A Comparative Analysis of Non-Linear Techniques in South African Stock-Selection

Nikhil Amaidas Hutheram

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

May 20, 2015

MPhil in Mathematical Finance,
University of Cape Town.

Declaration

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

Signed by candidate

Nikhil Amaidas Hutheram

May 20, 2015

Abstract

Forecasting stock performance has long been one of the primary objectives of financial practitioners. Literature has shown that the classical linear approach to modeling the interactions among company-specific factors and its stock market returns in time have become less suited for capturing the movements of the stock market. Hence, attempts to predict the performance of a stock have become associated with additional layers of complexity. This has led to the adoption of non-linear approaches to forecast stock performance. This dissertation explores the performance of some non-linear models in the South African market. These were classification and regression trees (CART), logistic regression and a random forest approach compared against a linear regression model. Moreover, a hybrid model between CART and logistic regression was considered. The models fell into two categories (i.e., static and dynamic models). Using a set of classification and portfolio performance metrics it was found that that a dynamic modeling approach outperformed a static approach. Overall, the logistic and linear regression models dominated in terms of performance against the tree-based models and hybrid approaches. The results also demonstrated that a hybrid approach offered an improvement over a stand-alone CART.

Acknowledgements

"It always seems impossible until it's done." — Nelson Mandela
"If I have seen further, it is only because I have stood on the shoulders of Giants." — Sir Isaac Newton

First and foremost, I would like to express my gratitude to my supervisor, Petrus Bosman for his patience and understanding throughout this research project. I would like to also express my gratitude to Dan Golding and Byran Taljaard of Avior Capital Markets. This study would not have materialised if it were not for their continuous support and interest in this research project. Last but not least I would like to thank Mario Giuricich, who assisted me in the early stages of this research project, from exchanging ideas to understanding concepts.

I would like to acknowledge my parents for their patience, love and support, without which I would have been unable to complete this research project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The financial market is complex and evolutionary, with the performance of the stock market being influenced by, *inter alia*, numerous economic factors, high degrees of uncertainty and unknown relationships. As a result, financial forecasting is difficult.

Ben Bernanke, the former chairman of the Federal Reserve Bank of the United States defined emerging markets as follows: "generally speaking, emerging market economies are defined as those economies in the low to middle-income category that are advancing rapidly and are integrating with global capital and product markets" (Bernanke, 2011). The Fragile Five and BRICS (Brazil, Russia, India, China, South Africa) are amongst the most well-known groupings of emerging market economies, with South Africa being a member of both.

After the 2008 global financial crisis, emerging markets were considered to be the engines of global growth and world financial stability, however, investor confidence in these countries has declined over the past few years. Because of the turbulence observed on the stock markets of most of these emerging market economies and increased access to information, stock selection techniques have become increasingly pertinent in identifying relationships between company-specific factors in order to predict stock market performance (Hargreaves and Hao, 2013; Karami and Talaeei, 2013).

According to current research, interactions amongst company specific factors and stocks market returns have been postulated using linear relationships in both an emerging and developed market setting (Sun, 2008; Zhu *et al.*, 2012; Giuricich, 2013). However in practice, movements in the market prices appear random and behave in a highly non-linear, dynamic manner (Lahmiri, 2011). Hence, stock selection models based on the classical linear framework are less suited to capturing higher order relationships between stock returns and company specific factors (Zhu *et al.*, 2011). The advantages of employing non-linear models, as outlined by Zhu *et al.* (2011), are that these models offer a higher degree of model diversification than linear models and that, if it is posited that the structural relationship between company

characteristics and stock returns are non-linear, it may be assumed that unexplained profit opportunities could be identified using non-linear models.

Technological advances have brought to the fore a number of methodologies for stock selection and prediction. Amongst non-linear models are machine learning techniques, such as neural networks and decision tree based models. In literature focusing on forecasting stock returns using non-linear methods, little attention has been given in the past to the South African market. More recently, however, the South African market has received greater interest due to its emerging market status and limited literature (Bonga-Bonga and Makakabule, 2010; Khanna and Palepu, 2010).

In the South African context, Bonga-Bonga and Makakabule (2010) posited the superiority of a Smooth Transition Regression Model, which is non-linear by nature, over an Ordinary Least Square and Random Walk model for modeling South African stock returns. A study by Kruger (2011), on the JSE All Share Index (ALSI) and its constituents found evidence of return predictability using company fundamental data through the use of linear and non-linear models. Hodnett *et al.* (2012) used a blended modeling approach between a linear and non-linear model as a stock selection technique. By using a wide range of company fundamental data their study concluded that the blended approach outperformed in out-of-sample forecasting ability against its linear and non-linear constituents. A recent study by van Gysen *et al.* (2013) examined a broader range of linear and non-linear models, which included AR, ARMA, GARCH and EGARCH models in forecasting returns on the JSE. Their results show that the linear models outperformed their non-linear counterparts over the 2008/09 financial crisis period.

A range of non-linear models have become increasingly used in international financial literature. These include models such as logistic regression, classification and regression trees (CART) and random forests. Logistic regression analysis has been used in many areas of corporate finance, such as assisting in default prediction and performance-based company classification (see Hua *et al.*, 2007; Gong and Sun, 2009; Chen, 2011; Hargreaves and Hao, 2013). Logistic regression has often been selected as a model of choice because it assists in the formation of a multivariate regression analysis between a response variable and several explanatory variables by empirically estimating coefficients for each of the explanatory variables (Lee *et al.*, 2007; Huang *et al.*, 2007; Upadhyay *et al.*, 2012). The advantages of logistic regression are that firstly the model itself is preferred when working with a binary response variable; secondly, the normality of the variables does not need to be assumed; thirdly, the model can be analysed with a mix of predictor variables (continuous, discrete and dichotomous) (Upadhyay *et al.*, 2012; Hargreaves and Hao, 2013). Additionally,

Zhu *et al.* (2011) found that logistic regression is highly effective at capturing the global features of a data set and the model is able to produce smooth probabilities through its continuous inputs.

The CART model was first proposed by Breiman *et al.* (1984). Their seminal works provide a detailed overview of the theory and methodology of CART, with additional examples from many disciplines. CART is a statistical technique also known as a recursive partitioning algorithm, whose purpose is to improve prediction (Sorensen *et al.*, 2000; Zhu *et al.*, 2012). The CART model makes use of a set of training data, also referred to as a "learning sample" (Lewis, 2000), to construct a decision tree. Each level of the decision tree represents a yes/no question such as "Is the patient a smoker?" or "Is the company expensive?", which repeatedly splits the tree into two further levels. Afterwards this hierarchy of decision rules is used to make predictions on unseen data (i.e., data not included in the training set). The CART methodology was first introduced in the early 1980s for use in the medical field, where it was initially employed to generate clinical decision rules (Lewis, 2000; Schroders, 2009). However, due to its non-linear nature and flexibility, over the years CART has been identified as a more robust process for analysing financial time series data (Schroders, 2009). The advantages of CART lie in its ability to identify non-linearities and complex interactions in data (Sorensen *et al.*, 2000; Zhu *et al.*, 2012). Additionally, CART offers a high degree of interpretability with its ability to compress large data volumes into understandable output; the model is non-parametric in nature, requiring no distributional assumptions regarding the variables in the model; the model is robust to the effect of outliers; it is invariant to monotone transformations of the independent variables, and the model has been noted to be adequate at handling non-homogeneous relationships with regard to conditional information (Breiman *et al.*, 1984; Lewis, 2000; Sorensen *et al.*, 2000; Timofeev, 2004; Schroders, 2009; Zhu *et al.*, 2011).

A combination of the logistic and CART models was successfully implemented by Zhu *et al.* (2011) and termed the hybrid model. The authors found the two models to be complimentary and discovered that the hybrid model delivered better forecasts of future stock returns than either the logistic and CART models. This was as a result of the hybrid model incorporating linear relationships through the use of the logistic model, which should in theory provide a superior model outcome to CART on its own. A natural extension to CART is a random forest, which is an ensemble learning method introduced by Breiman (2001). The random forest technique offers an improvement over a stand-alone CART, by constructing a collection of trees in place of the single CART tree and aggregating their output to determine the overall forest-based prediction (Cutler *et al.*, 2012). The advantages of a random forest

lie in its ability to be trained relatively quickly, having to depend on only two or three parameters, it is able to generate its own unbiased generalization error and it also solves the problem of the very discrete probability space that CART produces (Cutler *et al.*, 2012). Caruana and Niculescu-Mizil (2006) empirically tested the performance of a wide range of learning algorithms including decision trees, logistic regression and random forests. They concluded that a decision tree and logistic model performed relatively poorly against a random forest.

The objective of this dissertation was to compare the performance of eight linear and non-linear methods of stock selection on the Johannesburg Stock Exchange (JSE). The methods compared include a linear and logistic regression model (First and Second-Order), CART, a random forest approach and two versions of a hybrid model (combining CART & logistic regression) posited by Zhu *et al.* (2011). While this study attempted to distinguish between the performance of the aforementioned models using a traditional approach of having a fixed training and fixed testing set, this was extended to a dynamic technique of model estimation termed the "evolving approach" by Sorensen *et al.* (2000). The models implemented in this dissertation may be separated into two categories, which will be termed static (i.e., finite training/testing set) and dynamic (i.e., re-estimation at each period) models. The predictive ability of each model were assessed out-of-sample following two approaches. The first approach was to test the classification ability of each model using a range of classification metrics commonly employed in practice (i.e., accuracy, precision, recall, etc.). The second approach was to construct long and long-plus-short portfolios with their performance measured using a range of risk and non-risk adjusted metrics.

The structure of the dissertation is as follows. Section 2 presents the theoretical development of each model implemented in this study. Section 3 provides detail on the data utilized, that is, where it was sourced, issues that arose, and how the data was tailored for the purposes of the study. This is followed by an explanation of how each model was fitted for the purposes of South African stock selection, in addition to a description of the performance metrics used in testing each model out-of-sample. The final section concludes with a representation of the results and a detailed comparison of the model performance followed by conclusions.

Chapter 2

Theoretical development

2.1 Regression models

Regression analysis looks to explain the association between a single response variable and multiple explanatory variables. Generalized linear models (GLM) encapsulate a larger class of models, originally introduced by Nelder and Wedderburn (1972) and further popularized by McCullagh and Nelder (1989). A GLM, as described by Dobson and Barnett (2008), is characterized by a random component, a systematic component and a link function. The random component specifies the probability distribution associated with the response variable Y_i. The distribution of the response is a member of the exponential family, which includes the Gaussian (normal), binomial, Poisson, negative binomial and gamma distributions. The systematic component specifies the explanatory variables (X-variables) in the model. More specifically, the linear combination of explanatory variables is used to construct a linear function of the form:

$$Z_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}, \tag{2.1}$$

where k is the number of explanatory variables, X_{ij} is the j^{th} predictor for the i^{th} case and β_j is the j^{th} coefficient.
The link function is used to describe the association between the random and systematic components. It is expressed as $g(\mu_i)$=$g(E(Y_i))$ and represents the transformation of the mean of the response that will be modeled as a linear function of the X-variables. For the purposes of this research only two simple regression models were used, a linear and logistic regression model. The following illustrations of the models below have been adapted from Dobson and Barnett (2008).

2.1.1 Linear and logistic regression models

The linear model is the simplest case of a GLM, and is expressed as:

$$E(Y_i) = \mu_i = \boldsymbol{x_i^T \beta}, \tag{2.2}$$

or more commonly as

$$Y_i = \beta_0 + \boldsymbol{\beta} X_i + \boldsymbol{\epsilon}_i, \tag{2.3}$$

where the response variable Y_i is independent and normally distributed with mean μ_i and variance σ^2 and $\epsilon_i \sim N(0, \sigma^2)$ is the error term in the model. The X-variables constitute the systematic components in the linear model and may be continuous or discrete. The identity link, $g(E(Y_i))=E(Y_i)$, is used as the mean is modeled directly.

A GLM for a binary response variable which measures the probability of a particular outcome, is linked to the linear predictor function by the following expression:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_j \beta_j x_{ij} = Z_i, \tag{2.4}$$

which is termed the logit link, where equation (2.4) represents the logistic regression expression. As a result, the relation between Z_i and the probability of the event of interest can be illustrated and simplified as follows:

$$\begin{aligned} \log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \sum_j \beta_j x_{ij} \\ \Rightarrow \log\left(\frac{p_i}{1-p_i}\right) &= Z_i \\ \Rightarrow \frac{p_i}{1-p_i} &= e^{Z_i} \\ \Rightarrow p_i &= \frac{e^{Z_i}}{1+e^{Z_i}} \end{aligned}$$

The response variable in the logistic regression expression above is binomially distributed, where $p_i \in (0, 1)$ is the probability of success. Analogous to the linear model, the X-variables may be continuous or discrete. The link function used is the logit link, $\text{logit}(p) = \log(\frac{p}{1-p})$.

2.2 CART model

A CART classification may be separated into two categories, depending on the nature of the response variable (i.e., continuous or discrete). *Classification trees* are used for discrete variables, while *regression trees* are used for continuous variables. The trees referred to in this research were all regression trees as the response variable was continuous and the purpose of the tree was to produce the probability that a stock would outperform over the next period. At a high level James *et al.*

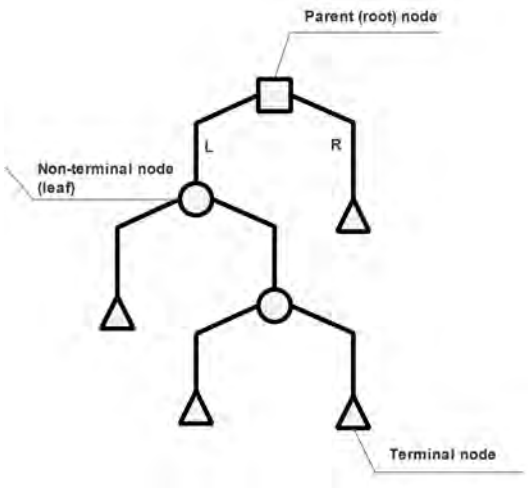

Fig. 2.1: A graphic illustration of the structure of a CART model.

(2013) explained the construction of a regression tree in two steps. Let Y denote a continuous response variable and let $X_1, ..., X_k$ be a set of predictor (explanatory) variables, which represent the predictor space. The first step is to partition the predictor space into J non-overlapping regions, $R_1, ..., R_J$. Secondly, for every observation that lands in a region R_j its associated predicted value is simply the average of the responses for each training observation within R_j.

In greater detail, Breiman *et al.* (1984) and Zhu *et al.* (2011) denote a learning sample $\boldsymbol{L}$ as, $\boldsymbol{L}$=$\{(x_1, y_1),...,(x_n, y_n)\}$ which comprises of a vector of explanatory variables $\boldsymbol{x}_i$ and associated responses y_i, where n denotes the number of observations. Let the learning sample denote the *root node*. The objective as outlined by Breiman *et al.* (1984) and Zhu *et al.* (2011) is to recursively partition this space into two descendant nodes. Figure 2.1 illustrates the structure of a CART model and its various levels.

For a given predictor variable and splitting value s, a non-terminal node can be split into a left and right region represented as R_L=$\{X|X_j < s\}$ and R_R=$\{X|X_j \geqslant s\}$. The splitting criterion for a regression tree is to minimize the mean squared error (MSE) at the node:

$$\text{MSE} = \vartheta = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \tag{2.5}$$

where $\hat{y}$ is the average of the training observations at the node. Out of all the X variables and splitting values s the candidate split chosen to partition a non-terminal

node into two descendant nodes is the split that minimizes the equation:

$$\vartheta_{split} = \sum_{i=1}^{n_L}(y_i - \hat{y_L})^2 + \sum_{i=1}^{n_R}(y_i - \hat{y_R})^2 \tag{2.6}$$

$$= n_L\vartheta_L + n_R\vartheta_R \tag{2.7}$$

This process is repeated recursively until no node can be further split (i.e., there is no reduction in MSE) and as a result those nodes are then referred to as *terminal nodes*. This procedure does however lead to the model over-fitting the data (Breiman *et al.*, 1984; Hastie *et al.*, 2009; Cutler *et al.*, 2012; James *et al.*, 2013), this is overcome by a process called *pruning*, which is addressed in Section 2.5.

2.3 The Hybrid model (CART & Logistic regression)

Zhu *et al.* (2011) constructed the hybrid model following a two step procedure. Firstly, they apply the CART model to uncover the higher-order interactions in the data and then they modify those probabilities using a logistic regression approach. The hybrid approach by construction requires a dynamic CART model to be trained at each quarter, because a logistic model is used to modify each stock's terminal node specific probability, where at each quarter different stocks may lie in different terminal nodes of the constructed tree. The number of stocks which require classification will be denoted by n. Let $\boldsymbol{X}$ denote the matrix of explanatory variables for each of these n stocks with dimensions $n \times k$, where k represents the number of explanatory variables. $\boldsymbol{X}_j$ is a column vector for the j^{th} explanatory variable for each of the n stocks.

Assume that at the end of a given quarter the optimal tree trained on the set $\boldsymbol{X}$ has r terminal nodes with n_r stocks in the r^{th} terminal node. Each of the r terminal nodes has an associated probability, p_r, which denotes the probability of outperformance for each of the stocks in that node. CART associates the same probability to all the stocks that fall in a specific node, while a more appropriate method would be to fine-tune each stock's associated probability through a stock specific component which can be represented as follows:

$$p_{ri} = p_r + \phi_{ri}, \;\; i = 1, 2, ..., n_r, \tag{2.8}$$

where p_{ri} is the probability associated with the i^{th} stock in the r^{th} terminal node, and ϕ_{ri} is a stock specific component used to modify the probability p_r for stock i. The second step describes the method of fine-tuning the probabilities by implementing a logistic regression to model the stock specific component ϕ_{ri}. The probability of interest p_{ri} is obtained by the following expression:

$$\text{logit}(p_{ri}) = \text{logit}(p_r) + \mathbf{X}_i^r \boldsymbol{\beta}_r, \tag{2.9}$$

where $\boldsymbol{X}_i^r$ represents the i^{th} row in the explanatory variable matrix $\boldsymbol{X}^r$ for the n_r stocks in the r^{th} terminal node. After constructing the CART model, Zhu *et al.* (2011) described two possible methods in proceeding with the hybrid model, of which Zhu *et al.* (2011) only implemented the second. The two methods are described below:

1. Fit a logistic regression model for each of the r terminal nodes to obtain a vector of coefficients $\boldsymbol{\beta}_r$, which is specific to that r^{th} node. In greater detail, the terminal node specific regression model is fitted using only the data for the observations in that node and using that terminal nodes logit transformed

probability, logit(p_k), as an offset parameter in the regression. An offset parameter is a term in a regression model whose coefficient is not estimated by the model but is rather constrained to one.

2. Fit a single logistic regression model trained on all the data, as used in constructing the CART model, while simultaneously using each observations associated logit transformed probability, logit(p_k), as an offset parameter. This results in a single vector of coefficients denoted by $\boldsymbol{\beta}$ replacing $\boldsymbol{\beta}_r$ in equation (2.9) above.

In an approach to the method implemented by Zhu *et al.* (2011), if the logistic regression model adds no further information to a stocks probability of outperformance (p_r), the probability remains the CART-based probability. For the purposes of this research the former procedure of fitting a terminal node specific logistic regression is referred to as the Hybrid model 1 and the latter method of fitting a single universal logistic regression model is called the Hybrid model 2.

2.4 Random forest approach

CART may be an adaptive non-linear model, but it is not without its limitations. The binary recursive partitioning algorithm (considered to be a greedy[1] algorithm) used to construct the tree may cause it to partition the predictor space in a way that may at first be promising, but may lead to a less optimal configuration than another configuration that had used an initial partition that was suboptimal. This naturally leads to the notion of not being restricted to using a single tree, but rather considering many trees. This approach is commonly known as the random forest method, as popularised by Breiman (2001).

A random forest is an adaption of CART that uses a technique known as *bagging*, which refers to bootstrap aggregation (Breiman, 1996). The term "bootstrap" in this context refers the procedure of resampling (with replacement) from the training set (Singh and Xie, 2008).

Theorem 2.1. *(James et al., 2013: p. 316) For a given set of n independent observations $Z_1, ..., Z_n$ each with a variance σ^2, the variance of the mean $\hat{Z}$ of the observations is given by $\frac{\sigma^2}{n}$.*

Theorem 2.1 shows that a reduction in variance may be achieved by aggregating across a set of observations. Bagging, as described by James *et al.* (2013) is the

[1] The construction algorithm of CART is termed "greedy", because it is locally optimal. In that it chooses the locally best predictor variable (splitting rule) at each stage in its process—usually one node at a time (Bennett, 1994).

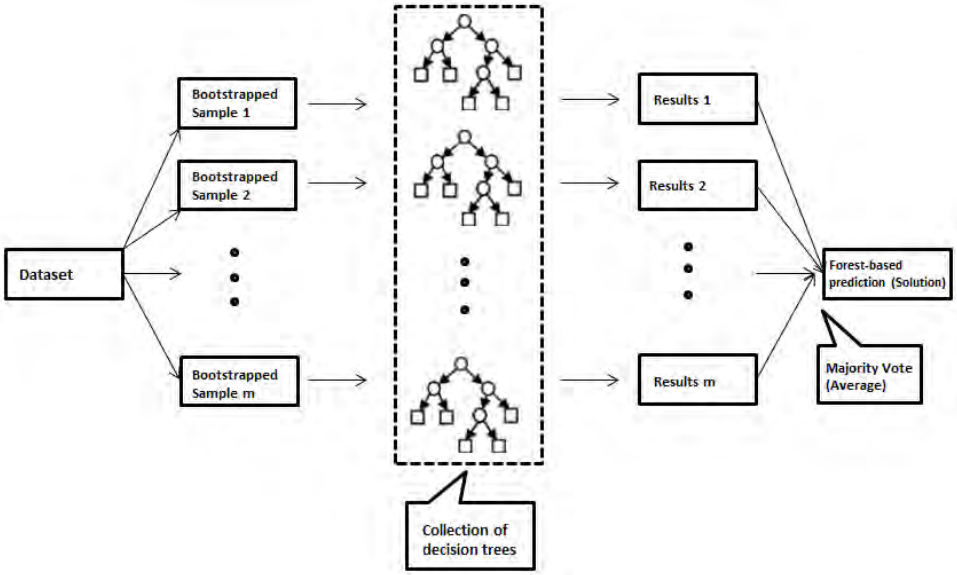

Fig. 2.2: A graphic illustration of a random forest (Zhu, 2012).

procedure of reducing the variance of a decision tree by growing a full tree on each of the bootstrapped training sets and averaging the predictions, as seen in the above theorem. The in-sample observations not used in the growing of the decision tree (i.e., the observations not included in the bootstrapped training sets) are referred to as out-of-bag (OOB) observations. These OOB observations are used to compute an overall OOB mean-squared-error for a given random forest model, represented as:

$$\text{MSE}_{oob} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{oob}(x_i))^2, \tag{2.10}$$

where $\hat{y}_{oob}(x_i)$ is the average predicted response for the i^{th} out-of-bag observation.

James *et al.* (2013), Cutler *et al.* (2012) and Breiman (2001) concluded that this OOB error is a valid test error as it is computed using those observations not used in the growing of a tree. This OOB error can then be used to decide on the value for the varying parameters in a random forest model, which include the number of trees, minimum number of observations per node and the number of predictor variables considered at each split in the individual trees (q).

The following is a summary of the steps taken in constructing a random forest, with an intuitive graphic illustration represented in Figure 2.2, adapted from Cutler *et al.* (2012), Zhu (2012) and James *et al.* (2013).

- Create M bootstrapped samples with replacement from the full set of training

data, where each sample is approximately the same size as the original dataset.

- For each sample grow a full regression tree as outlined in section 2.2. However, at each non-terminal node randomly select $q < k$ out of the k explanatory variables to split the node into two descendants, imposing variation in the constructed trees.
- The forest-based regression forecast for a given set of independent variables is taken to be the average of the predictions for the individual trees.

2.5 Addressing overfitting

2.5.1 Variable selection

A learning algorithm that exactly fits the idiosyncrasies exhibited by a set of in-sample data will extrapolate those idiosyncrasies out-of-sample, which is known as overfitting. Hawkins (2004) described overfitting to be when too many features are included in a model learned on a set of in-sample data causing the resulting model to fail in generalizing out-of-sample. Overfitting may be considered as violating the principle of parsimony (Occam's Razor). That is, 'all other things being equal the simplest solution is best' (William of Okham, 14th Century).

Regression analysis in conjunction with the principle of parsimony implies that the smallest model fitting the data is the best, as a result a regression model fitted with all the explanatory variables may lead to overfitting (Zucchini, 2000; Faraway, 2004; James *et al.*, 2013). Hence, a stepwise regression procedure which involves sequentially adding variables that add the greatest improvement to the model fit is used. More specifically, the Akaike information criterion (AIC) is a commonly employed method in predictive modeling in conjunction with stepwise regression to select the most important variables (Shtatland *et al.*, 2001; Arnold, 2010; Zhu *et al.*, 2011). A study by Stone (1977) shows that AIC is asymptotically equivalent to cross-validation, a method used in the CART framework (i.e., in the pruning procedure of the CART model). The linear and logistic regression models used in this research were all constructed using the AIC variable selection procedure.

2.5.2 Pruning a regression tree

The process described in Section 2.2 results in an overly large tree being grown, which fits the training data well, but is unlikely to perform well out-of-sample (Lewis, 2000). The process of *pruning* produces a sub-tree out of the large tree grown by

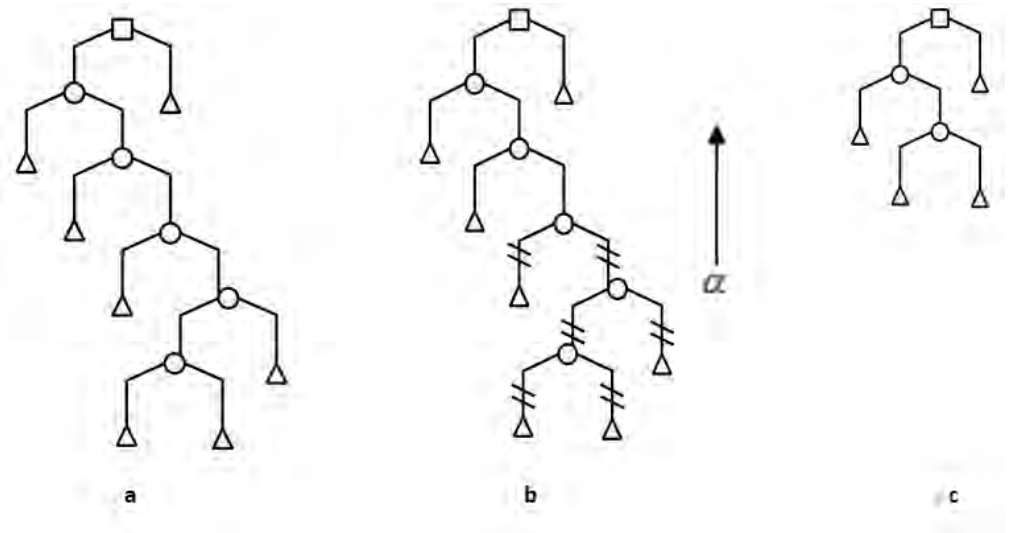

Fig. 2.3: A graphic illustration of a large tree being pruned to a smaller subtree. Subfigure (a) shows a maximal tree, subfigure (b) shows which nodes are pruned for a given value of alpha and subfigure (c) displays the final pruned tree.

trimming off lower level nodes of the tree. This is illustrated in Figure 2.3. A pruned decision tree allows for easier interpretation in addition to its improved out-of-sample performance (Breiman *et al.*, 1984; Torgo, 2000). The chosen sub-tree is the tree with the lowest validation error rate, which is estimated using cross-validation (described below). The methodology used in achieving this goal is termed, *error-complexity pruning* (Breiman *et al.*, 1984).

The method begins by growing a maximal tree (T_{max}) by the process described in Section 2.2. During the pruning process, a complexity (tuning) parameter, $\alpha \geqslant 0$, is gradually increased. As α increases from zero, the terminal nodes of the maximal tree get pruned (removed) in a nested fashion. As a result, a sequence of sub-trees are produced as a function of α. In accordance to the principle of parsimony, by selecting an α a sub-tree is selected that does not overfit the in-sample set. K-fold cross validation is used to choose this α (and hence sub-tree) that has the lowest cross-validation error, and therefore the best out-of-sample predictive performance (Breiman *et al.*, 1984; James *et al.*, 2013; Lacerda, 2014).

The following is an outline of the method adapted from the literature (Breiman *et al.*, 1984; Torgo, 2000; James *et al.*, 2013). For a more detailed exposition refer to Brieman *et al.*, 1984: p. 66 – 78, 232 – 237.

Definition 2.2. (Brieman *et al.*, 1984: p. 66) For any subtree $T \preccurlyeq T_{max}$ (i.e., T is a pruned subtree of T_{max}), define its complexity as $\mid \tilde{T} \mid$, the number of terminal nodes in T. Let $\alpha \geqslant 0$ be a real number called the complexity parameter and define the error-complexity measure $C_\alpha(T)$ as

$$C_\alpha(T) = C(T) + \alpha \mid \tilde{T} \mid \tag{2.11}$$

$$= \sum_{m=1}^{|\tilde{T}|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha \mid \tilde{T} \mid \tag{2.12}$$

In this definition $C(T)$ is the sum of squared errors for tree T, also known as the resubstitution error[2] in this context, where R_m represents a subset of the predictor space, referred to in Section 2.2, associated with the m_{th} terminal node and $\hat{y}_{R_m}$ is the average predicted response of the training observations in R_m. $\mid \tilde{T} \mid$ is the number of terminal nodes for any subtree of the maximal tree, $T \preccurlyeq T_{max}$. α is the non-negative complexity parameter controlling the trade-off between a subtrees complexity and its fit to the training data. For $\alpha = 0$ the subtree is equivalent to the maximal tree T, and as α increases there is penalty for having a complex (large) tree, hence equation (2.11) is minimized for a smaller subtree.

The next step is to find the pruned subtree $T(\alpha) \preccurlyeq T_{max}$ which minimizes $C_\alpha(T)$ in equation (2.11) for a given value of α (Breiman *et al.*, 1984), which can be represented as:

$$C_\alpha(T(\alpha)) = \min_{T \preccurlyeq T_{max}} C_\alpha(T) \tag{2.13}$$

The goal is to find an increasing sequence of α values where each value has an associated subtree with a terminal node recursively trimmed off. This results in the maximal tree being decomposed into a finite set of subtrees and associated increasing sequence of α values which can be represented as:

$$T_{max} \succ T_1 \succ T_2 \succ ... \succ t_0 \tag{2.14}$$

and

$$0 = \alpha_0 < \alpha_1 < \alpha_2 < ... < \alpha_{t_0} \tag{2.15}$$

where t_0 is the root node of the tree. Even though α can range through a continuum of values, there is a decreasing sequence of pruned trees such that each subtree is optimal for a given α , as a result there exists only a finite set of *interesting* α values (see Breiman *et al.*, 1984; Torgo, 2000).

[2] The error rate for a decision tree computed from the training sample (Bradford *et al.*, 1998).

The next step is to use K-fold cross validation to choose α. That is, divide the learning sample $\boldsymbol{L}$ into K subsets, $\boldsymbol{L}_k$, $k = 1, 2, ..., K$. K maximal trees are then grown, where each tree is trained on the remaining K-1 folds. Each tree is then pruned according to the aforementioned procedure (i.e., error complexity pruning) resulting in a sequence of best subtrees as a function of α, generating a parametric family of pruned trees $T^k(\alpha)$. Estimates of the MSE for each tree in each of the k sequences is obtained using the k_{th} left-out fold, as a function of α. Average the errors for each α and choose the α corresponding to the minimum average error/cross-validation error. There are two ways to proceed from here: either choose the subtree from the sequence in (2.14) that corresponds to the chosen α or choose the smallest subtree in the sequence whose cross-validation error lies within one standard error of the minimum cross-validation error, that is, within the interval $E^{CV}_{min} + SE(E^{CV}_{min})$. The latter method is most commonly used in practice and known as the one-standard error (1-SE) rule (Breiman *et al.*, 1984; Torgo, 2000). The CART models used in this research were all pruned using 10-fold cross validation and selected using the 1-SE rule. Upon investigation it was found that increasing the number of folds made the procedure more time-consuming without adding any benefit to the models' performance.

Chapter 3

Data and Methodology

3.1 Data

Quarterly stock data for the period from January 2000 (2000:Q1) to July 2014 (2014:Q3), for 184 companies listed on the JSE were used. This was a summation of all the stocks listed on the JSE Top40 and JSE MidCap indices for the aforementioned period. Access to company fundamental data is restricted to the frequency at which it is reported. Quarterly stock data is selected for this research, as it was found that company fundamental data would, at best, be updated quarterly and mostly only be updated every 6 months. For this research, it was necessary to also source data from several different sources namely Bloomberg, Thomson Reuters (Datastream) and McGregor BFA, which raised the task of cross-checking data across these providers.

A common practice in detecting outliers in data is to highlight those values lying three standard deviations away from the mean. However, Leys *et al.* (2013) highlight the disadvantages of using such a method one of them being that the outliers themselves influence the mean and standard deviation. Hence, in addition to the aforementioned method the data was bucketed into quantiles and the values lying in the upper and lower two percent tails were highlighted. After some preliminary cross-checking it was decided that a third of the data should be collected off Bloomberg and the remainder off Datastream. The latter was found to provide consistent output for certain fundamental ratios in addition to the systems' ability to provide more complete information for companies that had ceased to exist since the period of review.

Look-ahead bias can be thought of as "Leakage of future information" (Aronson, 2011). Chan (2013) described look-ahead bias as the notion of future information being used to make a "prediction" at the current time. All company fundamental data was lagged appropriately to account for the look-ahead bias. Gilbert and Strugnell (2010) showed that survivorship bias may lead to incorrect inferences being

made from financial data. Survivorship bias was also avoided by considering all the stocks over the period in question and not just a surviving sample.

All the analyses for this dissertation were performed in MathWorks MATLAB.

Two separate studies by Zhu *et al.* (2011, 2012) and a more recent study by Giuricich (2013) utilized nine composite factors to forecast future stock performance with each comprising of an equally weighted average of distinct fundamental stock ratios. Hence, for this research attention was given to collecting the ratios posited by those studies in order to assist with forecasting the future performance of a stock.

In further detail, the fundamental stock characteristics used as inputs into the forecasting models can be broadly categorised into five groups (see Zhu, 2012). Firstly, value factors (dividend earnings, book value and cash flows). Fama and French (1988), Lewellen (2004) and Ang and Bekaert (2007) illustrate that dividend earnings offer some explanatory power in forecasting future stock returns. Pontiff and Schall (1998) provide significant evidence illustrating the ability of book value in forecasting a firms future cash flows. Additionally, a study by Lakonishok *et al.* (1994) illustrates that value–based strategies using a firms cash flows has power in forecasting future stock returns. Secondly, momentum factors, which refers to the historical performance of a stock. The popularity around the use of momentum factors comes from its supportive literature. Bondt and Thaler (1985, 1987) illustrate the predictive power of stock returns of a 3–5 year time horizon, while later studies by Jegadeesh and Titman (1993, 2001) show that a portfolio strategy of being long winning stocks and being short losing stocks based on returns over the previous 6–12 months generate excess returns. Asness (1997) and Dhrymes and Guerard Jr (2012) also emphasise the usefulness of momentum–based strategies in forecasting future stock returns. Thirdly, profitability factors (return-on-equity, pre-tax margins and asset turnover ratio). Chen *et al.* (2011) illustrates the predictive power of a firms profitability factors, however Campbell and Thompson (2008) found that even though its predictive ability is warranted, it is weak. Fourthly, financial strength factors (debt to equity ratio, debt to market capitalization ratio, interest cover and free cash flow to debt ratio), which measure a firms current debt capacity and its ability to service that debt. Bhandari (1988) concluded that a positive correlation exists between a companies future stock returns and its debt/equity ratio, with supportive evidence from a study by Barbee Jr *et al.* (1996). Davis *et al.* (2012) also emphasises the importance of a firms debt attributes as a signal for measuring its current and future performance. Finally, analyst forecast factors (brokers forecasts and revisions of those forecasts). Opinions from company outsiders through analyst/broker forecasts has become an increasingly popular topic in academic literature. A study by Trueman (1994) found an association between analyst forecasts

Tab. 3.1: Composition of the nine explanatory variables with each of their respective fundamental data inputs.

Explanatory variable	Composition
Value (VAL)	The mean of earnings to price, dividends to price, cashflow to price, sales to price, book to price.
Profitability (PROF)	The mean of return-on-equity, pre-tax margin, asset turnover.
Leverage (LEVERAGE)	The mean of debt to equity, debt to market cap.
Debt service (DEBT.SERVICE)	The mean of interest cover, free cashflow to debt.
Momentum (MOM)	Relative Strength Index (RSI) 14 day.
Stability (STAB)	The mean of volatility in corporate earnings, sales (revenues) and operating cashflows.
Historic growth (HIST.GROWTH)	The mean of 3 year historic growth in earnings, sales (revenues) and operating cashflows.
Forward growth (FWD.GROWTH)	Broker forecasts of earnings per share (EPS) two years ahead.
Earnings change (EREV)	Change in broker forecasts for EPS.

and a firms future stock returns. Ang and Ciccone (2001), Davis *et al.* (2012) and Shan *et al.* (2014) also emphasise the usefulness of "other information" in analysts' forecasts as potential predictors of the future cash flows of a company. Additionally, as noted by Zhu (2012), further studies by Givoly and Lakonishok (1980) and Lys and Sohn (1990) show that outperformance can be achieved by actively using analysts' earnings revisions over a simple buy and hold strategy.

A total of 21 fundamental stock ratios were collected and bucketed into their respective categories as shown in Table 3.1. The JSE All Share Index (ALSI) was used as the benchmark index for this research, however data for the ALSI was only available from 2002 mid-year. To manage this gap in the data, by using the market capitalization of each of the 184 stocks and their total returns cum-dividends, a market-cap weighted index was formed as a proxy benchmark for the period January 2000 to mid-year 2002. This compiled benchmark (market-cap weighted and ALSI) was used in the training and testing phases of each model.

3.1.1 Multicollinearity

Before proceeding with the model implementation it was necessary to check whether any of the explanatory variables were correlated. Multicollinearity has been defined as the condition where the explanatory variables in a multiple regression model are highly correlated (Heiberger and Holland, 2004). The presence of multicollinearity is problematic when implementing linear or GLM models. James *et al.* (2013) explained that the presence of collinearity between explanatory variables results in unstable estimates for the regression coefficients. The correlation matrix (Table 3.2) for the explanatory variables should not be inspected in isolation as it provides no information on the possibility of the existence of collinearity among three or more variables. For this reason the variance inflation factor (VIF) for each variable may be examined. The VIF provides a measure of the impact of multicollinearity among variables in a regression model, while a correlation matrix or scatter-plot provides only a bivariate measure (James *et al.*, 2013). For each explanatory variable the VIF can be computed as:

$$\mathrm{VIF}_{X_j} = \frac{1}{1 - R^2_{X_j|X_{j-1}}}, \tag{3.1}$$

where j=1...k explanatory variables and the $R^2_{X_j|X_{j-1}}$ is the R-squared obtained by regressing the j^{th} explanatory variable on to the remaining explanatory variables. The VIF is bounded below by 1 with no upper bound. No standard cutoff value exists to identify how high a VIF value should be to warrant a problem, however studies by Rogerson (2001), Hair *et al.* (2006), Obrien (2007) and Pan and Jackson (2008) indicated that VIF factors in excess of 4 or 5 should be cause for concern. The Pearson correlation matrix in Table 3.2 illustrates that the correlation between the explanatory variables was quite low with 0.36 (FWD.GROWTH and EREV) being the greatest positive value, which is plausible as one is a function of the other. Additionally, the VIF values were all below the cut-off value of 4. Had severe correlations been present between the explanatory variables, a method of grouping or exclusion of the explanatory variables would need to have been considered (James *et al.*, 2013).

Tab. 3.2: Pearson correlation matrix and variance inflation factors (VIF) for the nine explanatory variables for the period Q1 2000:Q1 – 2014:Q3.

	VAL	PROF	LEVERAGE	DEBT SERVICE	MOM	STAB	HIST. GROWTH	FWD. GROWTH	EREV
VAL	1.00	-0.08	0.05	0.02	-0.05	-0.12	0.06	-0.01	-0.01
PROF	-0.08	1.00	-0.23	0.29	-0.01	-0.06	0.15	0.10	0.03
LEVERAGE	0.05	-0.23	1.00	-0.45	-0.07	0.01	-0.03	-0.05	-0.04
DEBT SERVICE	0.02	0.29	-0.45	1.00	0.05	-0.03	0.06	0.07	0.02
MOM	-0.05	-0.01	-0.07	0.05	1.00	0.04	0.03	0.04	0.03
STAB	-0.12	-0.06	0.01	-0.03	0.04	1.00	0.09	-0.04	-0.03
HIST. GROWTH	0.06	0.15	-0.03	0.06	0.03	0.09	1.00	0.04	0.02
FWD. GROWTH	-0.01	0.10	-0.05	0.07	0.04	-0.04	0.04	1.00	0.36
EREV	-0.01	0.03	-0.04	0.02	0.03	-0.03	0.02	0.36	1.00
VIF	1.04	1.15	1.29	1.33	1.01	1.03	1.04	1.16	1.15

3.2 Model fitting and testing

Tab. 3.3: A list and description of the models implemented in this study.

Models	Description
CART	CART model pruned using 10-fold cross validation.
CART-NP	CART model with no pruning.
Random forest (RF)	RF model with a forest size of 500. The predictor set size considered at each split was 4.
First-order linear and logistic model	Polynomial of order one (X_k^1)
Second-order linear and logistic model	Polynomial of order two (X_k^2) and interaction terms ($X_k X_j$)
The Hybrid model 1 & 2	Terminal node specific logistic model; Single logistic model
Random classifier	Uniformly distributed random numbers on the open interval (0,1).

A study by Sorensen *et al.* (2000) found that an evolving CART model performed better than a CART learned on a fixed training sample. The hybrid model posited by Zhu *et al.* (2011) imposed a logistic model on top of an evolving tree, however the study did not extend this "evolutionary" technique onto the logistic, random forest and standalone CART model it used for comparison. A model evolving over time seems to make more sense, as it can progressively adapt to market dynamics (Sorensen *et al.*, 2000). Hence, for every model employed in this research its evolving counterpart was also calculated. Consequently, the models in this research can be divided into two broad categories, namely static (i.e., predetermined training and testing set) and dynamic models (i.e., evolving approach).

Each model's performance was then assessed out-of-sample, using the data for the period January 2000 – December 2010 (2000:Q1 – 2010:Q4) for training, while retaining January 2011 – July 2014 (2011:Q1 – 2014:Q3) for testing. The in-sample period was deliberately chosen so as to capture the period of the 2007/08 Global Financial Crisis and part of the recovery period, so as to incorporate information on the turbulent market activity exhibited over that period. The dynamic models implemented in the study were sequentially formed using all the data from January 2000 up to each quarter end, concluding at the end of the second quarter of 2014.

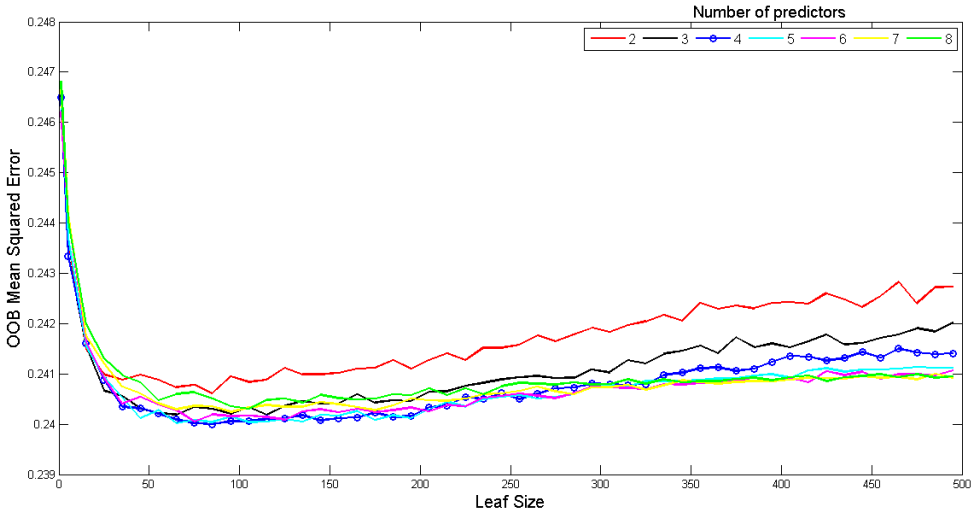

Fig. 3.1: Plot of the out-of-bag (OOB) mean-squared-error (MSE) for the period 2000:Q1 – 2010:Q4. The MSE, for a forest size of 500, is displayed as a function of leaf size. Each line corresponds to a different number of predictors available for splitting at each non-terminal node. The minimum error is achieved at a predictor and leaf size of 4 and 85 respectively.

Initially, for the purposes of model implementation, each fundamental factor input in Table 3.1 was standardized[1] and then averaged to form each of the nine explanatory variables used in this study. In proceeding with the initial training of each model, the excess stock returns first needed to be calculated. These excess returns were calculated by subtracting the benchmark return from the stock return at each quarter. Consequently, and as in the studies implemented by Sorensen *et al.* (2000), Zhu *et al.* (2011, 2012) and Giuricich (2013) a positive excess return was associated with an outperforming stock, a negative excess return with an underperforming stock. This generated categorical variable was then used as the response (dependent variable) in the model training phases.

Earlier studies by Sorensen *et al.* (2000), Zhu *et al.* (2011, 2012) and Giuricich (2013) did not consider regression models with higher order polynomials and interaction factors, thus making the robustness of their results limited to the small number of models used against which CART was compared. The linear and logistic regression models used in this research are categorized as first and second-order, where a first-order model indicates that only independent variables in the first power

[1] Each input factor (X) in the dataset is transformed by $\frac{X-E(X)}{\sigma_X}$. As a result, all the inputs have been centred (i.e. their mean is set to zero) and have standard deviations of one. Thus equal numerical importance is given to all factors.

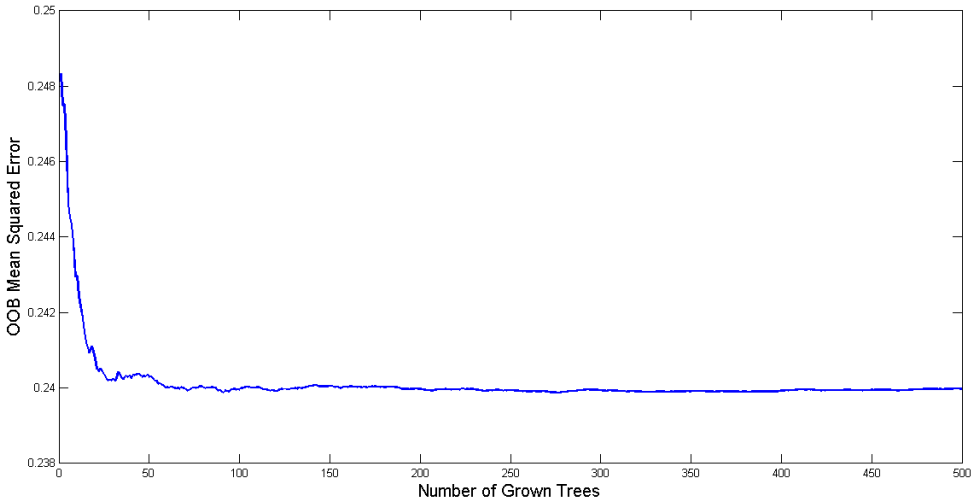

Fig. 3.2: Plot of the out-of-bag (OOB) mean-squared-error (MSE) as a function of the number of trees (forest size) for a predictor size and leaf size of 4 and 85 respectively.

are included in the model, while a second-order model allows for interactions and squared variables. Zhu *et al.* (2011) described two approaches in constructing a hybrid model, but implemented only a single approach, thus failing to empirically show whether one approach outperforms another. For this reason both approaches were implemented in this research and respectively termed Hybrid model 1 and Hybrid model 2.

Upon implementing of the random forest model as mentioned in Section 3.4, three parameters needed to be set, these being the forest size (number of trees), the number of explanatory variables to use at each split (q) and the minimum number of observations per node. For regression trees a forest size of 500 was selected as is commonly used in practice, in addition to being the recommended size for regression on many statistical software programs, such as Matlab, R, Weka and Statistica (Breiman, 2001; Liaw and Wiener, 2002; Hastie *et al.*, 2009; Cutler *et al.*, 2012).

The minimum number of observations per node (leaf size) and number of predictors q were determined in preliminary trials with the objective being to minimize the out-of-bag mean-squared-error (OOB MSE) (Breiman, 2001; Cutler *et al.*, 2012). The number of predictors needs to be less than the number of explanatory variables (i.e., which for this study is nine). By construction, once a value for q has been determined (i.e., by identifying the value of q and leaf size which minimizes the OOB

MSE), that value of q determines the number of explanatory variables, out of the nine available, that are randomly chosen for splitting at each node. A recommended starting point for the number of predictors in regression trees is $\frac{\text{no. of X variables}}{3}$ (Hastie *et al.*, 2009; Cutler *et al.*, 2012). Upon investigation of various predictor and leaf sizes as illustrated in Figure 3.1, it became clear that having a too few or too large predictor size results in a larger error rate than values in-between. It was found that a predictor size and leaf size of 4 and 85 minimized the OOB MSE respectively. From Figure 3.2 it can be seen that from a forest size of 250 onwards the OOB error rate is stable. Breiman (2001) and Hastie *et al.* (2009) showed that a large forest size should not effect the performance but rather the computation time, and that there would be no overfitting issue. Upon investigation it appeared that there was no concerning additional cost in using a forest size of 500, while increasing the forest size added no benefit to the models performance. Table 3.3 summarises the models implemented in this study.

Note that two additional models have been added to those already discussed; a random classifier and CART without pruning (CART-NP). These two models act as a "base hurdle" to benchmark the performance of the other models against.

The following subsection explains the classification metrics used to test the out-of-sample performance of each model and the portfolio measures used to test the four portfolio strategies formed on each model.

3.2.1 Classification measurement

Tab. 3.4: Confusion Matrix

Class	Predicted Negative	Predicted Positive
Actual Negative	**TN**	**FP**
Actual Positive	**FN**	**TP**

Classification models which produce continuous outputs can also be referred to as *probabilistic classifiers* (Fawcett, 2004). A probabilistic classier can be discretized by combining its output with a finite threshold, such that values above the threshold are classified as positive, else they are classified as negative. For the purposes of this study, the CART, logistic, hybrid and random classification models distinguished between outperforming stocks and underperforming stocks at a threshold value of 0.5. The linear regression models distinguished between the stocks by associating a positive excess return with an outperforming stock and a negative excess return with an underperforming stock. This induced categorical variable allowed each model to

behave as a binary classification system. For a given binary classification system, the predicted outcomes could either be positive or negative classes, denoted as class-1 and class-0 objects respectively. For the procedure four possible states exist, which include two correct predictions; True Negative (TN), True Positive (TP) and two incorrect prediction; False Negative (FN) and False Positive (FP). The outcomes can be summarized in Table 3.4, commonly referred to as the confusion matrix.

In order to asses the performance of the models in this dissertation, various classification metrics were calculated for the testing period 2011:Q1 – 2014:Q3.

- **Accuracy:** this measure can be interpreted as the probability of correct classification. It is defined as the sum of the number of true positive and true negatives, divided by the total number of tested objects (Lundstrom, 2013).

$$\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{3.2}$$

- **Specificity:** refers to the proportion of real negative values correctly classified as negative (i.e., the probability of correctly classifying a class-0 object correctly) (Lundstrom, 2013). It is defined as the number of true negatives divided by the sum of the number of true negatives and false positives.

$$\frac{\text{TN}}{\text{TN+FP}} \tag{3.3}$$

- **Precision:** can be thought of as, of all the observations the algorithm classified as a class-1 object, what fraction were actually class-1 objects. It is defined as the number of true positives divided by the number of predicted positives (Lundstrom, 2013).

$$\frac{\text{TP}}{\text{TP+FP}} \tag{3.4}$$

- **Recall (Sensitivity):** refers to the probability of correctly classifying a class-1 object correctly (Lundstrom, 2013). It can also be thought of as, of all the observations that are class-1 objects, what fraction did the algorithm correctly detect as a class-1 object. It is defined as the number of true positives divided by the number of actual positives.

$$\frac{\text{TP}}{\text{TP+FN}} \tag{3.5}$$

The above metrics are among the most commonly used in practice (see Chen *et al.*, 2004; Davis and Goadrich, 2006; Sokolova *et al.*, 2006; Tang *et al.*, 2009).

A study by Chen *et al.* (2004) explained that the use of accuracy as a standalone metric is inappropriate, because a trivial (non-learning) classifier can achieve high accuracy by simply predicting the majority class for each observation, as a result rendering the usefulness of accuracy quite low. For that reason, another measure termed *balanced accuracy* was included, which is given by

$$\frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{P}}+\frac{\mathrm{TN}}{\mathrm{N}}\right), \tag{3.6}$$

where P=FN+TP and N=FP+TN and the full expression represents the average between the true positive and true negative rate. Brodersen *et al.* (2010) described it to be a more robust metric than accuracy especially when the dataset is imbalanced. When a given class is more prominent in a testing set it is known as a skewed (imbalanced) dataset—it was found that a third of the observations in the testing data (true values) used in this study were class-1 objects, and the remainder class-0 objects. In this situation a Precision-Recall (PR) curve is considered to provide more information about a classifier's performance (Davis and Goadrich, 2006; Murphy, 2007; Tang *et al.*, 2009). There is an inverse relationship between precision and recall, and a PR curve plots the precision and recall of a classification algorithm as a function of the threshold. In order to compare classifiers the area under the PR curve (PR-AUC) is calculated. A higher PR-AUC indicates a better classifier (Murphy, 2007; Tang *et al.*, 2009). The usefulness of PR-AUC as a metric is emphasized by it not being restricted to a single threshold unlike the aforementioned classification metrics. In addition to the above classification metrics, the performance of each model were also assessed on mean-squared error (MSE). Minimising the MSE forms the basis on which regression models were constructed in this study, making it a natural measure on which to assess a model's quality in a regression setting (James *et al.*, 2013).

3.2.2 Portfolio measurement

In addition to the classification metrics used to compare the out-of-sample model performance, for each model, four portfolio strategies were formed based on each model's forecast.

- **Long portfolio strategy:** a portfolio comprising of long positions in the stocks forecasted to outperform, with weights for the linear regression models given by their excess return, and for the remaining models given by their forecasted probability of outperformance. Another long portfolio was formed by enforcing an equally weighted strategy in each stock forecasted to outperform.

- **Long-plus-short portfolio strategy:** a portfolio comprising of short positions in the stocks forecasted to underperform while simultaneously going long on the benchmark, with weights for the linear regression models given by their excess return, and for the remaining models given by their forecasted probability of underperformance. Another long-plus-short portfolio was formed by enforcing equally weighted strategy in each stock forecasted to underperform.

The long-plus-short strategy, also known as a market neutral strategy strips out the exposure to market swings (systematic risk), hence delivering positive returns regardless of the market direction (Jacobs *et al.*, 1999). That is in contrast to a long-only portfolio, which has full exposure to market movements. The objective of the long-plus-short strategy is to deliver returns in excess of some proxy of the market risk-free rate. Firer and McLeod (1999) and Strydom and Charteris (2009) suggested that an three month Negotiable Certificate of Deposit (3M-NCD) rate is a suitable proxy for a South African risk-free rate.

The portfolios were rebalanced quarterly, with profits being reinvested and transaction costs incorporated at each quarter. Transaction costs of 0.2% nominal traded were factored in at each rebalancing period. According to Yu (2008) and Kruger (2011) a 20 basis point (bps) transaction fee is a conservative proxy in South Africa.

To assess the performance of these portfolios, various portfolio performance metrics were used. These metrics included the annualised portfolio return, annualised excess portfolio return, mean excess portfolio return, volatility, tracking error, Sharpe ratio, information ratio and Sortino ratio (see Table A.1 in Appendix A for formulae).

Chapter 4

Model comparison and discussion

The following section can be divided into roughly four subsections. The first deals with a discussion around the output of the hybrid model. The second provides an analysis of the classification and performance metrics of each model. Thirdly, a brief look at what the models looked like before and after the 2007/08 Global Financial Crisis. Finally, conclusions are drawn.

4.1 Hybrid model results

As outlined by Zhu *et al.* (2011), the hybrid model extends the dynamic CART approach proposed by Sorensen *et al.* (2000), by building an evolving tree model and adjusting its CART-based probabilities at each period using a logistic model. Firstly, for the purposes of this study, a regression tree was trained using data from 2000:Q1 through 2010:Q4 and re-estimated at each quarter moving forward, where each tree was used to forecast the probability of outperformance for each subsequent quarter. Finally, the tree-based probabilities were modified using stock-specific information through a logistic regression model. As in the study by Zhu *et al.* (2011), the AIC variable selection procedure was used to choose the explanatory variables included in the logistic model. For illustrative purposes of Hybrid model 2, the regression tree estimated up to 2014:Q2 is shown in Figure 4.1, and the coefficients of the logistic model estimated as at 2014:Q2 is shown in Table 4.1. Zhu *et al.* (2011) explained that through the use of a single logistic regression model any global features of the dataset that CART may have overlooked should now be incorporated.

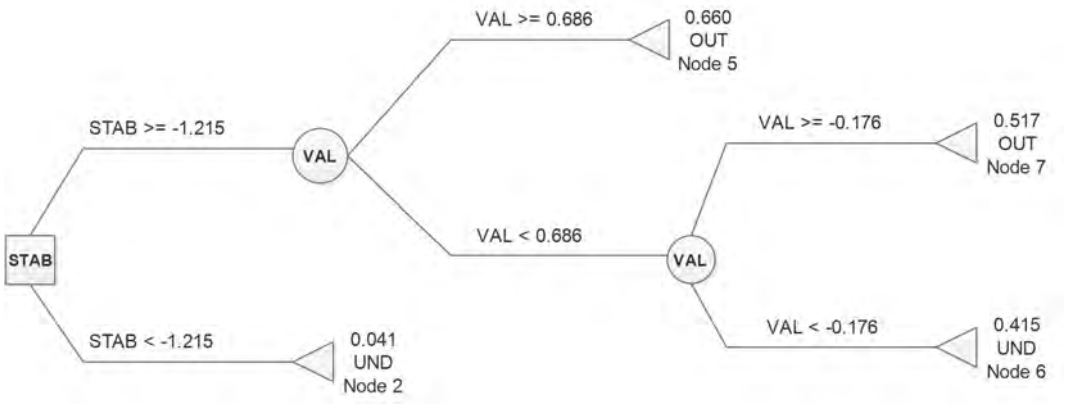

Fig. 4.1: CART model for the universe of stocks for the period 2000:Q1 – 2014:Q2. Note: moving upward in the tree represents a right-hand split and downward a left-hand split. While UND stands for underperforming and OUT outperforming.

Tab. 4.1: Variables from the logistic regression model (including offset parameters) estimated for the period 2000:Q1 – 2014:Q2. Significance levels: 0–0.001 '***'; 0.001–0.01 '**'; 0.01–0.05 '*'; 0.05–0.1 '·'.

X **Variable**	β **Coefficient**	**P-value**
VAL	0.487	0.000***
DEBT.SERVICE	0.184	0.001**
FWD.GROWTH	0.131	0.004**
EREV	0.092	0.044*
STAB	-0.095	0.057·
HIST.GROWTH	0.105	0.060·
LEVERAGE	0.097	0.082·

It was clear from examining Figure 4.1, the primary split was on Stability, distinguishing between stocks of low stability (high volatility in earnings, sales and operating cashflows over the previous 3 years) and high stability. Additionally, the explanatory variable Value appears at two non-terminal nodes with two different splitting values. However simplistic it may be, the regression tree classified stocks with low stability (high volatility) and high value as outperformers. Also, it is evident that CART identified Value as an important splitting determinant in distinguishing between out/underperforming stocks as it occurred at two consecutive levels in the tree.

Seven variables were selected for the logistic regression model using the AIC criteria. Value, Earnings Revenue, Profitability, which are analogous to those selected in the study by Zhu *et al.* (2011), in addition to Leverage, Stability, Historical and Forward growth. All the selected variables were significant at a 10% level of significance, with most significant at 5%. The p-values are not of real interest as the explanatory variables were not selected using them, but it was comforting to note that no extremely insignificant variables were included in the model. From the coefficients selected by the logistic regression model it was clear that CART did not capture the effects of up to five predictors, emphasizing the benefit of blending the two models.

Upon further investigation on the capabilities of the hybrid approach, it was clear that the CART model provided only four unique probabilities (as illustrated in Figure 4.1). For this particular dataset and period, it was found that by imposing the logistic model on top of the CART, over sixty unique probabilities were produced. This allowed for a more diversified weighting strategy when creating probability weighted portfolios as opposed to equally weighted portfolios as implemented by (Sorensen *et al.*, 2000; Zhu *et al.*, 2011, 2012).

4.2 Out-of-sample model comparison

4.2.1 Classification metrics

The following analysis is based on the information exhibited in Table 4.2. On balanced-accuracy, the linear regression models displayed the strongest performance. More specifically, a dynamic first-order linear model displayed the strongest performance. However, the Hybrid model 2 showed promising performance, deviating from the first-order linear model by less than a percent. What stood out is that each model, both static and dynamic, had a much higher specificity than sensitivity, implying that each model was more likely to identify an underperforming stock correctly than a outperforming stock. On this basis, CART performed the poorest in its ability to correctly identify outperforming stocks and the strongest in correctly identifying underperforming stocks. However, the robustness of these results is limited to the finite threshold upon which they were calculated, making them highly sensitive to any threshold changes.

For that reason more attention was given to the PR-AUC metric. The PR-AUC of 0.33 for the random classifier was proportional to the class imbalance exhibited in the testing set (i.e., a third of the testing set was class-1 objects (outperformers) and the remaining two-thirds was class-0 objects (underperformers)). This value acted as an artificial benchmark to compare the other models against, with a better model

Tab. 4.2: Accuracy, Balanced Accuracy, Specificity, Recall(Sensitivity), Precision, Precision Recall - Area under curve (PR-AUC) and Mean-Sqaured Error (MSE) for each of the static and dynamic models. These are calculated for the testing period 2011:Q1 – 2014:Q3.

Category	Model	Accuracy	Balanced Accuracy	Specificity	Recall(Sensitivity)	Precision	PR-AUC	MSE
	Random classifier	0.493	0.512	0.489	0.502	0.329	0.326	0.339
Static	CART-NP	0.640	0.569	0.783	0.355	0.450	0.314	0.307
	CART	0.664	0.515	0.963	0.067	0.473	0.391	0.239
	Random forest	0.663	0.558	0.874	0.242	0.490	0.420	0.235
	Linear-1	0.656	0.565	0.838	0.291	0.473	0.485	0.007*
	Logistic-1	0.662	0.552	0.882	0.223	0.486	0.481	0.132
	Linear-2	0.660	0.561	0.857	0.266	0.481	0.487	0.007*
	Logistic-2	0.659	0.554	0.871	0.237	0.478	0.483	0.131
Dynamic	CART-NP	0.640	0.528	0.804	0.311	0.443	0.318	0.294
	CART	0.658	0.558	0.917	0.139	0.456	0.399	0.232
	Random forest	0.665	0.552	0.881	0.233	0.494	0.422	0.232
	Linear-1	0.673	0.571	0.877	0.265	0.518	0.532	0.006*
	Logistic-1	0.670	0.558	0.896	0.220	0.513	0.511	0.114
	Linear-2	0.666	0.569	0.861	0.276	0.499	0.503	0.007*
	Logistic-2	0.664	0.566	0.861	0.271	0.494	0.497	0.120
	Hybrid-M1	0.659	0.553	0.873	0.233	0.478	0.414	0.233
	Hybrid-M2	0.664	0.570	0.853	0.286	0.493	0.425	0.234

being associated with a higher PR-AUC. The results under PR-AUC suggest that the linear and logistic models dominated over the tree-based models. A random forest performed better than the CART model, highlighting the benefits of using more than one tree for predictive modeling. The static second-order linear and logistic models provided only a marginal improvement over there first-order alternatives, while the dynamic second-order linear and logistic models provided no improvement over there first-order alternatives, suggesting that the added complexity of regression models with higher order terms and interactions may have overfitted this dataset.

In an attempt to assess how well model predictions match the actual data, the MSE metric was next considered. Note that the linear regression models did not allow for a fair comparison against the other models, as the linear regression models made direct use of a stock's excess return relative to the benchmark as a response variable in their training phases. With that in mind, it might well have been expected for the linear regression models to achieve the lowest overall error. Focusing on the remaining models, the logistic regression models performed the best, more specifically the dynamic first-order logistic model exhibited the lowest error. Analogous to the observation made on the classification metrics, a random forest model showed an improvement over the CART model. Since the objective of the hybrid models was to simply combine a logistic and CART model and not minimize the MSE, it was not an unusual observation for the hybrid models to produce a higher error than its constituents.

It was no surprise that a pruned CART performed better than an unpruned one (CART-NP) under all the metrics. This finding supported the principle of parsimony for both the dynamic and static models. It was also found that a random classifier and CART-NP model were indeed poor models. In summary, it was clear that the dynamic models were better classifiers than their static counterparts. The linear models showed the strongest performance, however this can be attributed to the linear models being trained on more complete information regarding a stock's performance relative to the benchmark. Also, the logistic and hybrid models showed promising performance.

4.2.2 Portfolio analysis

While the performance of the dynamic models' was encouraging when measured using the classification metrics, the performance of each model was now compared under a simulated portfolio setting. This approach was similar to that implemented by Sorensen *et al.* (2000), Zhu (2012) and Giuricich (2013). However, those studies did not consider transaction costs and as a result the conclusions drawn may not have truly reflected the performance of a model in an actual market setting. In

order to assess the performance of the stocks each model predicted to outperform, a long portfolio of those outperforming stocks was constructed with transaction costs of 0.2% nominal traded being incorporated at each quarter. Note that, analogous to the studies implemented by (Sorensen *et al.*, 2000; Zhu, 2012; Giuricich, 2013), for the logistic, hybrid, CART and random forest model, stocks with a probability of outperformance above 50 percent were classified as outperformers, otherwise they were classified as underperformers. The linear regression models associated a positive excess return with an outperforming stock and a negative excess return with an underperforming stock. Note that the predictor sizes and leaf sizes for the dynamic random forest were re-estimated at each quarter. This information added no benefit to the long-only portfolio, however it did benefit the long-plus-short strategy. For that reason, when implementing the dynamic random forest, the parameters for the dynamic long-only portfolios were held constant at a predictor size and leaf size of 4 and 85 respectively. While they were re-estimated at each quarter for the dynamic long-plus-short portfolios.

The following analysis was based on the information exhibited in Table 4.3, which reported the annualised return, annualised excess return, annualised mean excess return, the Sharpe ratio, the information ratio, the tracking error, the Sortino ratio and the annualised portfolio return volatility for each of the static and dynamic models. (Note: Table 4.3 – Table 4.6 are displayed at the end of this section).

From the probability weighted long portfolios (Table 4.3), it was evident that on a non-risk adjusted basis (i.e., mean excess return) all but one of the dynamic models considerably outperformed their static counterparts. Even though this does not hold for the random forest, such an observation may have been due to chance since the excess return was very close to zero. Focusing on the dynamic models, the top three on a non-risk adjusted basis were the first-order linear model (9.69%), the CART model (8.01%) and the first-order logistic model (5.98%).

The Sharpe and Sortino ratios were used together to evaluate the performance of the models on a risk adjusted basis. The standard deviation used in calculating the Sharpe ratio does not discriminate between "good" or "bad" volatility, while the Sortino ratio caters for this by considering only the "harmful" (downside) volatility (Sortino and Price, 1994). Note that the Sharpe ratio and Sortino ratio for the benchmark (JSE ALSI) is 1.31 and 1.77 respectively. On a risk adjusted basis, the dynamic first-order logistic and linear models are the most successful at converting the risk taken on, into a return. A static random forest outperformed CART on a risk-adjusted basis, as the returns exhibited by the CART model were greatly offset by its high volatility. This observation was not consistent with the dynamic setting. For both of the static and dynamic categorizations, the first-order linear

and logistic regression models outperformed their second-order counterparts on a risk and non-risk adjusted basis, analogous to the observation made on the classification performance of each model.

This apparent superiority of the first-order regression models suggests that the added complexity of interaction and squared factors reduced the models predictive ability, even though theoretically it was expected that these second-order models would not lead to overfitting as only the most important variables were added to the model. What is particularly encouraging is that both of the hybrid models exhibited strong risk adjusted performance, outperforming a stand-alone CART in addition to the benchmark. It was also evident that the Hybrid model 2 convincingly outperformed the Hybrid model 1 on a risk and non-risk adjusted basis even with a near identical level of risk (Volatility).

In order to evaluate whether the risk-level assumed for each model relative to the benchmark was sufficiently rewarded, the tracking error and information ratio were analysed. Most of the model tracking errors were found to be relatively stable at between five to eight percent, with CART-NP having the lowest tracking error, which was suggestive of a lower risk-level relative to the benchmark. The first-order linear and logistic models more consistently outperformed the benchmark, with information ratios greater than 1. Additionally, the first-order linear and logistic models achieved this level of consistency with tracking errors only marginally different to the other models. The high tracking errors associated with the CART models could be attributed to the bigger positions (or "bets") taken relative to the benchmark. These large bets could be a chief reason for the strong non-risk adjusted performance of the CART models.

In order to gauge the degree of deviation between a portfolio's holdings relative to a benchmark, the "Active Share" was examined. Active share was popularised through the work of Cremers and Petajisto (2009) as a new measure for active management. The active share is defined as

$$\text{Active Share} = \frac{1}{2}\sum_{i=1}^{n} |w_{portfolio,i} - w_{benchmark,i}|, \tag{4.1}$$

where $w_{portfolio,i}$ is the weight of stock i in the portfolio, $w_{benchmark,i}$ is the weight of the same stock in the benchmark index. Intuitively, the active share describes the fraction of the portfolio that is invested differently to its benchmark (Cremers and Petajisto, 2009).

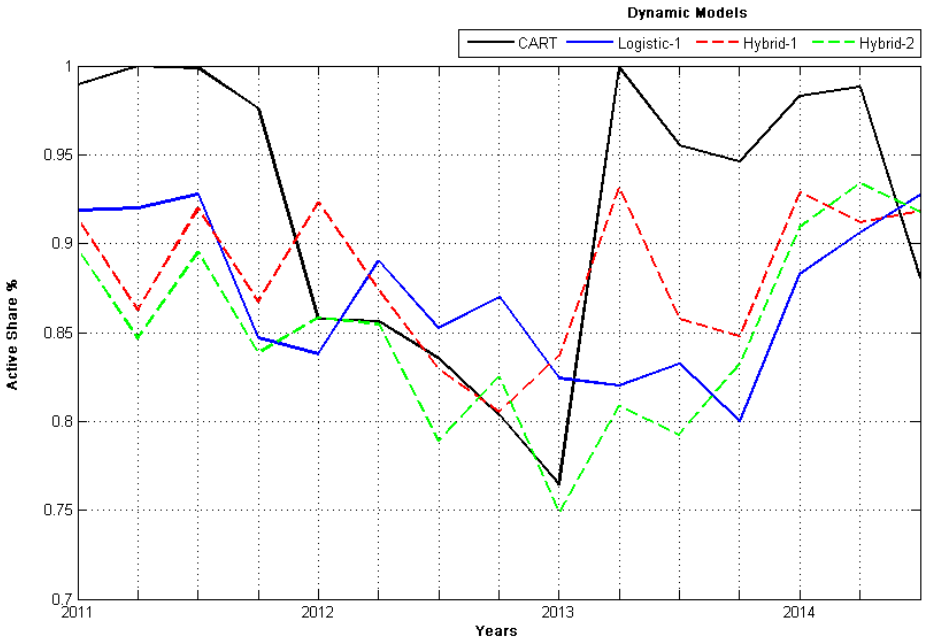

Fig. 4.2: Active share for the dynamic first-order logistic model, the dynamic CART model, Hybrid model 1 and Hybrid model 2 for the period 2011:Q1 – 2014:Q3 (Long-only strategy).

Figure 4.2 shows the Active share as a function of time for the hybrid models and their constituents. What is immediately apparent is that the holdings in CART deviated significantly from the benchmark, in some cases by up to 100% — a chief reason for this being that the actual number of stocks in the CART-based long portfolio only ranged between 1 and 20, over the testing period. From a holdings perspective it would not be feasible to have such high concentrations in a single stock, which is indicated by the high tracking error and volatility of the CART-based long portfolio. The hybrid models exhibited a less erratic active share than CART, with stock holdings ranging between 20 to 60 over the testing period. Even though the hybrid models might not have outperformed the first-order logistic model, they did show an improvement over CART, allowing for a more diversified stock holding strategy.

On considering the equally weighted long portfolio in Table 4.4, similar observations can be made as compared with the probability weighted long portfolios. Hence, more attention will be given to the dissimilar results. Firstly, attention must be given to the observation that the performance of the equally weighted and probability weighted CART-based long portfolios had identical performance. Upon

investigation, this was attributed to the small trees formed in training the CART model. Only a single node in the tree had a probability of outperformance above 50 percent, hence all the stocks predicted to outperform (i.e., the stocks chosen for long positions) were allocated the same probability. As a result, regardless of whether the portfolio was equally weighted or probability weighted they would be identical. This highlights one of the drawbacks of the CART model: depending on the tree size, the probability space may become very discrete.

The portfolios based on the linear regression models no longer exhibited the same strong risk and non-risk adjusted performance observed in the probability weighted setting — a chief reason for this was that these models no longer had the magnitude of excess returns as weights. That is, in training the linear models the response variable was excess stock return, relative to the benchmark. Hence the predicted values from the linear regression models reflected more complete information regarding stock performance. This may have contributed to strong portfolio performance when weighting by those predicted values, while in an equally weighted portfolio setting the models were on a more level playing field. However, as in the probability weighted setting the dynamic first-order logistic model was more successful at taking on risk relative to the benchmark (i.e., higher information ratio). Analogous to the observation made on the classification metrics, the CART-NP and random classifier performed the poorest regardless of the performance metric for both portfolios, which is not unusual as the former violates the principle of parsimony while the latter is simply a non-learning algorithm.

The natural extension would be to assess each model's ability in selecting underperforming stocks. Upon reviewing both the probability and equally weighted long-plus-short strategy (Table 4.5 and Table 4.6), most of the dynamic models outperformed their static counterparts. In support of the literature touching on the benefits of a random forest over CART, the random forest did indeed outperform CART regardless of the metric used. The hybrid models offered enhanced risk and non-risk adjusted performance over a stand-alone CART. However, as in the long portfolios, the dynamic linear and logistic regression models exhibited the strongest overall performance, in terms of consistency over beating the benchmark (i.e., higher information ratio) and risk and non-risk adjusted performance. Notably, although the volatility for the long-plus-short portfolios were in some instances marginally lower than the long-only strategies, the risk-adjusted performance on the long-plus-short portfolios were quite low. Typically when implementing a long-plus-short strategy one would expect to have a lower volatility and hence a lower return than a long-only strategy (Northern Trust Global Investments, 2012). In this study the returns of the long-plus-short strategies are lower, but the volatility of these re-

turns are still relatively indifferent to the volatility of the returns in the long-only strategies.

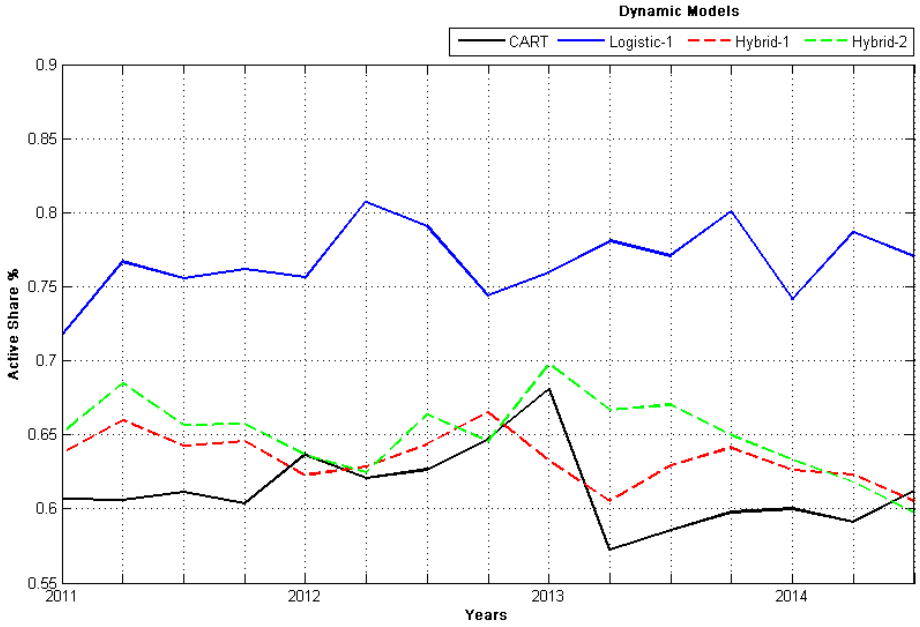

Fig. 4.3: Active share for the dynamic first-order logistic model, the dynamic CART model, Hybrid model 1 and Hybrid model 2 for the period 2011:Q1 – 2014:Q3 (Long-plus-short strategy).

From a holdings perspective, having a portfolio consisting of very few stocks is more risky in a short portfolio setting than in a long setting. Taking a short position in stock exposes the investor to unlimited downside risk, even more so if the bets are concentrated in only a few stocks. Upon investigating the amount of stocks each model had in their respective portfolios, no concentrated holdings were found as opposed to in the CART-based long portfolio (i.e., stock holdings ranged between 35–125 over the period in question). With reference to the active share, it is clear from Figure 4.3 that the hybrid models benefited by having a more diversified portfolio relative to CART, while still outperforming CART regardless of the metric.

In summary, the portfolio analysis showed that dynamic models outperformed their static counterparts. Overall, the logistic regression model displayed the strongest performance, with a linear regression model coming in close, regardless of the performance metric. In most instances the random forest approach outperformed CART on a risk-adjusted basis, however this did not hold on a non-risk adjusted basis. This disappointing performance may be partly due to overfitting. Segal (2004) showed

that overfitting is more pronounced in random forests when performing noisy regression or classification tasks. In addition the findings by Gashler *et al.* (2008) showed that a random forest is less adept at handling a large number of irrelevant features of a dataset as opposed to a collection of entropy-reducing decision trees.

In agreement with the observation made by Zhu *et al.* (2011), the hybrid model offered an improvement over a stand-alone CART model. However, in contrast to the observations made by Zhu *et al.* (2011), the hybrid approach for this study (i.e., for the particular time–frame and dataset used in this study) was unable to outperform a logistic regression model. More specifically, the Hybrid model 2 offered the greatest improvement to CART. This observation supports the explanation provided by Zhu *et al.* (2011), that the second hybrid model would be more robust than the first. The advantages of the hybrid model as claimed by Zhu *et al.* (2011) include its ability to produce a smoother probability surface by incorporating a logistic model, as opposed to the highly discrete surface produced by a stand-alone CART. By imposing a single logistic model the second hybrid approach is able to pick up any global features of the dataset uncaptured by CART. This benefit over CART is largely due to the so called "greedy" recursive partitioning algorithm used to construct a CART model.

From an implementation and interpretation perspective, the hybrid model may not be a "black-box" method. The interpretable output of CART together with the refinement of observation specific probabilities using a logistic model was found to not only overcome the local optimum problem of CART but also outperform a random forest approach. This is interesting seeing that a random forest is one of the more popular tree-based machine learning techniques used in practice (see Kumar and Thenmozhi, 2006; Strobl *et al.*, 2008; Grömping, 2009; Su *et al.*, 2011; Qi, 2012).

A closer inspection on the different attributes exhibited by the CART, logistic and linear model over two different time–frames revealed that the information in the training set may have poorly reflected the information in the testing set. The time–frames upon which each model was fitted was 2000–2007 and 2008–2014. The trees of the CART model for these two periods not only had different shapes but also had very different splitting values. The main splitting value for CART was Value pre–2008 and Stability post–2008. Also the tree post–2008 was more skewed to the left, as illustrated in Figure B.1 and Figure B.2 in Appendix B. The linear and logistic regression models also exhibited different attributes in terms of the variables selected and their factor loadings (beta coefficients) (see Table B.1 – B.8 in Appendix B). Similar observations were found for the specific training and testing set used in this study. The training set for this research (i.e., from 2000:Q1 – 2010:Q4), was selected so that the models could be given information during and

just after the Global Financial Crisis, and to use that information to identify changes in trends or patterns and apply that to the testing set. It would remain for further research to identify whether each of the models in this study, based on either their splitting values or selected variables, could provide information on the time–frame or economic regime they are in. Such an analysis would warrant the inclusion of economic indicators as explanatory variables, so that the modeler could identify the important splits and variables as well as their impact over different time–frames.

Tab. 4.3: Annualised return, annualised excess return, annualised mean excess returns, Sharpe ratios, Information ratios, tracking errors, Sortino ratios and annualized return volatility for each of the static and dynamic models **long probability weighted portfolio** strategy. Portfolios are rebalanced quarterly with a 20bp transaction fee. Note that the benchmark is the JSE ALSI and 3M-NCD is used as a proxy for the risk-free rate. These are calculated for the testing period 2011:Q1 – 2014:Q3. The annualised return, Sharpe ratio, Sortino ratio and volatility for the benchmark is 19.71%, 1.31, 1.77 and 10.14% respectively.

Category	Model	Ann. Ret. (%)	Ann. Exc. Ret. (%)	Mean Exc. Ret. (%)	Sharpe Ratio	Information Ratio	TE (%)	Sortino Ratio	Vol (%)
	Random classifier	14.94	-4.77	-4.31	1.04	-0.83	5.34	1.41	8.54
Static	CART-NP	13.57	-6.15	-5.36	0.72	-0.82	7.19	0.98	10.22
	CART	22.41	2.70	2.76	1.15	0.16	13.84	1.62	13.45
	Random forest	19.17	-0.54	-0.53	1.34	-0.06	7.80	1.84	9.58
	Linear-1	22.45	2.74	2.27	1.82	0.35	7.81	2.50	8.78
	Logistic-1	22.81	3.10	2.53	2.01	0.46	6.77	2.77	8.15
	Linear-2	20.63	0.92	0.96	1.21	0.13	5.73	1.76	11.53
	Logistic-2	20.99	1.28	0.96	1.74	0.22	6.39	2.40	8.43
Dynamic	CART-NP	14.72	-4.99	-4.42	0.91	-0.89	5.29	1.32	9.41
	CART	28.18	8.47	8.01	1.41	0.49	14.99	1.93	14.67
	Random forest	19.14	-0.57	-0.57	1.37	-0.07	6.44	1.85	9.34
	Linear-1	30.77	11.05	9.69	2.32	1.64	6.33	3.03	10.17
	Logistic-1	26.74	7.03	5.98	2.43	1.22	5.60	3.28	8.25
	Linear-2	23.14	3.43	3.11	1.47	0.57	5.52	2.00	11.17
	Logistic-2	21.22	1.51	1.19	1.69	0.28	5.50	2.26	8.80
	Hybrid-M1	19.20	-0.51	-0.55	1.42	-0.07	5.53	2.06	9.07
	Hybrid-M2	20.61	0.90	0.70	1.53	0.17	5.56	2.20	9.28

Tab. 4.4: Annualised return, annualised excess return, annualised mean excess returns, Sharpe ratios, Information ratios, tracking errors, Sortino ratios and annualized return volatility for each of the static and dynamic models **long equally weighted portfolio** strategy. Portfolios are rebalanced quarterly with a 20bp transaction fee. Note that the benchmark is the JSE ALSI and 3M-NCD is used as a proxy for the risk-free rate. These are calculated for the testing period 2011:Q1 – 2014:Q3. The annualised return, Sharpe ratio, Sortino ratio and volatility for the benchmark is 19.71%, 1.31, 1.77 and 10.14% respectively.

Category	Model	Ann. Ret. (%)	Ann. Exc. Ret. (%)	Mean Exc. Ret. (%)	Sharpe Ratio	Information Ratio	TE (%)	Sortino Ratio	Vol (%)
	Random classifier	14.84	-4.87	-4.39	1.02	-0.86	5.22	1.38	8.60
Static	CART-NP	14.13	-5.59	-4.97	0.88	-0.79	6.61	1.21	9.12
	CART	22.41	2.70	2.76	1.15	0.16	13.84	1.62	13.45
	Random forest	19.17	-0.54	-0.53	1.35	-0.06	7.76	1.86	9.51
	Linear-1	17.99	-1.72	-1.57	1.25	-0.24	6.42	1.78	9.39
	Logistic-1	22.50	2.78	2.25	1.99	0.43	6.60	2.75	8.11
	Linear-2	18.10	-1.62	-1.56	1.41	-0.28	4.97	2.00	8.46
	Logistic-2	20.65	0.94	0.67	1.69	0.17	6.38	2.34	8.47
Dynamic	CART-NP	14.57	-5.14	-4.55	0.89	-0.90	5.37	1.30	9.47
	CART	28.27	8.56	8.09	1.41	0.50	15.00	1.93	14.67
	Random forest	19.03	-0.68	-0.68	1.38	-0.09	6.32	1.87	9.23
	Linear-1	23.02	3.30	2.79	1.82	0.54	5.95	2.43	9.08
	Logistic-1	26.20	6.49	5.50	2.39	1.11	5.69	3.25	8.20
	Linear-2	18.79	-0.92	-0.88	1.33	-0.17	4.74	1.80	9.38
	Logistic-2	20.49	0.77	0.56	1.58	0.15	5.57	2.13	8.94
	Hybrid-M1	17.78	-1.94	-1.77	1.24	-0.31	5.58	1.82	9.28
	Hybrid-M2	19.23	-0.49	-0.51	1.40	-0.06	5.61	2.08	9.23

Tab. 4.5: Annualised return, annualised excess return, annualised mean excess returns, Sharpe ratios, Sortino ratios and annualized return volatility for each of the static and dynamic models **long-plus-short probability weighted portfolio** strategy. Portfolios are rebalanced quarterly with a 20bp transaction fee. Note that the benchmark is the JSE ALSI and 3M-NCD is used as a proxy for the risk-free rate. These are calculated for the testing period 2011:Q1 – 2014:Q3. Note that the tracking error and information ratios have been left out since the benchmark is the risk-free rate.

Category	Model	Ann. Ret. (%)	Ann. Exc. Ret. (%)	Mean Exc. Ret. (%)	Sharpe Ratio	Sortino Ratio	Vol (%)
	Random classifier	5.41	-0.04	0.30	-0.21	-0.35	8.70
Static	CART-NP	5.66	0.21	0.70	-0.15	-0.27	10.44
	CART	7.73	2.29	2.59	0.07	0.12	9.16
	Random forest	7.94	2.49	2.75	0.10	0.19	8.77
	Linear-1	10.23	4.79	4.96	0.38	0.62	8.61
	Logistic-1	8.91	3.46	3.72	0.19	0.32	9.30
	Linear-2	11.12	5.67	6.08	0.34	0.53	11.54
	Logistic-2	8.25	2.80	3.10	0.12	0.19	9.44
Dynamic	CART-NP	5.19	-0.26	0.20	-0.20	-0.35	9.87
	CART	7.21	1.76	2.07	0.01	0.02	8.95
	Random forest	7.75	2.31	2.59	0.08	0.14	9.00
	Linear-1	12.11	6.66	6.88	0.51	0.79	9.96
	Logistic-1	10.47	5.03	5.25	0.36	0.62	9.46
	Linear-2	11.96	6.52	6.85	0.44	0.68	10.99
	Logistic-2	8.26	2.81	3.11	0.12	0.22	9.37
	Hybrid-M1	7.32	1.87	2.19	0.02	0.04	9.14
	Hybrid-M2	7.86	2.41	2.73	0.08	0.14	9.41

Tab. 4.6: Annualised return, annualised excess return, annualised mean excess returns, Sharpe ratios, Sortino ratios and annualized return volatility for each of the static and dynamic models **long-plus-short equally weighted portfolio** strategy. Portfolios are rebalanced quarterly with a 20bp transaction fee. Note that the benchmark is the JSE ALSI and 3M-NCD is used as a proxy for the risk-free rate. These are calculated for the testing period 2011:Q1 – 2014:Q3. Note that the tracking error and information ratios have been left out since the benchmark is the risk-free rate.

Category	Model	Ann. Ret. (%)	Ann. Exc. Ret. (%)	Mean Exc. Ret. (%)	Sharpe Ratio	Sortino Ratio	Vol (%)
	Random classifier	6.17	0.72	1.05	-0.11	-0.21	8.87
Static	CART-NP	6.05	0.60	1.06	-0.11	-0.22	10.30
	CART	7.87	2.42	2.70	0.09	0.16	8.95
	Random forest	8.35	2.90	3.15	0.15	0.28	8.71
	Linear-1	7.70	2.25	2.56	0.06	0.11	9.23
	Logistic-1	8.92	3.47	3.72	0.19	0.33	9.16
	Linear-2	8.07	2.62	3.01	0.09	0.13	10.31
	Logistic-2	8.50	3.05	3.33	0.15	0.24	9.31
Dynamic	CART-NP	5.27	-0.17	0.27	-0.19	-0.34	9.70
	CART	7.50	2.05	2.33	0.05	0.08	8.84
	Random forest	8.21	2.77	3.03	0.13	0.24	8.92
	Linear-1	10.51	5.06	5.30	0.36	0.59	9.60
	Logistic-1	10.66	5.21	5.44	0.37	0.64	9.51
	Linear-2	8.98	3.53	3.87	0.19	0.31	10.05
	Logistic-2	8.53	3.08	3.39	0.15	0.26	9.61
	Hybrid-M1	7.68	2.23	2.52	0.07	0.11	8.99
	Hybrid-M2	8.59	3.15	3.42	0.17	0.29	9.20

Chapter 5

Conclusion

The purpose of this research was to compare the performance of non-linear techniques in stock selection. The performance of a CART, random forest, first/second-order logistic and linear regression and two hybrids of a logistic and CART model were assessed. From examining both classification and portfolio performance metrics, this study showed that a dynamic modeling approach outperformed a static approach. The logistic regression model exhibited the overall strongest performance on both metrics, producing superior results to the random forest and hybrid models. The Hybrid model 2 exhibited a significant improvement over a stand-alone CART model.

Based on the portfolio analysis it was found that all the models performed better (on a risk-adjusted basis) in a long-only strategy than they did in a long-plus-short strategy (market neutral strategy). This would imply that, for the testing period used in this study (i.e., 2011:Q1 – 2014:Q3) the model's were to an extent more successful in identifying outperforming stocks than underperforming stocks. In turn, that suggests that these models can aid in an investors portfolio construction process. The model's output can be used to reinforce the types and sizes of bets an investor takes based on his own fundamental analysis. An asset manager can use the magnitude of predicted outperformance from each model as a multiple of how much to either overweight or underweight a stock relative to the index. "Good models only add strength to an individual person, just like adding wings to a tiger" (Wei, 2012).

It may be noted that the results presented in this research were a function of the specific input data utilised and the time-frame over which the performance of each model was assessed: the time-frames used for the training sets may not have been optimal (e.g. by spanning over known regime changes) and thus an interesting avenue for further research would be to find a more optimal window length for this training sample. In this research the process of forecasting a stocks future performance was restricted to the use of company fundamental data as the

primary source of predictor variables. McConnell *et al.* (1986) found that qualitative data, such as macroeconomic variables and management factors can also assist in forecasting a stocks future performance. A study by Lahmiri (2011) used technical indicators and macroeconomic variables in predicting stock market trends. It would remain for future research to identify the optimal blend between qualitative and company fundamental data in South Africa that will efficiently aid in forecasting a companies' future stock performance.

Bibliography

Ang, A. and Bekaert, G. (2007). Stock return predictability: Is it there?, *Review of Financial studies* **20**(3): 651–707.

Ang, J. S. and Ciccone, S. J. (2001). Analyst forecasts and stock returns.
URL: *http://ssrn.com/abstract=271713*

Arnold, T. W. (2010). Uninformative parameters and model selection using Akaike's information criterion, *The Journal of Wildlife Management* **74**(6): 1175–1178.

Aronson, D. (2011). *Evidence-based technical analysis: applying the scientific method and statistical inference to trading signals*, Vol. 274 p. 29, John Wiley & Sons.

Asness, C. S. (1997). The interaction of value and momentum strategies, *Financial Analysts Journal* **53**(2): 29–36.

Barbee Jr, W. C., Mukherji, S. and Raines, G. A. (1996). Do sales-price and debt-equity explain stock returns better than book-market and firm size?, *Financial Analysts Journal* **52**(2): 56–60.

Bennett, K. P. (1994). Global tree optimization: A non-greedy decision tree algorithm, *Computing Science and Statistics* pp. 156–156.

Bernanke, B. (2011). Lessons from emerging market economies on the sources of sustained growth.
URL: *http://www.federalreserve.gov/newsevents/speech/bernanke20110928a.pdf*

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. (1985). Does the stock market overreact?, *The Journal of finance* **40**(3): 793–805.

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.

Bradford, J. P., Kunz, C., Kohavi, R., Brunk, C. and Brodley, C. E. (1998). Pruning decision trees with misclassification costs, *Machine Learning: ECML-98*, Springer Science, pp. 131–136.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**(2): 123–140.

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

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and regression trees*, CRC press.

Brodersen, K. H., Ong, C. S., Stephan, K. E. and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution, *Pattern Recognition (ICPR), 2010 20th International Conference on*, IEEE, pp. 3121–3124.

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

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms, *Proceedings of the 23rd international conference on Machine learning*, ACM, pp. 161–168.

Chan, E. (2013). *Algorithmic trading: winning strategies and their rationale*, Vol. 1 p. 22, John Wiley & Sons.

Chen, C., Liaw, A. and Breiman, L. (2004). Using random forest to learn imbalanced data, *University of California, Berkeley* .

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

Chen, M.-Y. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression, *Expert Systems with Applications* **38**(9): 11261–11272.

Cremers, K. M. and Petajisto, A. (2009). How active is your fund manager? a new measure that predicts performance, *Review of Financial Studies* **22**(9): 3329–3365.

Cutler, A., Cutler, D. R. and Stevens, J. R. (2012). Random forests, *Ensemble Machine Learning*, Springer Science, pp. 157–175.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves, *Proceedings of the 23rd international conference on Machine learning*, ACM, pp. 233–240.

Davis, J. H., Aliaga-Díaz, R. and Thomas, C. J. (2012). Forecasting stock returns: What signals matter, and what do they say now, *Valley Forge, Pa.: The Vanguard Group* .

Dhrymes, P. J. and Guerard Jr, J. B. (2012). Additional evidence of the risk and return of stocks.
URL: *http://www.columbia.edu/ pjd1/mypapers/mycurrentpapers/DhrymesGuerard3.pdf*

Dobson, A. J. and Barnett, A. (2008). *An introduction to Generalized Linear Models*, CRC Press.

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

Faraway, J. J. (2004). *Linear models with R*, CRC Press.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers, *Machine learning* **31**: 1–38.

Firer, C. and McLeod, H. (1999). Equities, bonds, cash and inflation: historical performance in South Africa 1925-1998, *Investment Analysts Journal* (50).

Gashler, M., Giraud-Carrier, C. and Martinez, T. (2008). Decision tree ensemble: Small heterogeneous is better than large homogeneous, *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*, IEEE, pp. 900–905.

Gilbert, E. and Strugnell, D. (2010). Does survivorship bias really matter? an empirical investigation into its effects on the mean reversion of share returns on the JSE (1984-2007), *Investment Analysts Journal* (72): 31–42.

Giuricich, M. N. (2013). *The benefits of a tree-based model for stock selection in a South African context*, Master's thesis, University of Cape Town.

Givoly, D. and Lakonishok, J. (1980). Financial analysts' forecasts of earnings: Their value to investors, *Journal of Banking & Finance* **4**(3): 221–233.

Gong, J. and Sun, S. (2009). A new approach of stock price prediction based on logistic regression model, *New Trends in Information and Service Science, 2009. NISS'09. International Conference on*, IEEE, pp. 1366–1371.

Goodwin, T. H. (1998). The information ratio, *Financial Analysts Journal* **54**(4): 34–43.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest, *The American Statistician* **63**(4).

Hair, J. F., Tatham, R. L., Anderson, R. E. and Black, W. (2006). *Multivariate data analysis*, Vol. 6, Pearson Prentice Hall Upper Saddle River, NJ.

Hargreaves, C. and Hao, Y. (2013). Prediction of stock performance using analytical techniques, *Journal of Emerging Technologies in Web Intelligence* **5**(2): 136–142.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J. and Tibshirani, R. (2009). *The elements of statistical learning*, Vol. 2, Springer Science.

Hawkins, D. M. (2004). The problem of overfitting, *School of Statistics, University of Minnesota, Minneapolis, Minnesota* **44**(4): 1–12.

Heiberger, R. M. and Holland, B. (2004). *Statistical analysis and data display: an intermediate course with examples in S-Plus, R, and SAS*, Springer Science & Business Media.

Hodnett, K., Hsieh, H.-H. and van Rensburg, P. (2012). Nonlinearities in stock return prediction: A blended approach, *Journal of Applied Business Research (JABR)* **29**(1): 7–22.

Hua, Z., Wang, Y., Xu, X., Zhang, B. and Liang, L. (2007). Predicting corporate financial distress based on integration of support vector machine and logistic regression, *Expert Systems with Applications* **33**(2): 434–440.

Huang, Q., Cai, Y. and Peng, J. (2007). Modeling the spatial pattern of farmland using GIS and multiple logistic regression: A Case Study of Maotiao River Basin, Guizhou Province, China, *Environmental Modeling & Assessment* **12**(1): 55–61.

Jacobs, B. I., Levy, K. N. and Starer, D. (1999). Long-short portfolio management: An integrated approach, *The Journal of Portfolio Management* **25**(2): 23–32.

James, G., Daniela, W., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning: With applications in R*, New York, Springer Science.

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.

Jegadeesh, N. and Titman, S. (2001). Profitability of momentum strategies: An evaluation of alternative explanations, *The Journal of Finance* **56**(2): 699–720.

Karami, G. R. and Talaeei, L. (2013). Predictability of stock returns using financial ratios in the companies listed in Tehran Stock Exchange, *International Research Journal of Applied and Basic Sciences* **5**(3): 360–372.

Khanna, T. and Palepu, K. G. (2010). Emerging giants, *Rivals from developing countries are invading your turf. How will you fight back?* p. 35.

Kruger, R. (2011). *Evidence of return predictability on the Johannesburg Stock Exchange*, PhD thesis, The University of Cape Town.

Kumar, M. and Thenmozhi, M. (2006). Forecasting stock index movement: A comparison of support vector machines and random forest, *Indian Institute of Capital Markets 9th Capital Markets Conference Paper*.

Lacerda, M. (2014). Analytics, University of Cape Town: Department of Statistical Sciences–University lecture.

Lahmiri, S. (2011). A comparison of PNN and SVM for stock market trend prediction using economic and technical information, *International Journal of Computer Applications* **29**(3): 24–30.

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

Lee, S., Ryu, J. and Kim, L. (2007). Landslide susceptibility analysis and its verification using likelihood ratio, logistic regression, and artificial neural network models: Case Study of Youngin, Korea, *Landslides* **4**(4): 327–338.

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

Lewis, R. J. (2000). An introduction to classification and regression tree (CART) analysis, *Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California*, pp. 1–14.

Leys, C., Ley, C., Klein, O., Bernard, P. and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, *Journal of Experimental Social Psychology* **49**(4): 764–766.

Liaw, A. and Wiener, M. (2002). Classification and regression by random forest, *R news* **2**(3): 18–22.

Lundstrom, I. (2013). *Finding risk factors for long-term sickness absence using classification trees*, Master's thesis, Royal Institute of Technology–Stockholm, Sweden.

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.

McConnell, D., Haslem, J. A. and Gibson, V. R. (1986). The president's letter to stockholders: A new look, *Financial Analysts Journal* **42**(5): 66–70.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, Vol. 2, Chapman and Hall London.

Murphy, K. P. (2007). Performance evaluation of binary classifiers, *Technical report*, Technical Report, University of British Columbia.

Nelder, J. and Wedderburn, R. W. M. (1972). Generalised linear models, *Journal of the Royal Statistical Society* **135**(3): 370–384.

Northern Trust Global Investments, N. (2012). Managing equity risk in volatile markets.
URL: *https://pointofview.northerntrust.com/Investment-Objectives/Managing-Equity-Risk-in-Volatile-Markets*

Obrien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors, *Quality & Quantity* **41**(5): 673–690.

Pan, Y. and Jackson, R. (2008). Ethnic difference in the relationship between acute inflammation and serum ferritin in US adult males, *Epidemiology and infection* **136**(03): 421–431.

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

Qi, Y. (2012). Random forest for bioinformatics, *Ensemble machine learning*, Springer Science, pp. 307–323.

Rogerson, P. (2001). *Statistical methods for geography*, Vol. 1, London: Sage Publications Ltd.

Schroders (2009). Tree building at schroders modern investment tools for stock selection.
URL: *http://www.schroders.com/staticfiles/schroders/sites/QEP/QEP-Tree-Building-at-Schroders-09122009.pdf*

Segal, M. R. (2004). Machine learning benchmarks and random forest regression, *Center for Bioinformatics & Molecular Biostatistics* .

Shan, Y., Taylor, S. and Walter, T. (2014). The role of "other information" in analysts' forecasts in understanding stock return volatility, *Review of Accounting Studies* **19**(4): 1346–1392.

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

Shtatland, E. S., Cain, E. and Barton, M. B. (2001). The perils of stepwise logistic regression and how to escape them using information criteria and the output delivery system, *26th Annual SAS Users Group International Conference, Long Beach, California*.

Singh, K. and Xie, M. (2008). Bootstrap: a statistical method, *Unpublished Working Paper. Rutgers University* .
URL: *http://www.stat.rutgers.edu/home/mxie/stat586/handout/Bootstrap1.pdf*

Sokolova, M., Japkowicz, N. and Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation, *AI 2006: Advances in Artificial Intelligence*, Springer Science, pp. 1015–1021.

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 Price, L. N. (1994). Performance measurement in a downside risk framework, *the Journal of Investing* **3**(3): 59–64.

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

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 44–47.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. and Zeileis, A. (2008). Conditional variable importance for random forests, *BMC bioinformatics* **9**(1): 307.

Strydom, B. and Charteris, A. (2009). A theoretical and empirical analysis of the suitability of South African risk-free rate proxies, *Conference Paper, Cape Town: African Finance Journal Annual Conference July*.

Su, J., Cooper, K., Robinson, T. and Jordan, B. (2011). Customer retention predictive modeling in HealthCare Insurance Industry, *BlueCross BlueShield of Florida, Jacksonville, FL*.

Sun, C. (2008). *Stock market returns predictability: Does volatility matter?*, Master's thesis, Columbia University.

Tang, Y., Zhang, Y.-Q., Chawla, N. V. and Krasser, S. (2009). SVMs modeling for highly imbalanced classification, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* **39**(1): 281–288.

Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications*, Master's thesis, Humboldt University, Berlin.

Torgo, L. F. R. A. (2000). Inductive learning of tree-based regression models, *AI Communications* **13**(2): 137–138.

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

Upadhyay, A., Bandyopadhyay, G. and Dutta, A. (2012). Forecasting stock performance in Indian market using multinomial logistic regression, *Journal of Business Studies Quarterly* **3**(3): 16–39.

van Gysen, M., Huang, C.-S. and Kruger, R. (2013). The performance of linear versus non-linear models in forecasting returns on the Johannesburg Stock Exchange, *International Business & Economics Research Journal (IBER)* **12**(8): 985–994.

Wei, Z. (2012). *A SVM approach in forecasting the moving direction of Chinese Stock Indices*, Master's thesis, Lehigh University.

Yu, X. (2008). *The Investigation of Style Indices and Active Portfolio Construction on the JSE*, PhD thesis, University of Cape Town.

Zhu, M. (2012). *Return predictability and its implications for portfolio selection*, PhD thesis, The University of Sydney.

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.

Zucchini, W. (2000). An introduction to model selection, *Journal of Mathematical Psychology* **44**(1): 41–61.

Appendix A

First Appendix

A.1 Portfolio performance metrics

Tab. A.1: Formulae used to calculate the annualised return, annualised excess return, annualised mean excess returns, Sharpe ratios, Information ratios, tracking errors, Sortino ratios and annualized return volatility for quarterly return data. Note that $n = 15$ for the period 2011:Q1 – 2014:Q3. **Notation:** $r_{Pt}, r_{Bt}, r_{Ft}, \sigma_p, \sigma_{p(downside)}$ denote the quarterly portfolio returns at time t, quarterly benchmark returns at time t, quarterly risk-free rate at time t, portfolio standard deviation and portfolio downside standard deviation respectively.

Performance metric	**Formula**
Annualised return	$\big[\prod_{t=1}^{n}(1+r_{Pt})\big]^{\frac{4}{n}} - 1$
Annualised excess return	$\big[\prod_{t=1}^{n}(1+r_{Pt})\big]^{\frac{4}{n}} - \big[\prod_{t=1}^{n}(1+r_{Bt})\big]^{\frac{4}{n}}$
Annualised mean excess return	$[\text{mean}(1+(r_{Pt}-r_{Bt}))]^4 - 1$
Sharpe ratio (Sharpe, 1975)	$\dfrac{\big[\prod_{t=1}^{n}(1+r_{Pt})\big]^{\frac{4}{n}} - \big[\prod_{t=1}^{n}(1+r_{Ft})\big]^{\frac{4}{n}}}{\sigma_p}$
Sortino ratio (Sortino and Van Der Meer, 1991)	$\dfrac{\big[\prod_{t=1}^{n}(1+r_{Pt})\big]^{\frac{4}{n}} - \big[\prod_{t=1}^{n}(1+r_{Ft})\big]^{\frac{4}{n}}}{\sigma_{p(downside)}}$
Tracking error	$\sqrt{4\text{Var}(r_{Pt}-r_{Bt})}$
Information ratio (Goodwin, 1998)	$\dfrac{\text{Excess return}}{\text{Tracking error}}$
Volatility	$\sqrt{4\text{Var}(r_{Pt}+1)}$

Appendix B

Second Appendix

B.1 Models pre–2008

Tab. B.1: Variables from the first-order logistic regression model estimated for the period 2000:Q1 – 2007:Q4. Significance levels: 0–0.001 '***'; 0.001–0.01 '**'; 0.01–0.05 '*'; 0.05–0.1 '·'.

Variable	β Coefficient
VAL	0.55***
DEBT.SERVICE	0.16*

Tab. B.2: Variables from the first-order linear regression model estimated for the period 2000:Q1 – 2007:Q4. Significance levels: 0–0.001 '***'; 0.001–0.01 '**'; 0.01–0.05 '*'; 0.05–0.1 '·'.

Variable	β Coefficient
VAL	0.05***
DEBT.SERVICE	0.02**
LEVERAGE	0.01·
MOM	0.01·

Tab. B.3: Variables from the second-order logistic regression model estimated for the period 2000:Q1 – 2007:Q4.. Significance levels: 0–0.001 '***'; 0.001–0.01 '**'; 0.01–0.05 '*'; 0.05–0.1 '·'.

Variable	β Coefficient
VAL	0.55***
DEBT.SERVICE	0.24**
DEBT.SERVICE^2	-0.07·

Tab. B.4: Variables from the second-order linear regression model estimated for the period 2000:Q1 – 2007:Q4. Significance levels: 0–0.001 '***'; 0.001–0.01 '**'; 0.01–0.05 '*'; 0.05–0.1 '·'.

Variable	β Coefficient
VAL	0.05***
LEVERAGE:DEBT.SERVICE	0.05***
LEVERAGE^2	0.02***
DEBT.SERVICE	0.04***
VAL:LEVERAGE	0.02*
VAL:DEBT.SERVICE	0.02*
LEVERAGE	0.01

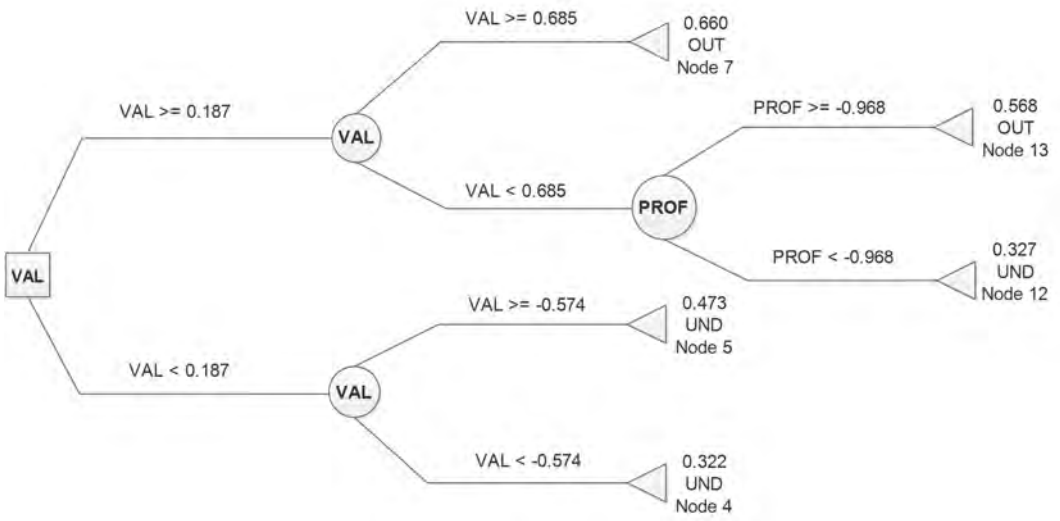

Fig. B.1: CART model for the universe of stocks for the period 2000:Q1 – 2007:Q4 (Pre–2008).

B.2 Models post–2008

Tab. B.5: Variables from the first-order logistic regression model estimated for the period 2008:Q1 – 2014:Q3. Significance levels: 0–0.001 '***'; 0.001–0.01 '**'; 0.01–0.05 '*'; 0.05–0.1 '·'.

Variable	β Coefficient
VAL	0.52***
FWD.GROWTH	0.12**
STAB	-0.16*
DEBT.SERVICE	0.13·
MOM	-0.09
EREV	0.06

Tab. B.6: Variables from the first-order linear regression model estimated for the period 2008:Q1 – 2014:Q3. Significance levels: 0–0.001 '***'; 0.001–0.01 '**'; 0.01–0.05 '*'; 0.05–0.1 '·'.

Variable	β Coefficient
VAL	0.04***
MOM	-0.01***
FWD.GROWTH	0.01 ***
STAB	-0.01·

Tab. B.7: Variables from the second-order logistic regression model estimated for the period 2008:Q1 – 2014:Q3. Significance levels: 0–0.001 '***'; 0.001–0.01 '**'; 0.01–0.05 '*'; 0.05–0.1 '·'.

Variable	β Coefficient
VAL	0.46***
FWD.GROWTH	0.20***
VAL:DEBT.SERVICE	-0.44**
VAL:FWD.GROWTH	0.17*
STAB	-0.14*
MOM	-0.09·
DEBT.SERVICE	0.04

Tab. B.8: Variables from the second-order linear regression model estimated for the period 2008:Q1 – 2014:Q3. Significance levels: 0–0.001 '***'; 0.001–0.01 '**'; 0.01–0.05 '*'; 0.05–0.1 '·'.

Variable	β Coefficient
VAL	0.03***
MOM	-0.01***
FWD.GROWTH	0.01***
LEVERAGE	0.01**
VAL:DEBT.SERVICE	-0.03**
VAL:FWD.GROWTH	0.01*
DEBT.SERVICE	0.01·
STAB	-0.01·
DEBT.SERVICE^2	-0.01·
MOM^2	-0.01·

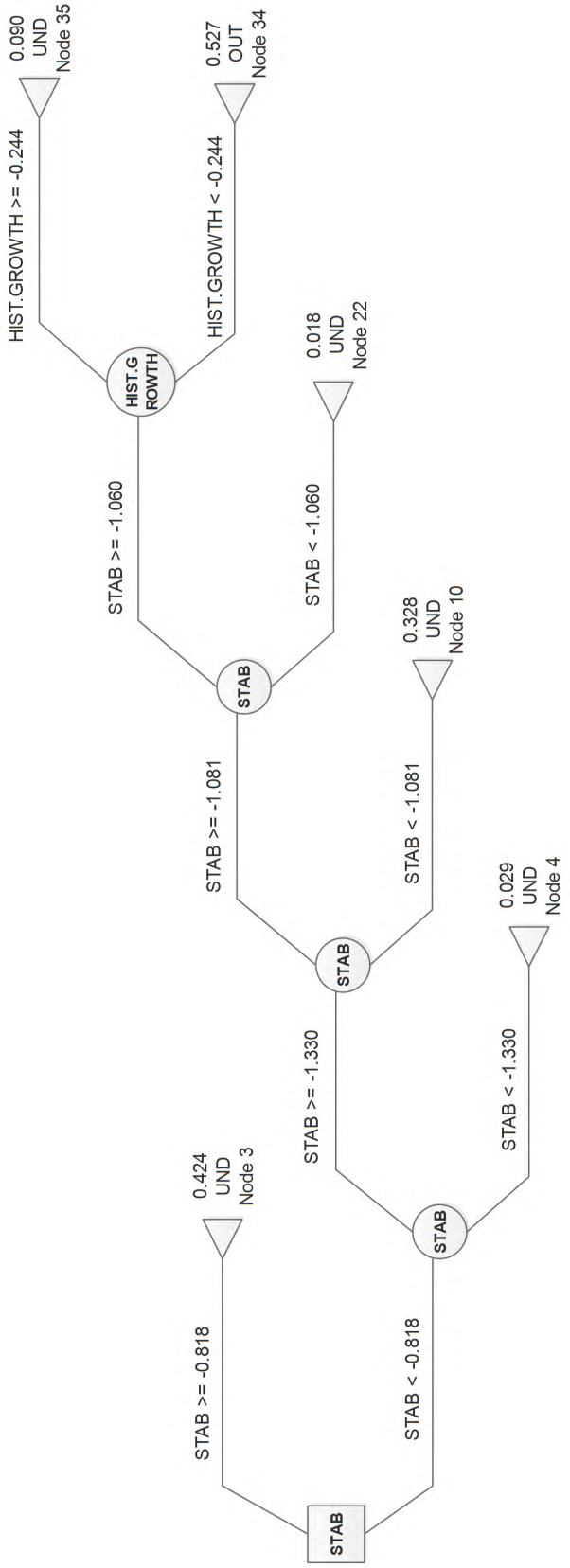

Fig. B.2: CART model for the universe of stocks for the period 2008:Q1 – 2014:Q3 (Post–2008).