The Benefits of a Tree-Based Model for Stock Selection in a South African Context

Mario Nicoló Giuricich

A dissertation submitted to the Department of Actuarial Science, Faculty of Commerce, at the University of the Cape Town, in partial fulfilment of the requirements for the degree of Master of Philosophy.

May 25, 2014

Master of Philosophy specializing in Financial Mathematics,
University of Cape Town,
South Africa.

Declaration

I declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Master of Philosophy in the University of the Cape Town. It has not been submitted before for any degree or examination in any other University.

_____

May 25, 2014

Abstract

Quantitative investment practitioners typically model the performance of a stock relative to its benchmark and the stock's fundamental factors in a classical linear framework. However, these models have empirically been found to be unsuitable for capturing higher-order relationships between a stock's return relative to a benchmark and its fundamental factors. This dissertation studies the use of Classification and Regression Tree (CART) models for stock selection within the South African context, with the focus being on the period from when the Global Financial Crisis began in early 2007 until December 2012. By utilising four types of portfolios, a CART model is directly compared against two traditional linear models. It is seen that during the period focused upon, the portfolios based on the CART model deliver the best excess return and risk-adjusted return, albeit in most cases modestly above the returns delivered by the portfolios based upon the linear models. This is observed in the hedge-fund style and long-only portfolios constructed. Moreover, it is observed that the CART-based portfolios' returns are not correlated with those from the linear-model-based portfolios. This observation suggests that CART models offer an attractive option to diversify model risk within the South African context.

Acknowledgements

I would like to sincerely thank my supervisor, Petrus Bosman, for all his help and input. My thanks also goes out to Dan Golding and Joel Wei from Prescient securities for the continued input and interest in my dissertation.

Finally I am deeply indebted to my mother - without her guidance and focus the work presented within this dissertation would have been impossible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the most explored topics in finance has been the predictability of stock returns. This topic is of significance not only within the realm of portfolio construction and management, but also in the fields of asset pricing as well as risk management. A former president of the American Finance Association, John. H. Cochrane, publicly addressed the topic of stock return prediction (Cochrane, 2011). Cochrane posited that modern finance had only begun to understand returns from a cross-sectional or time-series basis. By implication, a greater comprehension and therefore appreciation of stock returns is a young, developing field in finance. One of the broad aims of this dissertation is to add to the research in this young and developing field of finance.

For many years academics and financial practitioners alike have attempted to identify stock characteristics, from publicly available information, that influence their returns (Sorensen, Miller and Ooi, 2000; Zhu, Philpotts and Stevenson, 2012). Traditionally, when attempting to model stock returns, a linear relationship is assumed between the stock's forward returns and its fundamental characteristics. Long-established theories such as the Capital Asset Pricing Model (CAPM) (Fama and French, 1993; Carhart, 1997; Sharpe, 1964; Lintner, 1965), Arbitrage Pricing Theory (APT) (Ross, 1976) and the Fama-French three-factor model (Fama and French, 1993) seem to have inspired this linear assumption, as each of these models assume a linear relationship between a stock's or portfolio's mean return and the stock's fundamental characteristics. It must be noted, however, that the assumption of a linear relationship is not often observed in practice (Campbell, 1987; Fama and French, 1988). Therefore financial practitioners and academics using linear models to forecast stock returns and basing stock selection decisions on these models may need to research other generalised techniques of stock selection. As a consequence of this need, non-linear techniques for stock selection are beginning to be investigated by the investment community.

According to Zhu *et al.* (2012), there are two reasons for the attractiveness of non-

linear techniques. First, so-called unexplained profit opportunities may reasonably be assumed to arise if the true structural relationship between a stock's return and a stock's fundamental characteristics is indeed non-linear. Second, these non-linear techniques offer a broader model diversification from the more traditional linear approaches and could reduce the degree of commonality in portfolio holdings between investment managers.

In this dissertation, a non-linear technique for stock selection is applied to a wide universe of South African stocks. Specifically, a classification and regression tree (CART) is constructed and used in the process of stock selection. A CART is selected as the model of choice because it is non-parametric, is invariate to monotone transformations of independent variables, is not severely affected by outliers in the data used in the construction phase and does not need the modeler to specify the variables in advance (Breiman, Friedman, Olshen and Stone, 1984). Another benefit of a CART model is that it determines the interaction and formal hierarchy of stock characteristics. For example: does a value criterion prioritise over a momentum factor (Sorensen *et al.*, 2000)? The main focus of this dissertation is to compare, within the South African context, the performance of stock portfolios with constituent stocks selected by a CART model to that of portfolios with constituent stocks selected using a linear factor model. For each model, the portfolios initially investigated are simple, long-only ones. These investigations are naturally extended to form two portfolios for each model, that make use of both long and short positions in the stocks. The first is constructed using a 130-30 strategy and the second is a basic simulation of a hedge fund.

This dissertation is organised as follows: firstly, a review of the existing literature on the application of CART to stock selection is given. The main focus of the literature review is the research on US stocks by Zhu *et al.* (2012) - their paper served as an inspiration for this dissertation's research in South Africa. Secondly, a brief theoretical development of the CART methodology is presented together with a summary of linear weighting approaches. Following this theoretical exposition, the South African stock data is introduced. This is then followed by a description of how the methods described above are applied to stock selection in the South African market. The results are then presented: a CART and two linear models (one based on linear regression and the other on a mean-variance optimisation) are derived. In the final section of this dissertation, these three models are compared and assessed out-of-sample, and thereafter conclusions are drawn and presented.

Chapter 2

Literature Review

Most modern literature on the predictability of stock returns relies upon a linear regression approach to stock selection. The linear regression approach assumes a linear relationship between mean stock returns and the risk factors believed to explain those returns. Examples of such models include:

- The CAPM (Sharpe, 1964; Lintner, 1965),
- APT (Ross, 1976),
- The Fama-French, Three-Factor model (Fama and French, 1993), and
- The Cahart Four-Factor model (Carhart, 1997).

It must, however, be noted that there is no *a priori* motive suggesting that asset returns are linearly related to the risk factors. Today, there is an accumulating body of evidence indicating the need to research models allowing for a non-linear behaviour of the risk factors. Hsieh (1991) analysed American stock data and detected strong evidence against linearity in stock returns. Bansal, Hsieh and Viswanathan (1993) proposed an arbitrage pricing model that is non-linear, and highlighted that this model was more successful, when compared to APT-based models, in forecasting and explaining stock returns. Hiemstra and Jones (1994) researched the joint behaviour of trading volumes on the New York Stock Exchange and stock prices on the Dow Jones. They found that there exists non-linear behaviour between trading volumes and stock prices, and advocated that further research into these non-linear dynamics was necessary.

The Chartered Financial Analyst (CFA) Institute conducted a number of surveys of modeling approaches employed by large asset managers in Europe and the US (Fabozzi, Focardi and Jonas, 2008). The surveys found that the majority of techniques employed for financial modeling involved linear regression. Because of the extensive utilisation of similar data providers and linear models, Ang (2008) as well as Khandani and Lo (2011) observed an extensive similarity in quantitative

investor's trading strategies, especially before the so-called "quant-shock" that took place during July and August 2007. During the period 2007 to 2009, a large proportion of quantitative investors experienced poor, but similar, relative returns - this may be a consequence of the utilisation of common forecasting models.

The investigation conducted by the CFA Institute reported that only one-fifth of the large asset managers surveyed used non-linear methods of financial modeling (such as decision trees, random forests and neural networks). Because non-linear methods are not widely used, they are appealing because they seem to offer a large extent of model diversification compared to the traditional methods. Therefore, in view of the "quant-shock" previously referred to, non-linear methods appear to be important modeling tools for financial practitioners.

Many non-linear modeling approaches have been used in forecasting stock returns. In the US, neural networks, random forests and other non-parametric techniques have been employed. In South Africa, Bonga-Bonga and Makakabule (2010) applied a Smooth Transition Regression model to explain the smooth asymmetric response of stock returns on the Johannesburg Stock Exchange Securities Exchange (JSE) to macroeconomic variables. Another technique, that has not previously been employed in the South African context, is CART, and is possibly an ideal candidate to model stock returns because the technique is well suited to recognising complicated interactions within the data observations (Sorensen *et al.*, 2000).

Before the work of Sorensen *et al.* (2000) and Zhu *et al.* (2012), CART was not widely applied by quantitative investment practitioners. Frydman, Altman and Kao (1985) used CART to classify financially distressed companies. Sorensen, Mezrich and Miller (1998) developed a technique, based on a CART model, for tactical asset allocation. Thereafter Kao and Shumaker (1999) used a CART model to illustrate how equity-style stock rotations could be signaled, based upon changes in macroeconomic variables.

As was expressed in Cochrane (2011), a significant obstacle facing both investment practitioners and academics is the ability to discern the chief drivers of stock returns within a variety of stock-based variables. As posited by Breiman *et al.* (1984), CART is an instrumental technique to aid the modeler in selecting the most important explanatory variables that could influence the response. Therefore, tree-based models may be of interest when attempting to predict cross-sectional stock returns, for both investment practitioners and academics. The two papers that pioneered this work were those by Sorensen *et al.* (2000) and Zhu *et al.* (2012).

Sorensen *et al.* (2000) constructed a CART model using cross-sectional data on stocks comprising the USA's Russell 1000 index. The researchers specifically focused on technology stocks. The CART models constructed aimed to forecast whether a

stock's monthly return would outperform the median monthly market return over the period 1992 to 1995. The response variable was binary - the two outcomes being either outperformance or underperformance. The explanatory variables comprised a set of variables that the authors found to be most popular among money managers. They spanned profitability, price-momentum criteria, consensus earnings expectations and valuation. The first model Sorensen *et al.* (2000) constructed was a static CART model: the model was constructed using only data from 1992 to 1995. The CART model's primary split was on the change in consensus forecasted earnings. From their research, the authors concluded that out of all the possible explanatory variables analysed, the earnings estimate revisions could be the most powerful factors that explain the outperforming versus the underperforming stocks. The authors then went on to test this first model out of sample i.e. on the data from 1996 to 2000. Two portfolios were formed: firstly an equal weighting of those technology stocks expected to outperform and secondly an equal weighting of those expected to underperform. Without allowing for taxes and transaction costs, the authors found that on a monthly basis, stocks expected to be outperformers (as forecasted by the model) had an average excess return of 1.4 per cent per annum above those stocks expected to underperform. This difference was also seen to be statistically significant at a 95% level of confidence.

The second model Sorensen *et al.* (2000) constructed, termed the "evolving tree approach", differed from the first model. At each month during the period 1996 to 2000, the CART model was re-constructed using the latest available monthly data in addition to the data from 1992 to 1995. Therefore, for each month from 1996 to 2000 a different CART model was constructed. The evolving tree approach is more intuitive and sensible than the first model as it allows for gradual changes in market dynamics by including changes into the CART model over time. The primary split of this second model (as at October 1999) was also found to be the change in consensus forecasted earnings. After each of the evolving tree models were constructed, similar portfolios to those of the first model were built. The monthly return differential between the two portfolios was found to be slightly higher than the first model - an average excess return of 1.47 per cent per annum was found.

Sorensen *et al.* (2000) furthermore found that both CART-based stock selection methods produced portfolios with significantly higher returns as well as significantly better Sharpe ratios than portfolios produced by linear-based, single-factor stock selection models. Based on the Sharpe ratio, the authors also found that the evolving tree model performed the best, compared to the static CART approach and linear models.

Zhu *et al.* (2012) updated the work of Sorensen *et al.* (2000) by applying a CART model to a wider variety of stocks over a longer time period. The period considered spanned from December 1986 to August 2010 and almost all North American stocks were studied. The authors omitted financial stocks and stocks with market capitalisation less than $1 billion in 2010 (or its historical equivalent). Using response and explanatory variables similar to that of Sorensen *et al.* (2000), the authors constructed a single static CART model from data spanning December 1986 to April 2007. Their CART model's primary split was on a "value" variable: this variable comprised an equally-weighted average of five stock value metrics including book-to-price, sales-to-price, cashflow-to-price, dividends-to-price and earnings-to-price.

Zhu *et al.* (2012) then tested their CART model on data from April 2007 to August 2010, by constructing two portfolios similar to those that Sorensen *et al.* (2000) built. The authors specifically set out to test the performance of the CART-based portfolio over a time period associated with turbulence and uncertainty in stock returns. Moreover, for comparative purposes, Zhu *et al.* (2012) built two other portfolios in which the stocks in each of these portfolios were selected by a linear model: the first model being a simple multivariate linear regression and the second a simple mean-variance optimisation technique. Over the period from April 2007 to 2010, all three portfolios of stocks forecast to outperform were indeed found to outperform the median market return. However, the CART-based portfolio was found to have the highest outperformance of 2.6 per cent per annum (ignoring taxes and transaction costs), that was 1.6 per cent higher than each of the linear-based portfolios. The CART-based model also produced portfolios with the highest risk-adjusted returns. The authors reported that the CART-based portfolio delivered better performance without a higher turnover rate than the linear portfolios.

In view of all the research reviewed above, it is clear that a CART-model for South African stock data has to date not yet been constructed and reported on in the literature. However, portfolios based on other non-linear techniques of stock selection have been constructed within the South African context (see Bonga-Bonga and Makakabule, 2010). In the US context, portfolios constructed using CART models for stock selection have been shown to outperform median market returns. Therefore portfolios constructed using a CART-based model for stock selection, within the South African context, may be of interest to both quantitative investors and academics.

Chapter 3

Theoretical Development

CART is a subset of methodologies descended from a more general class of recursive partitioning algorithms (RPA). The purpose behind the CART intuition is to recursively partition a space of observations until all the resultant sub-spaces are homogeneous enough so that simple models can be applied to them (Zhu *et al.*, 2012). Over and above the recursive partitioning, CART models have other important features in common: firstly, an attempt at predicting the group membership of a dependent variable and secondly, the representation of a predictive process in the form of a "tree". Initially, CART was applied to medical diagnostics and prediction, but more recently researchers have employed the methodology to model the behaviour of financial markets. Kao and Shumaker (1999) used CART to solve a time series problem: the authors used a CART model to approximate a time series to distinguish between value stock and growth stock returns.

The CART methodology differs from simple linear regression where a distinct predictive formula is applied over the whole universe of observations. A CART model is not a global model and is better for modeling data containing multiple features that interact in a non-linear and complex manner, as often occurs in financial data sets (Andriyashin, Härdle and Timofeev, 2008). There are two important steps in constructing a CART model: firstly, a tree is constructed utilising a binary recursive splitting of nodes and secondly, the constructed tree is "pruned" so as to prevent over-fitting (Zhu *et al.*, 2012). Following the explication in Zhu *et al.* (2012), each of the two steps are discussed in detail below.

3.1 Binary recursive partitioning

Definition 3.1. (Breiman *et al.*, 1984, p.5) A **learning sample** comprises observations of the attributes (explanatory variables) and responses respectively $(\boldsymbol{x}_1, j_1), ..., (\boldsymbol{x}_N, j_N)$ on N cases where $\boldsymbol{x}_n \in \mathcal{X}$ and $j_n \in \{1, ..., J\}$, $n = 1, ..., N$. The learning

sample is denoted by $\mathcal{L}$, i.e.

$$\mathcal{L} = \{(\boldsymbol{x}_1, j_1), ..., (\boldsymbol{x}_N, j_N)\}.$$

Definition 3.2. (Breiman *et al.*, 1984, p.5) A variable is termed **numeric** or **continuous** if its measured values are real numbers. A variable is categorical if it takes values in a finite set not having any natural ordering.

The tree construction algorithm involves repeatedly splitting subsets of the learning sample $\mathcal{L}$ into two child subsets, commencing with $\mathcal{L}$ itself. For each node (i.e. a subset of the learning sample $\mathcal{L}$), a measure of the diversity of the observations within the node is required. The quantitative measure of diversity often employed by CART is the *diversity* or *Gini index*.

Definition 3.3. (modified from Breiman *et al.*,1984: p.38) The **diversity** or **Gini index** of any node is the probability that any two observations selected at random (with replacement) from those observations within the node will belong to separate groups [1]. When the dependent variable has q groups, the *diversity* or *Gini index* of any node is calculated as:

$$GI = 1 - \sum_{i=1}^{q} \rho_i^2 \tag{3.1}$$

where ρ_i, $i = \{1, ..., q\}$ is the proportion of observations within the node that belong to group i.

In order to construct the tree, the CART methodology proceeds as follows. The following steps are adapted from Wegner (2010):

1. Study each explanatory variable in turn.

2. For each of the explanatory variables, calculate every single split of the values of that independent variable (by the creation of two categories, C and C^C). For continuous variables, the allowable splits are of the form $x_i \geq f$ or $x_i < f$, for some constant f. For a categorical variable, the levels of that variable are separated into two classes. Hence for a categorical variable with D levels there will be 2^{D-1} allowable splits, ignoring order and the empty set Ω (Zhu *et al.*, 2012). These two classes will form the two child nodes.

3. For each of the two child nodes for each possible split, compute the diversity or Gini index: $GI_{\text{child 1}}$ and $GI_{\text{child 2}}$.

[1] The groups referred to here are the subsets one can form from the response variables: for example, a response variable could be an indicator variable comprising those stocks that outperformed their benchmark (group 1) and those that underperformed their benchmark (group 2).

4. Calculate the weighted average of the two diversity or Gini indices for each possible split, given by

$$WAGI = \frac{n_{\text{child 1}} GI_{\text{child 1}} + n_{\text{child 2}} GI_{\text{child 2}}}{n_{\text{child 1}} + n_{\text{child 2}}} \tag{3.2}$$

   where $n_{\text{child 1}}$ and $n_{\text{child 2}}$ are the number of observations in the first and second child nodes respectively.

5. The split to use in the tree is the one that leads to the greatest reduction in diversity from the parent node to the child nodes. Calculate the reduction in diversity, for each possible split, as

$$R = GI_{\text{parent}} - WAGI. \tag{3.3}$$

6. Repeat the above steps for each of the new child nodes.

The aim of CART is to ensure that the observations within each child node are maximally diverse: each node's GI should be maximised or at least near the theoretical maximum. The maximum value depends on the number of groups there are within each node.

Theorem 3.4. *Within a particular node, maximum diversity of the observations in the node will be achieved when the diversity or Gini index is equal to* $1 - \frac{1}{q}$.

Proof. In general, when q groups exist within a node, maximum diversity will arise when $\rho_i = \frac{1}{q}$ for all i. The diversity or Gini index therefore is:

$$\begin{aligned} GI &= 1 - q\frac{1}{q^2} \\ &= 1 - \frac{1}{q} \end{aligned}$$

□

3.2 Tree pruning

Classification trees are recursive algorithms. They begin with all explanatory variables in the root node (which is the very first node in the tree) and then split this root node into various child nodes. Each of these child nodes are further split into more nodes. In the extreme case, CART models can be split until there remains only a single case within each node. Having only one case in each of the bottom nodes is indeed an ultimate stopping condition for the tree as obviously the node cannot be partitioned further.

However, the splits further down the tree are likely to reflect specifics as well as random noise in the learning sample. Therefore, the CART model is unlikely to generalise to other sets of observations, which leads to the problem of 'overfitting' the model. The need arises to limit the tree size so that its findings are applicable to the broader population. Tree pruning is therefore a method to refine the robustness of the model: the method allows the tree to grow to full size, and then removes the nodes and branches of the tree that appear to lead to overfitting. According to Zhu *et al.* (2012), this is important if the model is being used for predictive purposes.

One of the more common techniques for pruning a tree is the pruning on misclassification error approach, as proposed by Breiman *et al.* (1984). This approach is considered to be useful when no testing or "hold-out" sample is available, and the learning sample is too small to have a hold-out sample extracted from it. The approach is based upon an assessment of misclassification error, termed the *cross-validation cost* (CV cost). The tree constructed by the binary RPA technique has an associated size or *complexity*, calculated by the tree's number of terminal nodes, and an associated CV cost, which is calculated as follows:

1. Divide the learning sample into K samples preferably equal in size.

2. Designate a particular 'size' for the CART model, commencing from the root node ($n = 1$) to the number of terminal nodes in the full un-pruned tree.

3. Construct the CART model, of a pre-specified size, K times and each time, omit one of the K sub-samples from the calculations and use it as a hold-out sample in order to compute a misclassification rate. For each node, the misclassification rate is defined to be one minus the overall hit rate.

4. Then calculate the average misclassification rate (i.e. the CV cost) for the tree, along with its standard error.

   The CV cost provides an approximation of a hold-out sample misclassification rate. In most circumstances when constructing a CART, it is advisable to select the tree with the minimum CV cost (Breiman *et al.*, 1984). Therefore, the following rule is used to stop the growth of the tree:

   Definition 3.5 (Pruning on CV cost rule)**.** (Breiman *et al.*, 1984, p.284): Select the tree with the fewest terminal nodes that has a CV cost equal to or less than

   $$\min(\text{CV cost}) + \Psi \times \text{standard error of tree with } \min(\text{CV cost}) \tag{3.4}$$

   where the term min(CV cost) is the smallest CV cost from all the possible CARTs that can be constructed and Ψ is a complexity parameter that is pre-

specified by the modeller. Note that Ψ is often considered to be the cost of each extra node added to the tree.

Note also that if $\Psi = 0$, then the tree with the lowest CV cost will be selected. Also, the higher the value of Ψ, the smaller the resulting tree is. In practice, a cross-validation sample is often used in determining the optimal value of Ψ, and a 30 percent cross-validation sample was used to calculate Ψ in this dissertation.

In sum, if properly used CART models are not black boxes. CART has a high degree of interpretability. It is able to compress a large set of data observations into a simple graphical format that identifies the important characteristics of the data. Therefore, this may add to the usefulness of CART in financial modeling.

3.3 Linear weighting approaches compared to CART

As discussed in the introduction, many modern stock selection models take the form of a linear weighting of factors that capture essential stock characteristics. For example, a traditional m factor linear model for stock returns (r) is of the form (Zhu *et al.*, 2012):

$$r = \alpha + \beta_1 f_1 + ... + \beta_k f_k + \epsilon \tag{3.5}$$

where α is a term specific to the stock being modeled, f_i and β_i , for $i = \{1, ..., k\}$, are the factors and the sensitivities to the factors respectively and ϵ is an error term. Most classical asset pricing theories such as the CAPM and APT follow this structure.

The linear approach may be convenient, simple and intuitive, however, it often does not fully gauge the level of complexity within the stock market between the factors. Indeed, linear approaches have weaknesses (Zhu *et al.*, 2012), such as:

- the assumption that stock returns react linearly to changes in predictor variables. Empirical studies, such as Hsieh (1991) and Shively (2003), suggest that this assumption is often breached in practice.

- the assumption that the underlying response variable follows a normal distribution. However, there exists much empirical evidence that return distributions are indeed not normal and therefore cannot be solely characterised by the first two moments. Much research, including Harvey and Zhou (1993), Cont (2001), Hueng and McDonald (2005) and Post, Van Vliet and Levy (2008) has found asymmetries in the distributions of empirical stock returns.

- the observation that traditional linear models may be affected by multicollinearity, missing values as well as outliers. Moreover, linear models are often not able to identify interactions between pertinent explanatory variables in the model, especially when the data set is noisy.

In view of the above, it is evident that CART offers a number of benefits. CART is a non-parametric model and does not require any distributional assumptions for the variables being modelled, in contrast to traditional linear models. CART models are also less affected by missing data and multicollinearity, and are robust towards outliers and noisy data sets (Breiman *et al.*, 1984). The latter two are particular features of financial data (Sorensen *et al.*, 2000).

CART models also differ from traditional linear modeling methodologies in that the former constructs a hierarchy of input variables that could be more commensurate with human decision-making processes. A significant advantage of CART models over the linear modeling techniques is that CART permits the modeller to show the various interactions that exist between explanatory variables, as well as the conditional relevance of variables (Van der Smagt and Lucardie, 1991). Generally speaking, conditional relevance arises when a factor (i.e. explanatory variable) is significant (in determining the response variable from a statistical model) only when this factor is conditioned upon another factor (i.e. another explanatory variable).

However, CART does have some weaknesses. A significant criticism of CART is the recursive nature of the tree construction process. According to Breiman *et al.* (1984), local optimisation at each step in the sequential node-partitioning method will not necessarily produce a global optimisation of the overall tree. Put differently, in constructing the hierarchy of rules that comprise the tree, the CART algorithm does not account for the nature of the branches and nodes further down into the tree. Breiman (2001) has developed an alternative that overcomes this aforementioned problem - the random forest approach.

Another major criticism of CART is the fact that the method discretises continuous variables. CART discretises continuous variables based upon a rule of the form $x_i \geq f$ or $x_i < f$. There are two consequences of this that are discussed below. For illustrative purposes, let the response variable be an indicator for outperformance of a stock relative to the median market return, and the explanatory variables be the price-to-book ratio and indicator variables each for loss-making companies and for technology companies. Notice that the former variable is continuous while the latter two are categorical.

- Firstly, the response variables are not sensitive to changes in a continuous variable (price-to-book ratio in this case) forming the splitting rule of a node. All observations within the node share the same probability of outperforming

the median market return regardless of the variation within the continuous variable. The consequence of this, overall, is that the tree leads to resultant discontinuous probabilities.

- Secondly, the discretisation leads to the over-sensitiveness of response variables to a continuous variable near a boundary. Continuing with the illustrative example, this observation implies that a small change in the value of the price-to-book ratio could lead to a disproportionate change in the response variable (Zhu *et al.*, 2012).

On balance, even though discretising continuous variables and the assignment of the same output to all observations within a node could be considered to be parsimonious, it could oversimplify the true complexity of the data observations. Moreover, cognisance needs to be taken of the fact that extreme shifts from one node to another, caused by the discretization, may not necessarily be realistic. In order to overcome these problems, Zhu, Philpotts, Sparks and J. Stevenson (2011) suggested the use of a "hybrid" model that combines both CART and logistic regression. The hybrid model could both uncover the non-linearities in the data observations and produce a smooth probability surface. The hybrid model will not be investigated in this dissertation, but does provide the opportunity for further research.

Chapter 4

Data and Methodology

4.1 Data

Monthly cross-sectional stock data from January 2000 to December 2012, for 84 of the top 100 stocks listed on the JSE were used. All Rand denominated data was adjusted to December 2012, using the Consumer Price Index, in order to remove possible inflationary effects. In total 118 086 observations, derived from the universe of 84 stocks, were employed. It must be noted that all analyses for this dissertation were performed in Mathworks MATLAB as well as Microsoft Office Excel.

Each stock studied had a market capitalisation of at least $R1$ billion in 2012 (or its historical equivalent) in order to ensure a representation of liquid companies. The data was sourced from numerous data providers including Bloomberg, I-Net, McGregor BFA and Thompson Reuter's Datastream.

For this research, it was necessary to decide upon fundamental stock metrics that could be employed to forecast future stock returns. Such stock metrics that had been found in historical studies were employed. The following broad categories for the stock metrics, posited by Zhu *et al.* (2012), were focused upon in this dissertation:

- **Profitability:** Campbell and Thompson (2005) and Chen, Novy-Marx and Zhang (2011) highlighted the predictive abilities of a stock's profitability metrics when attempting to forecast stock returns. The profitability ratios employed are set out in Table 4.1.

- **Financial strength:** These stock metrics provide an indication of a firm's ability to manage and sustain its debt. The financial strength metrics employed are listed in Table 4.1. Empirical studies by Bhandari (1988) and Campbell, Hilscher and Szilagyi (2008) found these metrics to be useful in forecasting stock returns.

- **Value:** This category quantifies a firm's cashflows, dividends, turnover and book values. Lewellen (2004) found significant empirical evidence that divi-

dend yields can forecast stock returns. Also Lakonishok, Shleifer and Vishny (1994) reported that firms' cashflows possibly vary in line with the systematic variation in their stock returns.

- **Momentum:** Momentum factors are metrics relating to a stock return's past performance. Bondt and Thaler (1987) as well as Jegadeesh and Titman (1993) documented that future stock returns could be forecasted using past stock returns.

- **Financial analysts' forecasts:** These metrics comprise analysts' consensus forecasts as well as revisions in their forecasts. Trueman (1994) reported that these forecasts are widely used when forecasting stock returns. Also, earnings revisions have received attention in the literature. Lys and Sohn (1990) reported that investors who consistently abide by analysts' earnings revisions are able to steadily outperform a traditional long-and-hold strategy.

Table 4.1 displays a list of nine categories of stock metrics, of which all nine were used in this dissertation in an attempt to forecast the universe of stocks' returns. The metrics comprising each of the nine categories are similar to those employed by Zhu *et al.* (2012). In this dissertation, a total of 21 stock metrics were used to construct the categories. For ease of expression, the categories of the stock metrics will simply be referred to, for the remainder of this dissertation, as factors. Each of the nine factors comprised an equally weighted average of its constituent stock metrics, as shown in Table 4.1. This was done in order to overcome the potential correlations between the 21 metrics.

The nine factors were calculated, at each month, for each of the 84 stocks. Table 4.2 shows the Spearman rank correlation matrix for the nine factors; most of the correlations lie between -0.2 and 0.1.

Category	Stock metrics
Value (VAL)	Earnings to price, dividends to price, cashflow to price, Sales to price, book to price
Profitability (PROF)	Return-on-equity, pre-tax margin, asset turnover
Leverage (LEVERAGE)	Debt-to-equity, debt-to-market cap
Debt service (DEBT.SERVICE)	Interest cover, free cashflow to debt
Momentum (MOM)	Relative Strength Index (RSI) 14 day
Stability (STAB)	Volatility in corporate earnings, sales and cashflows
Historical growth (HIST.GROWTH)	3 year historical growth in earnings, sales and cashflows
Forward growth (FWD.GROWTH)	Median broker forecasts of EPS two years ahead
Earnings change (EREV)	Change in median broker borecasts for earnings

Tab. 4.1: Input variables (factors) into each of the three models together with the factors comprising them. Note that each category of stock metrics (or factor as it will be termed in this dissertation) comprises an equally weighted average of the stock metrics constituting it - for example, the LEVERAGE factor comprises an equally weighted average of debt-to-equity and debt-to-market cap.

	STAB	FWD.GROWTH	DEBT.SERVICE	HIST.GROWTH	PROF	EREV	VAL	MOM	LEVERAGE
STAB	1.00	-0.11	-0.45	-0.10	-0.09	-0.02	0.02	0.05	-0.01
FWD.GROWTH	-0.11	1.00	-0.58	-0.21	-0.10	0.02	0.02	0.02	0.01
DEBT.SERVICE	-0.45	-0.58	1.00	-0.28	0.01	0.01	-0.14	-0.02	0.05
HIST.GROWTH	-0.10	-0.21	-0.28	1.00	-0.14	-0.18	0.06	-0.06	-0.05
PROF	-0.09	-0.10	0.01	-0.14	1.00	-0.39	-0.06	-0.05	-0.05
EREV	-0.02	0.02	0.01	-0.18	-0.39	1.00	-0.16	-0.05	-0.08
VAL	0.02	0.02	-0.14	0.06	-0.06	-0.16	1.00	-0.15	-0.29
MOM	0.05	0.02	-0.02	-0.06	-0.05	-0.05	-0.15	1.00	-0.73
LEVERAGE	-0.01	0.01	0.05	-0.05	-0.05	-0.08	-0.29	-0.73	1.00

Tab. 4.2: Spearman rank correlation matrix for all the factors calculated over the entire stock universe.

In order to refine the robustness of the analysis, rank orders were used for each of the factors. Following the methodology of Zhu *et al.* (2012), absolute factor values were not used. For each month and for each of the nine factors, the stocks were ranked from largest to smallest and subsequently the rank order for each stock was found. To reiterate, the ranking was done on a per-factor per-month basis. This rank was then divided by the total number of stocks (i.e. 84) in order to provide a scale between 0 and 1. It must be noted that some of these factors only change 2 to 4 times a year - in these cases, the same value for the factor was employed in consecutive months until it changed.

Moreover at each month-end, forward stock returns inclusive of dividends were computed for each of the stocks. The forward stock returns were calculated according to Definition 4.1 below.

Definition 4.1 (Forward Stock Return)**.** If Y_t denotes the Rand stock price at time t observed from the data and Y_{t+1} denotes the Rand stock price at time $t+1$ observed from the data, the forward stock return [1] is given by

$$\frac{Y_{t+1} - Y_t + D_{t,t+1}}{Y_t} \tag{4.1}$$

where $D_{t,t+1}$ denotes the Rand value of the dividends received between times t and $t+1$.

4.2 Linear weighting approaches

The purpose of this dissertation is to highlight the benefits of utilising CART methodologies over the more traditional linear modeling approaches. Two versions of a linear model were constructed and these models were compared with the CART model. Each of the models were constructed using data on the 84 stocks from January 2000 to April 2007. This was deliberate as the period from April 2007 onwards contained the Global Financial Crisis (Zhu *et al.*, 2011).

Before the linear models were constructed, it was necessary to calculate the excess stock returns. The excess return was calculated, for each month, as the forward stock return less the return on an appropriate benchmark. Since 84 stocks from the top 100 on the JSE were used, it was decided to construct a benchmark using the top 100 stocks on the JSE. An arithmetically weighted average return, using market capitalisations as weights, was calculated for each month using total stock returns on the top 100 JSE shares and this was used as a benchmark.

[1] The forward stock return is the modelled or forecasted return for the stock for the next time period $(t, t+1)$.

In the first model, the sensitivities β_i for each of the factors were derived from a multivariate regression of one month excess stock returns on the nine factors. Note in this model that the explanatory variables were the nine composite factors per-stock per-month and the response variable was the one-month excess stock return. Equation (4.2) below was used in the regression.

$$\begin{aligned} ER_{i,t+1} = \alpha &+ \beta_1 VAL_{i,t} + \beta_2 PROF_{i,t} + \beta_3 EREV_{i,t} \\ &+ \beta_4 MOM_{i,t} + \beta_5 LEVERAGE_{i,t} + \beta_6 STAB_{i,t} \\ &+ \beta_7 FWD.GROWTH_{i,t} + \beta_8 HIST.GROWTH_{i,t} \\ &+ \beta_9 DEBT.SERVICE_{i,t} + \epsilon_{i,t} \end{aligned} \tag{4.2}$$

where ER_t denotes the return of the i'th stock over month t, β_j denotes the sensitivity of the excess return to factor j (for $j = \{1, 2, ...9\}$) and $\epsilon_{i,t}$ is an error term.

The sensitivities were then converted to a weight such that the total weights for each of the factors summed to 100. Both positive and negative weights were permitted.

In the second model, a mean-variance optimization technique was employed. This technique was also employed by Zhu *et al.* (2012). For each of the nine factors

$$\begin{aligned} k = \{&VAL, PROF, EREV, MOM, LEVERAGE, \\ &STAB, FWD.GROWTH, HIST.GROWTH, DEBT.SERVICE\}, \end{aligned}$$

the factor-relative historical changes (HR_k) and the volatilities thereof (γ_k) were calculated over a period of time (i.e. January 2000 to April 2007). Then, Markowitz's mean-variance optimisation theory was applied to equations (4.3) and (4.4) below to find an empirically optimal weighting scheme:

$$\mathbb{E}[R_p] = \sum_{k=1}^{9} w_k HR_k \tag{4.3}$$

$$\gamma_p^2 = \sum_{k=1}^{9} w_k \gamma_k^2 + \sum_{k=1}^{9} \sum_{m \neq k}^{9} w_k w_m \gamma_k \gamma_m \rho_{km} \tag{4.4}$$

where $\mathbb{E}[R_p]$ is the sum of the equally-weighted factor-relative historical changes of all 84 stocks over the time period studied, γ_p is the volatility of the sum of the equally-weighted factor-relative historical changes, γ_k is the volatility of the factor-relative historical changes on the k'th factor and where ρ_{km} denotes the correlation between the k'th and m'th factors. The optimisation technique included maximising $\mathbb{E}[R_p]$ subject to minimising γ_p

The w_k's were then calculated subject to the restriction that they summed to one. Again, both positive and negative weights were permitted.

As an extension to their analyses, both Sorensen *et al.* (2000) and Zhu *et al.* (2012) constructed linear models for industry sub-sectors in order to assess the models' effectiveness on an industry level. Both a multivariate linear regression model and a mean-variance optimization model were constructed for 3 sub-sectors of the JSE top 100 shares, namely resources, financials and industrials using 10, 16 and 24 stocks respectively. The respective indices used as benchmarks were the JSE's RESI, FINI and INDI indices.

4.3 CART model

The CART model was constructed using the same data that was used for the linear models. This non-linear model was built and pruned by applying the methodology outlined in Chapter 3. The purpose of the CART model was not to model excess returns as in the case of the linear models, but to forecast whether a stock would be an outperformer (OUT) or an underperfomer (UND) relative to its benchmark. So instead of using excess returns, a categorical variable - with 1 denoting OUT and 0 denoting UND, was employed in the CART model as the response variable. The nine factors were used as explanatory variables. In addition it must be noted that the CART model provided, for each stock, the probability of outperformance.

CART models were also constructed for the three sub-sectors of the JSE top 100 shares given above. The data used was the same as that for the linear models.

4.4 Portfolio construction for model comparison

The efficacy of the three models was then assessed out-of-sample. Stock data from May 2007 to December 2012, was used to calculate the nine factors as well as forward stock returns for each month. These factors, converted into ranks (as outlined in Section 4.1 above), were then used as inputs into the three models in order to acquire fitted values for each. Note that data from January 2000 to April 2007 was used to construct the models. Even though Sorensen *et al.* (2000) found evolving models to perform the best, this dissertation did not construct models that were updated with data as time progressed (i.e. the evolving models). This omission does open up further research into evolving linear and CART models for South African stock data.

Firstly, in order to assess the extent of correlation between the outputs of the three models, all stocks in each month were ranked based upon the fitted values. Secondly, for each of the three models, three portfolio strategies were used to compare their performance. The following three portfolios were formed:

- **Long portfolio:** a portfolio comprising the stocks forecasted to outperform, with weights given by, in the case of the linear models, the excess return forecasted by the model and, in the case of the CART model, the probability of outperformance.

- **Short portfolio:** a portfolio comprising the stocks forecasted to underperform, with weights given by, in the case of the linear models, the excess return forecasted by the model and, in the case of the CART model, the probability of outperformance.

- **130-30 strategy:** a portfolio going short in 30 per cent of its stocks, going long in 30 per cent and investing 100 per cent in the portfolio's benchmark. The stocks comprising the 30 per cent long and 30 per cent short components are sourced from the top 10 stocks forecasted to outperform by the model and top 10 stocks forecasted to underperform by the model, respectively. A 130-30 strategy was specifically selected as it is most widely used in industry (Johnson, Ericson and Srimurthy, 2007). The weights of each of the stocks are found in the same way as for the long and short portfolios.

- **Hedge fund-style strategy:** a portfolio comprising both short and long positions in the top 10 shares forecasted to outperform by the model and top 10 shares forecasted to underperform by the model. The weights of the shares were found in exactly the same way as was done for the long and short portfolios above. A gearing ratio of 3 was used (*Symmetry: South African Hedge Fund Survey.*, 2013): this implies that for every 1 per cent in return from the positions taken, the portfolio return will be 3 per cent.

All four portfolios above were formed for the models based on the universe of 84 stocks, using the top 100 total return as a benchmark. Transaction costs were not accounted for. In addition, the long and short portfolios were also constructed for each the three models based on the FTSE/JSE sub-sectors, in each model using the FTSE/JSE Resources 10 Index (RESI), the FTSE/JSE Industrial 25 Index (INDI) and the FTSE/JSE Financials Index (FINI) as the benchmark returns. These portfolios were constructed to assess whether the CART model was superior, compared to the linear models, for each of the three sub-sectors.

To ascertain the performance of these portfolios, various performance measures were calculated for the period May 2007 to December 2012. The measures included the following:

- **Excess portfolio return:** this measure is concerned with the excess of the annual portfolio return above the benchmark return. Mathematically, the

excess portfolio return can be calculated as (Zhu *et al.*, 2012):

$$ER = \left[\Pi_{t=1}^{n}\left(1 + r_{Pt}\right)\right]^{\frac{12}{n}} - \left[\Pi_{t=1}^{n}\left(1 + r_{Bt}\right)\right]^{\frac{12}{n}} \tag{4.5}$$

where $n = 68$ (months), r_{Bt} and r_{Pt} are the monthly returns of the benchmark and the portfolio at time t respectively (for $t = \{1, 2, ...68\}$).

- **Tracking error:** this measure is annualised and is calculated as (Zhu *et al.*, 2012):

$$TE = \sqrt{12\mathbb{V}\left(r_{Pt} - r_{Bt}\right)} \tag{4.6}$$

where $\mathbb{V}\left(r_{Pt} - r_{Bt}\right)$ denotes the variance of the excess return of the portfolio.

- **Sharpe ratio:** this refers to the annualised excess return of the portfolio above the risk free rate, per unit of portfolio standard deviation. Mathematically, the Sharpe ratio can be calculated using equation (4.7) (Sharpe, 1975):

$$SR = \frac{\left[\Pi_{t=1}^{n}\left(1 + r_{Pt}\right)\right]^{\frac{12}{n}} - \left[\Pi_{t=1}^{n}\left(1 + r_{Ft}\right)\right]^{\frac{12}{n}}}{\sigma_p} \tag{4.7}$$

where r_{Ft} denotes the risk free rate at time t, $n = 68$ (months) and σ_p denotes the portfolio standard deviation over the period May 2007 to December 2012. A proxy to the risk free rate was used, this being the 3-month negotiable certificate of deposit (NCD) rate. According to Firer and Staunton (2002), the 3-month NCD rate is suitable as a proxy for a short-term default-free rate.

- **Information ratio:** This ratio is the annualised mean of the excess portfolio return over the annualised tracking error, given by (modified from Goodwin, 1998):

$$IR = \frac{ER}{TE} \tag{4.8}$$

with ER and TE as defined by equations (4.5) and (4.6) respectively.

- **Sortino ratio:** This ratio is the annualised excess return of the portfolio above the risk free rate, per unit of portfolio downside standard deviation. The Sortino ratio can be calculated by equation (4.9) (Sortino and Van Der Meer, 1991):

$$STR = \frac{\left[\Pi_{t=1}^{n}\left(1 + r_{Pt}\right)\right]^{\frac{12}{n}} - \left[\Pi_{t=1}^{n}\left(1 + r_{Ft}\right)\right]^{\frac{12}{n}}}{\sigma_{dp}} \tag{4.9}$$

where $n = 68$ (months) and σ_{dp} is the downside standard deviation of the portfolio.

- **Holding period:** The holding period is a metric that assesses a portfolio's stock turnover. By implication, it may be that the shorter the holding period for a single stock, the higher the portfolio transaction costs and portfolio turnover. The holding period for a stock is calculated as the average of the total numbers of consecutive months that a stock is held in a portfolio in the out-of-sample test data (Zhu *et al.*, 2012). For example, a stock is held for 5 months, then later on in the analysis the stock is held for another 7 months. Hence, the stock holding period in this case is 6 months. The holding period was included as a performance measure to provide some indication of the level of transaction costs when managing the portfolio. Intuitively, the higher the portfolio turnover, the higher the transaction costs as stocks are moved into and out of the portfolio more frequently.

Finally, a measure of the models' effectiveness in forecasting outperformers is required. Using the out-of-sample data, a "hit rate" is calculated. The number of times that the models correctly identify a stock as being an outperformer or an underperformer is calculated and expressed in percentage form.

Chapter 5

Discussion

5.1 Linear weighting and CART approaches

The first model constructed for the universe of 84 stocks was a multivariate linear regression of 1 month excess returns on the nine factors. The weights from this regression are shown in Table 5.1 below. It is evident from Table 5.1 that this regression-based weighting model favours Stability, Debt service, Forward Growth, Historical Growth, Profitability and Earnings Revision, although it slightly penalises stocks with high leverage.

The second model constructed for the universe of 84 stocks used the volatility and the historic return of each factor in a mean-variance optimisation to generate an optimal weighting scheme. The weights from this optimisation scheme are also shown in Table 5.1 below. In a similar way to the linear regression model, the mean-variance model favoured the Value and Leverage factors. Unlike the first model, the second placed a marked emphasis on Value and Leverage.

It is evident that both models placed some emphasis on Value factors, especially the mean-variance model. The emphasis on Value is consistent with the traditional focus of most quantitative investors and managers (Zhu *et al.*, 2012).

The third model that was developed for the universe of 84 stocks was the CART model. Figure 5.1 below shows the hierarchical structure of the CART model constructed using information up to and including April 2007. Following the methodology of Zhu *et al.* (2012), in the CART model shown in Figure 5.1 a stock is classified as an outperformer (OUT) if, according to the model, it has a 50 per cent or greater chance of outperforming, and an underperformer (UND) otherwise.

It is evident from Figure 5.1 that the primary split is on Value: the CART model therefore distinguishes between those stocks having a high Value factor and a low Value factors. As is also evident in Figure 5.1, an attractive node splits on Momentum: those stocks with a 14-day RSI greater than 82.446 have a 58.3 per cent chance of outperforming. Momentum does appear to be an important deciding

	Regression-based weights (%)	Mean-variance weights
STAB	18.80	-24.11
DEBT.SERVICE	18.51	0.05
FWD.GROWTH	17.53	-0.01
HIST.GROWTH	15.76	-6.56
PROF	13.57	-4.24
EREV	12.57	-1.38
VAL	6.26	20.00
MOM	-0.85	-3.18
LEVERAGE	-2.17	20.43

Tab. 5.1: The linear weights for the linear regression model and the mean-variance model, using data from January 2000 to April 2007.

factor in the CART model: all stocks forecasted to outperform have a 14-day RSI of at least 43.

Another important deciding factor in the CART model appears to be Stability. Stocks with low Stability (i.e a high volatility in their firm's earnings, sales and cashflows over the previous 5 years) as well as high Value seem to outperform. Stocks with high Value and both low Stability as well as low Momentum do not appear to outperform (Node 8).

An important benefit of using a CART model is conditional relevance. The mean-variance model presented above identifies Value as being the chief determinant of stock returns. However, the CART model suggests that stocks with low Value still have a good chance of outperforming the benchmark provided they possess strong Momentum (Node 4).

Finally, the CART model suggests the possibility of a complex and non-linear relationships between the underling factors. In order to reach the outperforming nodes (Nodes 4, 7, 9, 10, 12, 13, and 14) most often, a decision based on Value, Stability and Momentum factors is required more than once. For example, in order to reach node 14, two decisions based upon the Momentum factor are needed. It must also be noted that Value, Stability and Momentum are commonly used by many quantitative investors and managers when making stock selection decisions (Zhu *et al.*, 2011).

Linear regression, mean-variance and CART models were also constructed for each of the 3 sub-sectors (resources, financials and industrials) for the universe of 84 stocks. These models are presented in Tables A.1, B.1 and C.1 as well as Figures

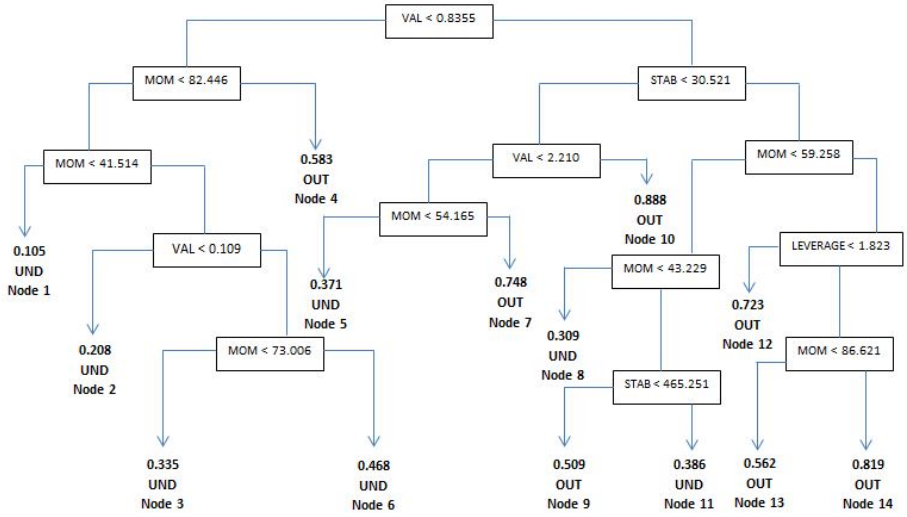

Fig. 5.1: CART model for the universe of the stocks, constructed using data from January 2000 to April 2007. This decision tree models the probability of a stock outperforming the benchmark (in this case, the benchmark is a market-capitalisation weighted average of the returns on the top 100 stocks on the JSE). The dependent variable is depicted to be an outperformer (OUT) if the stock achieves a positive excess return, and an underperformer (UND) otherwise. Note that at each node, should the factor be less than the deciding value (for example at the primary node if VAL $<$ 0.8355), then the decision tree will flow to the left-hand side. Finally, for a specific terminal node (i.e. nodes 1 to 14) the probability of a positive excess return is reported.

A.1, B.1 and C.1 in the appendices. On inspection, it is evident that the main factors upon which these CART models for these sub-sectors places emphasis are Value, Momentum and Stability.

5.2 Model Comparison

The efficacy of the three models in terms of their stock selection power was tested out-of-sample for the stock universe. For each month in the out-of-sample period (May 2007 to December 2012), all stocks were ranked based upon the fitted values provided by the three models. It is important to note that a historical period, associated with relatively poor stock performance for most quantitative investors and managers, was covered. Table 5.2 below reports the Spearman rank correlation

	Regression	Mean-Var	CART
Regression	1.00	0.11	-0.01
Mean-Var	0.11	1.00	-0.03
CART	-0.01	-0.03	1.00

Tab. 5.2: Spearman rank correlation matrix reporting the cross-sectional correlation between each of the model outputs out-of-sample (from May 2007 to December 2012).

matrix for the model forecasts during this turbulent period.

It is evident from Table 5.2 that there exists a low degree of positive correlation between the two linear models. However, it is evident that there is very little correlation between the linear and non-linear models, with the non-linear CART model presenting with a rank correlation of -0.01 and -0.3 for each of the two linear models. Therefore, it appears that the CART model shows diversification from the traditional linear models when it comes to stock selection. Similar observations were noted when the analysis was split by the three sub-sectors of the stock universe.

Furthermore, the hit rate calculated provides a measure of the efficiency of the three models in terms of selecting outperformers. The hit rate measured the accuracy of the three models in correctly identifying outperforming and underperforming stocks in the testing period. For the linear regression, mean-variance and CART models the hit rates were 59, 63 and 66 per cent respectively. These results seem to suggest that the CART model is slightly more accurate at identifying outperformers and underperformers when compared to the linear models. On closer inspection, it appeared that the CART model more accurately identified outperformers compared to underperformers - 72 per cent of the stocks observed to outperform were forecasted, by the CART model, to outperform. Only 42 per cent of those stocks that were observed to underperform were forecasted, by the CART model, to underperform. Ultimately, it must be borne in mind that this hit rate calculated is only applied to one testing sample - to make more robust conclusions, it may be necessary to calculate the hit rate for a number of samples.

According to Zhu *et al.* (2012), the true test of a model is often characterised by the model's performance out of sample. This was evaluated by forming the four portfolios introduced in Section 4.4 of Chapter 4. Table 5.3 on the next page reports the annualised excess stock return, the tracking error, the Sharpe ratio, the information ratio, the Sortino ratio and the portfolio holding period information for each of the various strategies.

Portfolio	Model	Excess Return (%)	Tracking Error (%)	Sharpe Ratio	Information Ratio	Sortino Ratio	Holding Period (months)
Long	Regression	0.40	23.71	-0.16	1.67	-0.31	7.2
	Mean-Var	0.66	23.29	-0.17	2.85	-0.22	7.4
	CART	1.02	20.73	-0.10	4.91	-0.15	7.5
Short	Regression	0.26	19.99	-0.46	1.30	-0.68	5.7
	Mean-Var	0.27	23.28	-0.17	1.17	-0.23	6.8
	CART	0.18	20.08	-0.53	0.90	-0.78	6.9
130-30 strategy	Regression	0.38	5.02	-0.16	7.25	-0.22	4.3
	Mean-Var	0.02	7.43	-0.23	0.22	-0.33	4.9
	CART	0.39	5.38	-0.17	6.68	-0.24	4.2
Hedge-fund style	Regression	0.61	38.98	-0.07	1.55	-0.09	4.1
	Mean-Var	-0.97	34.21	-0.02	-0.23	-0.03	3.6
	CART	3.70	29.47	0.37	12.53	0.56	3.1

Tab. 5.3: Annualised excess returns, tracking errors, Sharpe ratios, Information ratios, Sortino ratios and holding period calculated for each of the four portfolios. Note that transaction costs were not accounted for and the portfolios were rebalanced on a monthly basis. Stock turnover within a portfolio is indicated by the holding period.

Each of the long portfolios with their constituent stocks selected on the basis of a linear model outperformed the top 100 stocks benchmark, albeit a modest outperformance (0.40 and 0.66 per cent). It is also clear from Table 5.3 that each of the linear-based long portfolios' Sharpe ratios and Sortino ratios are negative. Given that the Global Financial Crisis occurred over most of the data testing period, the observation is not unusual: over much of this turbulent period the South African equity market underperformed risk-free investments.

The CART-based long portfolio outperformed better on both a non-risk adjusted basis (by considering the excess return) and risk adjusted basis (by considering the Sharpe and Sortino ratios), and had a lower tracking error than the linear models. It must be noted that high tracking errors observed for both the long and short portfolios, for each three models, are observed - a chief reason for this is that only between 14 to 36 stocks (out of the universe of 84 stocks) comprise each portfolio. It is also evident from Table 5.3 that the CART-based long portfolio had the highest information ratio, albeit negative. Therefore, it appears that the CART-based portfolio more consistently outperforms its benchmark, as compared to the linear-based approaches. The lower information ratios observed for the linear-based models are suggestive of positive yet erratic returns: on closer inspection of the monthly returns this was observed to be true for both the linear models, but the CART-based portfolio's outperformance was observed to be more consistent from month-to-month. Also, attention must be given to the observation that the higher Sharpe and Information ratios of the CART-based long portfolio were suggestive that the outperformance may not have been due to chance. Moreover, the CART-based portfolio has the highest Sortino ratio. Ultimately, this may be of importance for long-only investors since the risk of a capital loss may be lower when stocks are selected on the basis of a CART model compared to portfolios with stocks selected on the basis of either of the two linear models.

All these results seem to motivate the inclusion of CART-based methods of stock selection over the two linear-based methods within a long-only portfolio. However, before CART techniques are considered, cognisance of the fund manager's investment mandate should be accounted for.

On consideration of the short portfolios, positive excess returns are noted. The CART-based short portfolio does provide the lowest return when compared to the CART-based long portfolios, although it must be noted that the former is positive. Upon further analysis, it became clear that within each month, the CART model was selecting both stocks that underperformed and outperformed the benchmark. This evidence seems to be supportive of the earlier observation (stemming from the hit rates calculated) that the CART model constructed does not appear accurate

in identifying underperformers. Similar observations were noted for the two linear-based short portfolios. Therefore, it is evident that all three models may not be adequate in selecting underperforming stocks.

From Table 5.3, it is clear that for both the CART-based long and short portfolios, the holding period (in months) for a stock is longer than that for the linear-based models (7.5 against 7.2 and 7.4 months). This observation is suggestive of another possible benefit of adopting a CART-based selection procedure: the best outperformance of the CART-based portfolio was not achieved through a higher stock turnover rate. Therefore, it is possible that transaction costs would not be increased by adopting a CART-based stock selection procedure. The above analysis ascertained a CART model's effectiveness within the universe of 84 stocks. It was also necessary to ascertain whether the CART-based model for stock selection delivered superior performance, out-of-sample, within the three sub-sectors (resources, industrials and financials) of the universe or whether the CART models' superior performance was specific to certain sub-sectors. Table D.1 in Appendix D shows the performance of the long and short portfolios for each of the three sub-sectors. It was evident from the results that, for each of the three sub-sectors the CART-based long portfolios deliver returns superior to those of the linear-based long portfolios. Moreover, in most cases the CART model delivered the highest risk-adjusted return. It was within the industrials sub-sector that the CART model delivered the best excess return and risk-adjusted return. Finally, it was also observed that the CART-based portfolios do not seem to achieve a higher return at the expense of a higher stock turnover.

With reference to the long and short portfolios, a final - and important - point to note is that the constituent stocks comprising the CART-based portfolio were very different from the constituent stocks comprising the linear-based portfolios. This evidence is supportive of the observation noted earlier on: CART models provide diversification from traditional linear modeling approaches.

The natural extension to the long and short portfolios above would be the construction of a portfolio containing both long and short positions in the stocks. Such a portfolio is of more practical use to quantitative investors and managers. Therefore, the 130-30 strategy portfolio and hedge-fund style portfolio were constructed. It was observed (see Table 5.3 above) that the CART-based hedge-fund style portfolio had a higher excess return (3.70 per cent), Sharpe ratio (0.37), Information ratio (12.53) and Sortino ratio (0.56) than the all the other linear-based portfolios.

These observations could motivate the inclusion of CART-based stock selection procedures in a hedge fund setting: a manager may be able to employ CART techniques for stock selection and demonstrate skill relative to his or her peers. However,

the hedge-fund idea employed by this dissertation was simplistic. Further research would be necessary for the development of different possible stock allocation strategies hedge funds could use, with stocks selected using a CART model.

Another observation of the CART-based hedge-fund style portfolio is that the lowest stock holding period at 3.1 months. The low holding period was also observed for the linear-based hedge-fund style portfolio, when compared to all the other portfolios. This suggests that the high outperformance of the hedge-fund style portfolio could be attributed to a higher stock turnover.

Upon review of the 130-30 portfolios constructed, CART-based as well as linear regression-based stock selection techniques appear to lead to a low portfolio tracking error and a high portfolio information ratio. These observations together with the modest excess return demonstrated by these two portfolios may provide justification for the inclusion of both linear regression-based and CART-based stock selection techniques within a 130-30 strategy. The CART-based and linear regression-based 130-30 portfolios appeared to track the benchmark well and also provided a small degree of outperformance hence meeting one of the aims of the 130-30 strategy, which is to provide outperformance from a selected benchmark return (Johnson *et al.*, 2007).

Another observation was that both the 130-30 strategy and the hedge-fund style portfolio had high Sortino ratios, when compared to the long and short portfolios. These high Sortino ratios are suggestive of a lower risk of capital loss, compared to the long and short portfolios. However, the risk of capital loss is perhaps less of a concern for quantitative investors and asset managers holding a hedge fund-style or 130-30 portfolio.

In summary, the results of this dissertation have emphasised that CART-based stock selection models offer better or comparable out-of-sample performance when compared to more traditional linear models. However, it must be borne in mind that the chosen linear model alternatives should be viewed as a 'base hurdle" for the CART models, rather than a definitive outperformance test. This observation seems to suggest that further research may be necessary to test CART techniques against more sophisticated linear models, such as logistic models used by Zhu *et al.* (2011), sequential factor selection and higher-order polynomial and interaction factors (Carvalho *et al.*, 2011). Moreover, the research has highlighted the model diversification benefits that CART models can offer when compared to linear models. Non-linear modeling techniques such as CART have not been widely adopted by quantitative investment managers and investors for the purposes of constructing stock selection models. These aforementioned benefits of CART models seem to motivate strong grounds for their inclusion within a quantitative professional's toolkit, especially

within a hedge-fund setting.

Bibliography

Andriyashin, A., Härdle, W. K. and Timofeev, R. (2008). Recursive portfolio selection with decision trees, *Technical report*, SFB 649 discussion paper.

Ang, A. (2008). The quant meltdown: August 2007, *Columbia CaseWorks ID* **80317**.

Bansal, R., Hsieh, D. A. and Viswanathan, S. (1993). A new approach to international arbitrage pricing, *The Journal of Finance* **48**(5): 1719–1747.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence, *The Journal of Finance* **43**(2): 507–528.

Bondt, W. F. and Thaler, R. H. (1987). Further evidence on investor overreaction and stock market seasonality, *The Journal of Finance* **42**(3): 557–581.

Bonga-Bonga, L. and Makakabule, M. (2010). Modeling stock returns in the south african stock exchange: A nonlinear approach, *European Journal of Economics, Finance and Administrative Sciences* **19**: 168–177.

Breiman, L. (2001). Random forests, *Machine learning* **45**(1): 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). Classification and regression trees. wadsworth & brooks, *Monterey, CA* .

Campbell, J. Y. (1987). Stock returns and the term structure, *Journal of financial economics* **18**(2): 373–399.

Campbell, J. Y., Hilscher, J. and Szilagyi, J. (2008). In search of distress risk, *The Journal of Finance* **63**(6): 2899–2939.

Campbell, J. Y. and Thompson, S. B. (2005). Predicting the equity premium out of sample: Can anything beat the historical average?, *Technical report*, National Bureau of Economic Research.

Carhart, M. M. (1997). On persistence in mutual fund performance, *The Journal of finance* **52**(1): 57–82.

Carvalho, C. M., Lopes, H. F. and Aguilar, O. (2011). Dynamic stock selection strategies: A structured factor model framework, *Bayesian Statistics* **9**: 1–21.

Chen, L., Novy-Marx, R. and Zhang, L. (2011). An alternative three-factor model, *Available at SSRN http://ssrn.com/abstract=1418117* .

Cochrane, J. H. (2011). Presidential address: Discount rates, *The Journal of Finance* **66**(4): 1047–1108.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues, *Quantitative finance* .

Fabozzi, F. J., Focardi, S. M. and Jonas, C. L. (2008). On the challenges in quantitative equity management, *Quantitative Finance* **8**(7): 649–665.

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

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds, *Journal of financial economics* **33**(1): 3–56.

Firer, C. and Staunton, M. (2002). 102 years of south african financial market history, *Investment Analysts Journal* **56**.

Frydman, H., Altman, E. I. and Kao, D.-L. (1985). Introducing recursive partitioning for financial classification: the case of financial distress, *The Journal of Finance* **40**(1): 269–291.

Goodwin, T. H. (1998). The information ratio, *Financial Analysts Journal* pp. 34–43.

Harvey, C. R. and Zhou, G. (1993). International asset pricing with alternative distributional specifications, *Journal of Empirical Finance* **1**(1): 107–131.

Hiemstra, C. and Jones, J. D. (1994). Testing for linear and nonlinear granger causality in the stock price-volume relation, *The Journal of Finance* **49**(5): 1639–1664.

Hsieh, D. A. (1991). Chaos and nonlinear dynamics: application to financial markets, *The journal of finance* **46**(5): 1839–1877.

Hueng, C. J. and McDonald, J. B. (2005). Forecasting asymmetries in aggregate stock market returns: Evidence from conditional skewness, *Journal of Empirical Finance* **12**(5): 666–685.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency, *The Journal of Finance* **48**(1): 65–91.

Johnson, G., Ericson, S. and Srimurthy, V. (2007). An empirical analysis of 130/30 strategies: Domestic and international 130/30 strategies add value over long-only strategies, *The journal of alternative investments* **10**(2): 31–42.

Kao, D.-L. and Shumaker, R. D. (1999). Equity style timing, *Financial Analysts Journal* pp. 37–48.

Khandani, A. E. and Lo, A. W. (2011). What happened to the quants in august 2007? evidence from factors and transactions data, *Journal of Financial Markets* **14**(1): 1–46.

Lakonishok, J., Shleifer, A. and Vishny, R. W. (1994). Contrarian investment, extrapolation, and risk, *The journal of finance* **49**(5): 1541–1578.

Lewellen, J. (2004). Predicting returns with financial ratios, *Journal of Financial Economics* **74**(2): 209–235.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, *The review of economics and statistics* **47**(1): 13–37.

Lys, T. and Sohn, S. (1990). The association between revisions of financial analysts' earnings forecasts and security-price changes, *Journal of Accounting and Economics* **13**(4): 341–363.

Post, T., Van Vliet, P. and Levy, H. (2008). Risk aversion and skewness preference, *Journal of Banking & Finance* **32**(7): 1178–1187.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing, *Journal of economic theory* **13**(3): 341–360.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk*, *The journal of finance* **19**(3): 425–442.

Sharpe, W. F. (1975). Adjusting for risk in portfolio performance measurement, *The Journal of Portfolio Management* **1**(2): 29–34.

Shively, P. A. (2003). The nonlinear dynamics of stock prices, *The Quarterly Review of Economics and Finance* **43**(3): 505–517.

Sorensen, E. H., Mezrich, J. J. and Miller, K. L. (1998). A new technique for tactical asset allocation, chapter 12.

Sorensen, E. H., Miller, K. L. and Ooi, C. K. (2000). The decision tree approach to stock selection, *The Journal of Portfolio Management* **27**(1): 42–52.

Sortino, F. A. and Van Der Meer, R. (1991). Downside risk, *The Journal of Portfolio Management* **17**(4): 27–31.

Symmetry: South African Hedge Fund Survey. (2013). *Technical report*, Available at http://www.symmetry.co.za/media/94398/hedge-fund-survey-2013-12.pdf.

Trueman, B. (1994). Analyst forecasts and herding behavior, *Review of financial studies* **7**(1): 97–124.

Van der Smagt, T. and Lucardie, L. (1991). Decision-making under not-well-defined conditions: From data processing to logical modelling, *Tijdschrift voor economische en sociale geografie* **82**(4): 295–304.

Wegner, T. (2010). *Applied business statistics: Methods and Excel-based applications*, Juta Academic.

Zhu, M., Philpotts, D., Sparks, R. and J. Stevenson, M. (2011). A hybrid approach to combining cart and logistic regression for stock ranking, *The Journal of Portfolio Management* **38**(1): 100–109.

Zhu, M., Philpotts, D. and Stevenson, M. J. (2012). The benefits of tree-based models for stock selection, *Journal of Asset Management* **13**(6): 437–448.

Appendix A

Appendix

A.1 Resources (RESI) sub-sector: model outputs

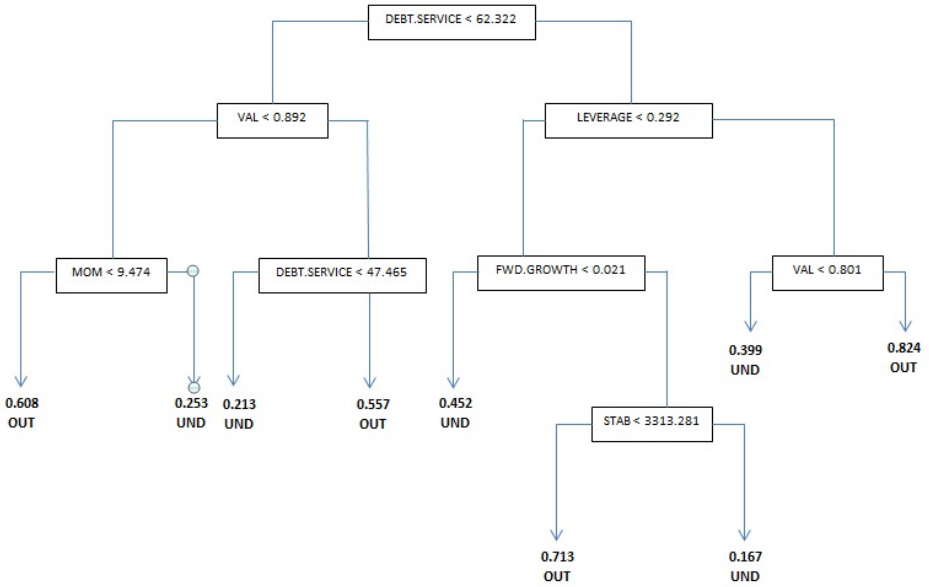

Fig. A.1: CART model for the 10 stocks from the RESI, constructed using data from January 2000 to April 2007. This decision tree models the probability of a stock outperforming the benchmark (in this case, the benchmark is the RESI). The dependent variable is depicted to be an outperformer (OUT) if the stock achieves a positive excess forward return, and an underperformer (UND) otherwise. Note that at each node, should the composite factor be less than the deciding value, then the decision tree will flow to the left-hand side. Finally, for a specific terminal node the probability of a positive excess forward return is reported.

	Regression-based weights (%)	**Mean-variance weights (%)**
STAB	-0.01	24.21
DEBT.SERVICE	-0.01	75.16
FWD.GROWTH	-0.01	88.15
HIST.GROWTH	14.50	30.44
PROF	-0.1	-32.23
EREV	44.42	3.10
VAL	42.62	80.58
MOM	-1.31	-13.48
LEVERAGE	-0.1	-155.93

Tab. A.1: The linear weights for the linear regression model and the mean variance model for 10 resource stocks from the RESI, using data from January 2000 to April 2007.

Appendix B

Appendix

B.1 Financials (FINI) sub-sector: model outputs

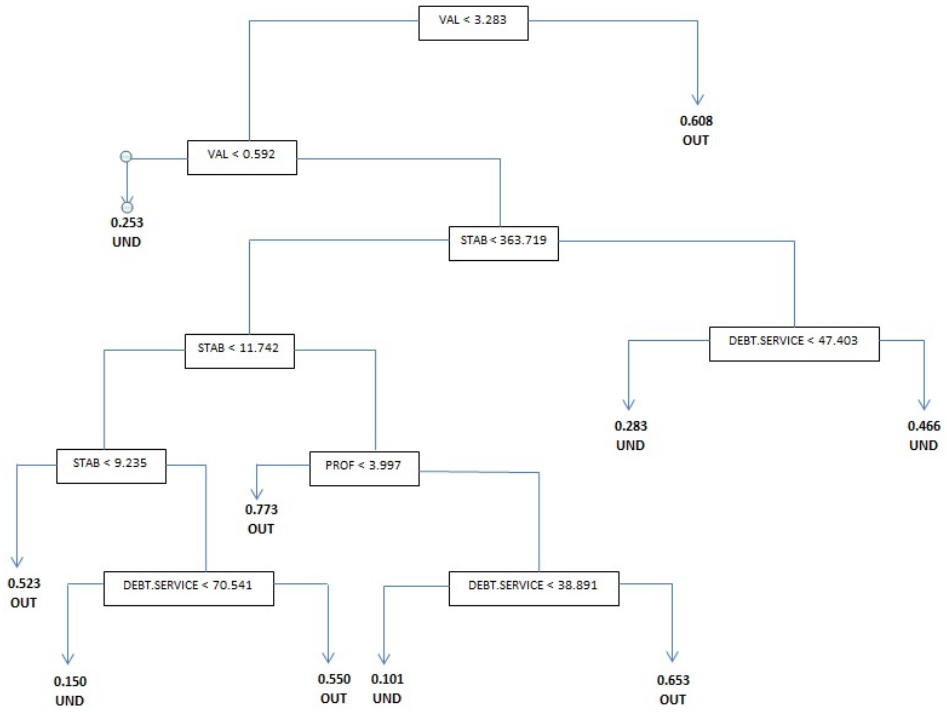

Fig. B.1: CART model for the 16 stocks from the FINI, constructed using data from January 2000 to April 2007. This decision tree models the probability of a stock outperforming the benchmark (in this case, the benchmark is the FINI). The dependent variable is depicted to be an outperformer (OUT) if the stock achieves a positive excess forward return, and an underperformer (UND) otherwise. Note that at each node, should the composite factor be less than the deciding value, then the decision tree will flow to the left-hand side. Finally, for a specific terminal node the probability of a positive excess forward return is reported.

	Regression-based weights (%)	**Mean-variance weights (%)**
STAB	0.1	-22.87
DEBT.SERVICE	-0.15	46.01
FWD.GROWTH	0.01	2.70
HIST.GROWTH	0.19	-11.36
PROF	-0.07	1.15
EREV	88.49	1.22
VAL	11.15	13.59
MOM	0.29	57.07
LEVERAGE	-0.01	12.49

Tab. B.1: The linear weights for the linear regression model and the mean variance model for 16 financial stocks from the FINI, using data from January 2000 to April 2007.

Appendix C

Appendix

C.1 Industrials (INDI) sub-sector: model outputs

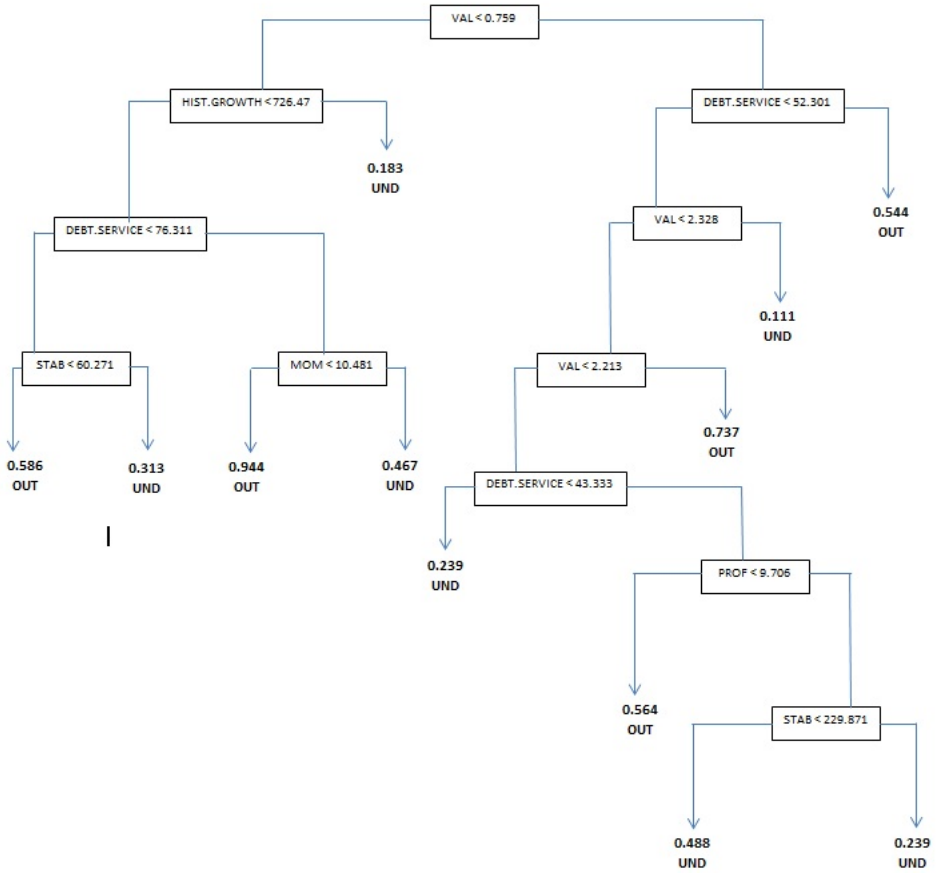

Fig. C.1: CART model for the 24 stocks from the INDI, constructed using data from January 2000 to April 2007. This decision tree models the probability of a stock outperforming the benchmark (in this case, the benchmark is the INDI). The dependent variable is depicted to be an outperformer (OUT) if the stock achieves a positive excess forward return, and an underperformer (UND) otherwise. Note that at each node, should the composite factor be less than the deciding value, then the decision tree will flow to the left-hand side. Finally, for a specific terminal node the probability of a positive excess forward return is reported.

	Regression-based weights (%)	**Mean-variance weights (%)**
STAB	-0.01	-10.14
DEBT.SERVICE	-0.02	-27.75
FWD.GROWTH	-0.02	81.26
HIST.GROWTH	32.26	38.29
PROF	0.78	85.59
EREV	26.49	-6.95
VAL	42.78	-38.13
MOM	-2.35	-9.16
LEVERAGE	0.09	-13.01

Tab. C.1: The linear weights for the linear regression model and the mean variance model for 24 industrial stocks from the INDI, using data from January 2000 to April 2007.

Appendix D

Appendix

D.1 Long and short portfolio performance for each of the three sub-sectors

Sub-sector	Portfolio	Model	Excess Return (%)	Tracking Error (%)	Sharpe Ratio	Information Ratio	Sortino Ratio	Holding Period (months)
Financials	Long	Regression	0.80	27.06	-0.42	0.17	0.22	7.7
		Mean-Var	0.75	26.19	-0.43	0.18	0.23	7.8
		CART	0.80	27.02	-0.39	0.19	0.20	6.9
	Short	Regression	0.49	21.11	-0.57	0.13	0.29	4.9
		Mean-Var	0.43	22.32	-0.57	0.12	0.23	6.5
		CART	0.17	20.19	-0.65	0.13	0.18	6.3
Industrials	Long	Regression	0.98	32.54	-0.73	0.47	0.40	7.7
		Mean-Var	0.99	33.71	-0.47	0.48	0.41	7.5
		CART	1.23	34.98	-0.42	0.55	0.32	7.6
	Short	Regression	0.47	31.01	-0.29	0.40	0.27	5.1
		Mean-Var	0.52	33.22	-0.39	0.45	0.26	6.3
		CART	0.23	32.45	-0.61	0.52	0.32	7.0
Resources	Long	Regression	0.53	27.01	-0.21	0.06	0.10	9.3
		Mean-Var	0.51	28.34	-0.15	0.08	0.10	7.5
		CART	0.61	27.67	-0.16	0.10	0.12	9.4
	Short	Regression	0.34	26.45	-0.16	0.02	0.13	3.5
		Mean-Var	0.41	24.77	-0.18	0.01	0.13	5.7
		CART	0.61	25.98	-0.27	0.04	0.13	7.2

Tab. D.1: Annualised excess forward returns, tracking errors, Sharpe ratios, Information ratios, Sortino ratios and holding period calculated for the long and short portfolios for each of the three sub-sectors. Note that transaction costs were not accounted for and the portfolios were rebalanced on a monthly basis. Stock turnover within a portfolio is indicated by the holding period.