coefficient significantly different from zero, whereas ns has not.

The R2 in the fixed effect regressions is near zero for $t-4$ to $t-17$, rising sharply from 0.0035 to 0.0141 in periods $t-3$ to $t-1$. The same pattern is shown for the McFadden-pseudo-R^2 calculated from the fixed-effect logit regressions. This notion leads to the conclusion that the coefficients, and therefore the data-generating process, change over time. When trying to fit time-constant coefficients to data containing 10 years of data, one achieves a worse fit than when only two years of data are utilized. This aspect, in tandem with the change in value and significance of coefficients, shows the time variability of factors on cross-sectional returns. Looking at Figure 4, the variability of the $|z|$ of the single-period regressions can still be observed. Some minor co-movements between the values of the OLS regressions and the logit regressions can be identified, (e.g., the spike of the $|z|$ of the influence of accruals in period $t-9$).

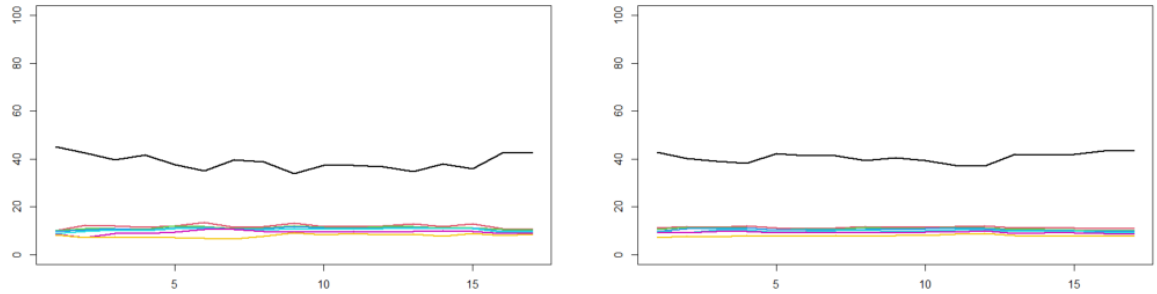

Figure 1. Variable Importance score of CARTs with backtesting

Description: Left side: regression trees, right side: classification trees, y-axis = variable importance, x-axis = number of periods included, black = momentum, red = accruals, green =market capitalization, blue = book-to-market ratio, turquoise = net stock issues, purple = change in assets, yellow = profitability

While $|z|$ in Figure 1 shows a lot of dependence on the included periods, the variable importance metric in the decision trees in Figure 2 does not. In both cases, momentum is the most important variable, with a variable importance score of approximately 40. All other variables have a variable importance score of around 15. In the regression and the classification setting, change in assets is the least important one. As shown in Figure 5 in the Appendix the same behavior of the |z|-score can be seen if the regressions are performed with only one period. Figure 5 also shows large changes in the |z|-score during the periods affected by the 2007-2008 financial crisis. This is in line with the results of Kozlowski and Lytle (2023), who showed the effect of recessions on the January anomaly.

Figure 6 illustrates that the variable importance metric of CARTs also behaves the same if the trees are trained with a single period.

There were some minor variations. However, as shown in Figure 2, the overall picture does not change. The dominance of return- or momentum-based predictors in variable importance has been confirmed in all studies using CART (Coqueret & Guida, 2018; Moritz & Zimmermann, 2016). As momentum is calculated from the past returns, its high importance supports the representation of returns with a time series model (e.g., an autoregressive process).

Unpruned trees were used to calculate the variable importance scores, this is due to not being able to calculate a variable importance score in every case otherwise. In all cases using a regression tree no split is left after pruning, leading to remarkable conclusions regarding the fit.

In Figure 3, the MCR and MSE of both methods are plotted. On the left, the regression tree has a lower MSE than the fixed-effect regression in every period. The spike of the fixed effects regressions MSE when only $t-2$ and $t-1$ are included is probably an anomaly. Nonetheless, the stump of the regression trees, using only the arithmetic mean of the last period as the predicted value for the next period, always has more predictive power than OLS derivatives. Thus, their usefulness for the future policy or investment decisions is disputed. The result of the parametric methods are consistent with the known phenomenon of ”pockets of predictability” in stock markets (Farmer, Schmidt, & Timmermann, 2023; Timmermann, 2008). CARTs do not show these pockets of predictability, which may be due to their less rigid structure in comparison to parametric methods. Hence, they can better adapt to changes in the data. Figure 7 in the Appendix portrays a similar picture for the MCR if only one period is used. Specifically, the MSE is more volatile when only one period is used. The first large spike in Figure 7 corresponds to the financial crisis in 2008 and its aftermath. The reasons for the second spike in Figure 7 and its slow ascent and rapid descent are unclear.

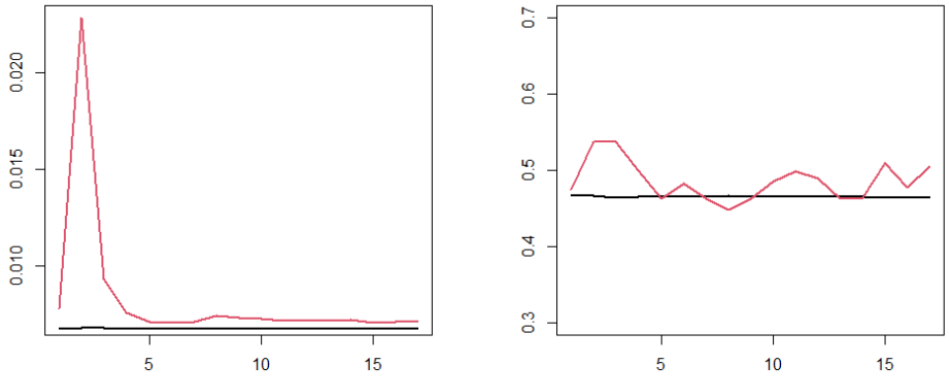

Figure 2. Precision metrics backtesting

Description: Left side: MSE, right side: MCR, y-axis = MSE respectively MCR, x-axis =periods included, black = trees, red = regressions

Figure 4 shows the classification trees with the lowest misclassification rate. The left tree was grown with all periods from $t-1$ to $t-8$. The tree on the right was grown with only with from the period $t-8$. The MCR of the left tree is 0.448, and the MCR of the right tree is 0.437. The MCR of the corresponding logit models are 0.463 and 0.422 respectively. In each node, the first number is the prevalent outcome, the second number is the probability of success, and the third is the size of the subsample in the node as a percentage of the full sample. The variable of the first split point was identical in both trees. In both cases, momentum is used. However, the direction of the split is different.

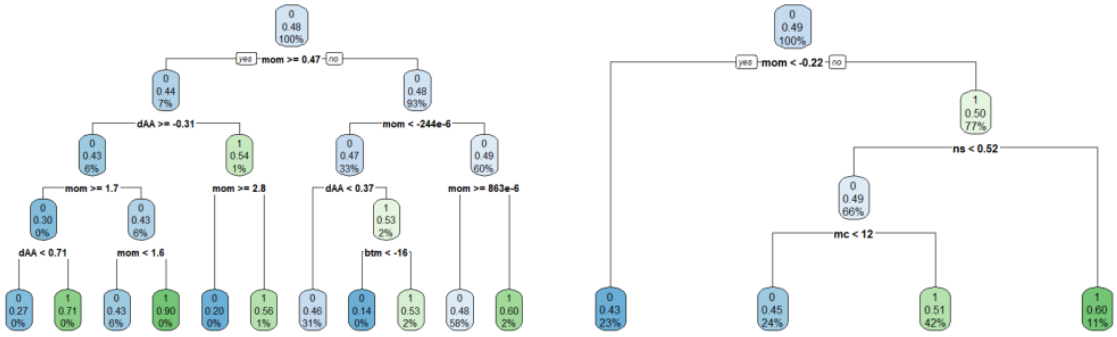

Figure 3. Pruned classification trees

Description: Left side: classification tree with 8 periods, right side: classification tree with one period. In the right tree momentum above 0.47 has a negative influence on the success probability, while on the left tree a momentum above −0.22 has a negative influence. Although the importance of the variables is more or less constant, the influence of the variables on the outcome changes over time.

An indication of the usefulness of the IF-ELSE structure of decision trees is the behavior of dAA in the left tree. In the left tree, $P(1|mom \leq 0.47 \wedge dAA < -0.31) > P(1|mom \leq 0.47 \wedge dAA \geq -0.31)$, but for a different range of momentum $P(1|mom < 0.000244 \wedge dAA < 0.37) < P(1|mom < 0.000244 \wedge dAA \leq 0.37)$. This notion is equivalent to a parameter change in a parametric regression depending on the covariates values. A similar behavior of dAA and ns can also be found in the left tree. This result is interesting as it confirms the thesis of time varying coefficients. The importance of variables does not change, but the split points do.

The leaves of both trees contained either large subsamples or very small subsamples. In both trees, the splits lead to a relatively small sample being split from a large sample or a small sample being split into two smaller samples. In the left example, the two largest leaves contain 87 percent of the original sample, and those leaves have a success probability of less than 0.02, away from 0.5. Only some of the small leaves can clearly distinguish observations belonging to the group with above-average returns or the group with below-average returns. The leaves in the right tree are more equally sized, but the success probability of the two largest leaves does not deviate more than 0.05 from 0.5. Therefore, decision trees cannot account for a large part of the variation. This may equivalent to the low R2 observed in the literature on cross-sectional returns.

5. Discussion

It can be concluded that the application of CARTs in the identification of cross-sectional returns is problematic. Especially when used as regression trees to determine factors influencing cross-sectional returns they cannot detect any structure and cannot split the sample. OLS derivatives can detect structures in the sample. A comparison of the predictive power of regression trees and OLS derivatives show a higher error for OLS derivatives. This error is a clear indication of overfitting. The best predictor of future stock returns is the mean of all returns from the last period. To obtain a reduction in forecasting error, it is imperative to employ more powerful methods such as artificial neural nets or the random forest algorithm. Both methods are regarded as black boxes and are not suitable for structural inference. When using classification trees to identify influencing factors in relation to the trend, decision trees can split the sample. Comparing the predictive power of CARTs and logit regressions, the results are inconclusive. Figures 3 and 7 are showing that the results of CARTs are often worse than the ones of the logit regressions. Only in some cases can CARTs have a lower prediction error than logit regression. As in the regression case the problem may be with the low signal separability. A possible explanation is that the interaction effects are not that important. Concluding on the structural analysis, the dominance of momentum is an influencing factor of returns and the trend of returns. This finding is in consistent with the concept of stock returns as autoregressive processes. This study also confirms the high level of time-variability in stock markets. Figures 1 and 6 specify the change in structure in the stock market over time. This problem cannot be alleviated using larger samples. Therefore, further research should include a broader sample covering the whole stock market and more possible factors on cross-sectional returns. Another possibility would be the grouping by industry as different industries have shown different behavior related to features (Zhu et al., 2023). The use of CARTs in econometrics is promising, as it offers new insights into the structure. Further development of guidelines and procedures is thus needed.

Acknowledgments

The authors thank Prof. Bernhard Schipp and the participants of the HERMES Seminar 2019 for their valuable comments and discussions.

Authors contributions

Not applicable.

Funding

This study was funded by the Saxon State Ministry of Education and Cultural Affairs.

Competing interests

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Redfame Publishing.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

References

Alpaydin, E. (2014). *Introduction to Machine Learning* (3rd ed.). *Adaptive Computation and Machine Learning series / Ethem Alpaydin*. Cambridge: MIT Press.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification And Regression Trees*. Routledge. https://doi.org/10.1201/9781315139470

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L., & Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology*, *108*(4), 299-307. https://doi.org/10.1093/oxfordjournals.aje.a112623

Bryzgalova, S., Pelger, M., & Zhu, J. (2019). Forest Through the Trees: Building Cross-Sections of Stock Returns. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.3493458

Canitz, F., Ballis-Papanastasiou, P., Fieberg, C., Lopatta, K., Varmaz, A., & Walker, T. (2017). Estimates and inferences in accounting panel data sets: comparing approaches. *The Journal of Risk Finance*, *18*(3), 268-283. https://doi.org/10.1108/JRF-11-2016-0145

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, 161-168. https://doi.org/10.1145/1143844.1143865

Chamberlain, G. (1980). Analysis of Covariance with Qualitative Data. *The Review of Economic Studies*, *47*(1), 225. https://doi.org/10.2307/2297110

Coqueret, G., & Guida, T. (2018). Stock Returns and the Cross-Section of Characteristics: A Tree-Based Approach. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.3169773

Dong, X., Li, Y., Rapach, D. E., & Zhou, G. (2022). Anomalies and the Expected Market Return. *The Journal of Finance*, *77*(1), 639-681. https://doi.org/10.1111/jofi.13099

Emmert-Streib, F., & Dehmer, M. (2019). Large-Scale Simultaneous Inference with Hypothesis Testing: Multiple Testing Procedures in Practice. *Machine Learning and Knowledge Extraction*, *1*(2), 653-683. https://doi.org/10.3390/make1020039

Fama, E. F., & French, K. R. (2008). Dissecting Anomalies. *The Journal of Finance*, *63*(4), 1653-1678. https://doi.org/10.1111/j.1540-6261.2008.01371.x

Farmer, L. E., Schmidt, L., & Timmermann, A. (2023). Pockets of Predictability. *The Journal of Finance*, *78*(3), 1279-1341. https://doi.org/10.1111/jofi.13229

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223-2273. https://doi.org/10.1093/rfs/hhaa009

Harvey, C. R. (2017). Presidential Address: The Scientific Outlook in Financial Economics. *The Journal of Finance*, *72*(4), 1399-1440. https://doi.org/10.1111/jofi.12530

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the Cross-Section of Expected Returns. *The Review of Financial Studies*, *29*(1), 5-68. https://doi.org/10.1093/rfs/hhv059

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). *The Elements of statistical learning: Data mining, inference, and prediction* (2nd ed., 12th repr). *Springer series in statistics*. New York: Springer.

He, X., Cong, L., Feng, G., & He, J. (2021). Asset Pricing with Panel Trees Under Global Split Criteria. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.3949463

Hoechle, D., Schmid, M., & Zimmermann, H. (2018). Do Firm Fixed Effects Matter in Empirical Asset Pricing? *European Financial Management Association, 2018 Annual Meetings*. Retrieved from https://www.cfr-cologne.de/download/kolloquium/2018/HoechleSchmidZimmermann.pdf

Hoetker, G. (2007). The use of logit and probit models in strategic management research: Critical issues. *Strategic Management Journal*, *28*(4), 331-343. https://doi.org/10.1002/smj.582

Isaksson, A., Wallman, M., Göransson, H., & Gustafsson, M. G. (2008). Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters*, *29*(14), 1960-1965. https://doi.org/10.1016/j.patrec.2008.06.018

Jagannathan, R., & Wang, Z. (1996). The Conditional CAPM and the Cross-Section of Expected Returns. *The Journal of Finance*, *51*(1), 3-53. https://doi.org/10.1111/j.1540-6261.1996.tb05201.x

James, G., Witten, D., Hastie, T. J., & Tibshirani, R. J. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). *Springer texts in statistics*. New York: Springer; Springer Science+Business Media.

Kozak, S., Nagel, S., & Santosh, S. (2017). Shrinking the Cross Section. *SSRN Electronic Journal*. Advance online publication. https://doi.org/10.2139/ssrn.2945663

Kozlowski, S., & Lytle, A. (2023). The January Anomaly and Anomalies in January. *Applied Finance Letters*, *12*(1), 2-10. https://doi.org/10.24135/afl.v12i1.615

Lewellen, J. (2015). The Cross-section of Expected Stock Returns. *Critical Finance Review*, *4*(1), 1-44. https://doi.org/10.1561/104.00000024

Long, W. J., Griffith, J. L., Selker, H. P., & D'Agostino, R. B. (1993). A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Research, an International Journal*, *26*(1), 74-97. https://doi.org/10.1006/cbmr.1993.1005

Markham, I. S. (2011). Assessing the prediction of employee productivity: a comparison of OLS vs. CART. *International Journal of Productivity and Quality Management*, *8*(3), 313. https://doi.org/10.1504/IJPQM.2011.042511

Moritz, B., & Zimmermann, T. (2016). Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns. *SSRN Electronic Journal*. Advance online publication. https://doi.org/10.2139/ssrn.2740751

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87-106. https://doi.org/10.1257/jep.31.2.87

Perlich, C., Provost, F., & Simonoff, J. (2003). Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. *Journal of Machine Learning Research*, *4*, 211-255. Retrieved from https://www.jmlr.org/papers/volume4/perlich03a/perlich03a.pdf

Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting*, *24*(1), 1-18. https://doi.org/10.1016/j.ijforecast.2007.07.008

Tsien, L., Christine, Fraser, F. S., Hamisch, & Long, J., William (1998). Using Classification Tree and Logistic Regression Methods to Diagnose Myocardial Infarction. *Studies in Health Technology and Informatics*, *52*.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28. https://doi.org/10.1257/jep.28.2.3

Zhu, L., Wu, H., & Wells, M. T. (2023). News-Based Sparse Machine Learning Models for Adaptive Asset Pricing. *Data Science in Science*, *2*(1). https://doi.org/10.1080/26941899.2023.2187895

Appendix

Table 1. Summary statistics

Variables	1st Quantile	Median	Mean	3rd Quantile
returns	0.041	0.008	0.0034	0.057
mom	-0.084	0.061	0.088	0.21
acc	-21.7	0.001	19.60	20.2
mc	10.68	12.68	13.12	16.00
btm	0.001	5.42	30.64	0.024
ns	-0.043	0.00061	0.029	0.12
dAA	-21.7	0.037	0.052	20.20
btmr	-0.026	0.064	20.20	0.148

Description: Summary Statistics, Datasource: Thomson Reuters Datastream

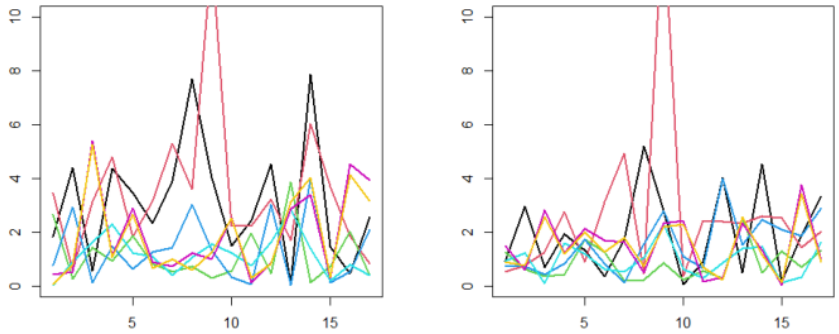

Figure 5. Z-score of parametric regressions with rolling window

Description: Right side: fixed effects regressions, right side: fixed effects logit regressions, y-axis = |z|, x-axis = period used, black = momentum, red = accruals, green =market capitalization, blue = book-to-market ratio, turquoise = net stock issues, purple = change in assets, yellow = profitability

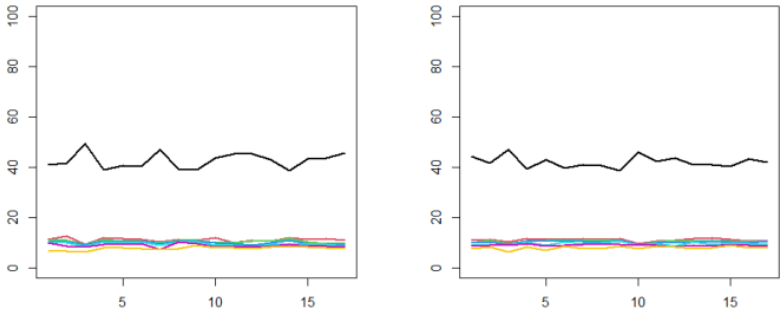

Figure 6. Variable Importance score of CARTs with backtesting

Description: Left side: regression trees, right side: classification trees, y-axis = variable importance, x-axis = period used, black = momentum, red = accruals, green =market capitalization, blue = book-to-market ratio, turquoise = net stock issues, purple = change in assets, yellow = profitability

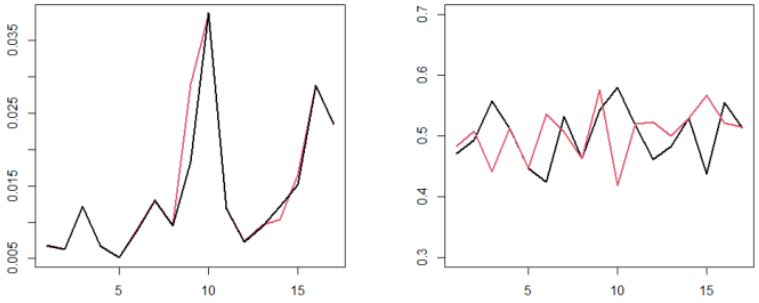

Figure 7. Precision metrics rolling window

Description: Left side: MSE, right side: MCR, y-axis = MSE respectively MCR, x-axis = period used, black = trees, red = regressions