CLASSIFICATION OF BORSA ISTANBUL FIRMS BASED ON MARKET PERFORMANCE DATA: COMPARISON OF CLASSIFICATION AND REGRESSION TREES

Şebnem Er*

University of Cape Town, Department of Statistical Sciences, Faculty of Science

Cape Town, South Africa

sebnem.er@uct.ac.za

Nuray Tezcan

Haliç University, Department of Business Informatics, Faculty of Management

Istanbul, Turkey

nuraytezcan@halic.edu.tr

Abstract

This study aims to classify Borsa Istanbul firms according to their return levels with one of the tree-based approaches known as Classification and Regression Trees (C&RT) using market performance data such as price to earnings ratio, market to book value ratio, risk measure of beta as well as firm level performance data such as debt ratio and profitability ratios. We aim to compare the results obtained from both Classification Trees and Regression Trees as well as providing a model that gives an optimal classification of the firms according to their return levels. After several runs of both of the algorithms, we find that the results obtained with Classification Trees are not consistent whereas the Regression Tree algorithm provides more consistent results. Moreover, both of the algorithms suggest different results. Classification trees suggest that profit per share and market to book value play a crucial role in classifying firms according to their returns whereas regression trees suggest that price to earnings and beta values play a crucial role.

Key Words: Classification and regression trees, Borsa Istanbul, stock returns

Topic Groups: Organizations and financing

INTRODUCTION

This study aims to classify firms according to their return levels using market performance data such as price to earnings ratio, market to book value ratio, risk measure of beta as well as firm level performance data such as debt ratio and profitability ratios using one of the tree-based approaches known as Classification and Regression Trees (C&RT). Besides providing a model that gives an optimal classification of the firms according to their return levels we aim to compare the results obtained from both Classification Trees and Regression Trees.

C&RT is a form of a nonparametric decision tree that can be used for either classification or regression estimation. Classification trees are mainly used when the predicted outcome variable is categorical and regression trees are used when it is numerical. Our predicted outcome is the return levels of Borsa Istanbul firms. It is common to evaluate stock returns using company financial ratios and market performance rates. The market performance rates used in evaluating company stock returns are risk, price to earnings, market to book value, and earnings per share ratios. We used both the numerical stock return variable and the binary coded return variable in order to compare the two C&RT algorithms.

The paper is organised as follows: The second section reviews the literature on market performance measures and their relationships with firm return levels. The third section gives a brief understanding of the C&RT algorithm. The fourth section discusses the data and the results. Finally the last section summarises the findings.

LITERATURE REVIEW ON MARKET PERFORMANCE RATES AND STOCK RETURNS

In an unpredictable environment of today, stock exchange investors are trying to find which of the stocks are generating more return than others. When investors are investing their money on common stocks, they monitor various classical financial ratios as well as market performance rates. The most important market performance rates are risk, price to earnings, market to book value, and earnings per share ratios. Apart from the market performance ratios, financial ratios such as debt and profitability ratios are also monitored.

Price to earnings ratio, which is measured by the ratio of market value per share to earnings per share, is an indicator of earnings growth of a company. If this value is high for a company, then investors expect higher earnings growth. Similarly, market to book value ratio which is a measure of how much investors are willing to pay in response of a book value of a share. If this ratio is higher than one, then that means the market value of the company is higher than the book value indicating that the company's stocks are overvalued. One of the important measures is earnings per share which measures a company's market value. The investors can have a feeling of how much of the total profit of a company is distributed to each of its shares.

There is a substantial literature that examines the associations between financial ratios and stock returns. Earlier in 1973 with a seminal work in this area, Fama and MacBeth find that there is a positive relationship between average stock returns and beta. This indicates that the higher the risky a common stock is the more return we expect to gain (Fama and MacBeth, 1973).

Regarding the book to market value ratio, there is again a wide range of studies. A research on US markets by Fama and French find that firm size, book to market equity ratio capture much of the cross-sectional variation in average stock returns (Fama and French, 1992). Another research from US markets by Rosenberg et al. find that average stock returns are positively related to book to market value ratio (Rosenberg, et al., 1985). Chan et al. examined Japanese markets and used four variables such as earnings yield, size, book to market ratio, and cash flow yield in order to explain stock returns. Their findings reveal that book to market value ratio and cash flow yield have the most significant positive impact on expected returns (Chan, et al., 1991). These findings indicate that the less overvalued (in other words the more undervalued) a common stock is the more return investors expect. This is supported in many other research papers from Turkey as well (Canbaş, et al., 2007; Yıldırım, 1997).

The effect of price to earnings (PE) ratio that is an indicator of earnings growth of a company is examined in several markets. Basu finds that stocks with high PE ratios generate lower stock returns (Basu, 1977). Aydoğan and Güney (1997) and Ege and Bayrakdaroğlu (2007) also investigate the effect of PE ratios in the Turkish market. They find evidence of a negative relationship between stock returns and PE ratios (Aydoğan and Güney, 1997; Ege and Bayrakdaroğlu, 2007).

As it can be summarised from the literature, we know that returns are associated with the risk measure beta, market-to-book value ratio, earnings to price ratio, size and other firm specific financial measures. In this paper our main aim is to find the explanatory variables of returns for the Turkish companies using nonparametric tree-based approaches.

METHODOLOGY: CLASSIFICATION AND REGRESSION TREES

Classification is a general term that can be used for statistical methods that aim to classify the sampling units given a set of measurements on those units. These methods mainly differ from each other based on the fact that the classes are predefined or not. If the classes are not predefined indicating that the class a sampling unit belongs to is unknown, then cluster analysis will be an appropriate option. On the other hand, if the classes are known in advance, then one can prefer to use discriminant analysis or a nonparametric alternative of tree-based approaches. Here in this paper, we use one of the tree-based approaches known as Classification and Regression Trees (C&RT). The main advantages of tree-based approaches over other methods can be summarised as follows:

- Tree-based approaches are nonparametric approaches that do not require data distribution specification like normality of the explanatory variables (Chang and Wang, 2006, p.1019).
- Tree-based approaches can be used for both classification and regression.
- Scaling of the variables is not necessary since the decision on the selection of variables is made sequentially.
- Variables that are numerical and categorical can be analysed together.
- Missing value handling of tree-based approaches is simple. Missing values are handled by substituting the value of a variable with similar splitting characteristics as the variable with the missing value (Nisbet, et al., 2009, p.243).
- Tree-based approaches offer a visual representation of the classification structure.
- The final results of tree-based approaches are summarized in a logical if-then format.

Considering all the advantages, we adopt classification and regression trees (C&RT) to classify Borsa Istanbul firms according to their return levels based on market performance data in order to produce an accurate classifier and to understand what variables or interactions of variables drive to that classification (Breiman et al., 1984).

There are several other tree-based classification algorithms which can be used as well rather than C&RT. Some of these are Chi-squared automatic interaction detection (CHAID), random forests and boosted trees, artificial neural networks and support vector machines. Here we focus on C&RT method and compare the results of classification and regression trees. The main difference of a C&RT algorithm is that it is a binary splitting algorithm. By "binary" we mean that the algorithm splits the tree into only two branches whereas for example with CHAID algorithm the tree can be split into more than two branches. Details on the algorithms that are not considered in this paper can be found in Nisbet, et al. (2009).

C&RT algorithm is a form of a decision tree that can be used for either classification or regression estimation (Nisbet, et al., 2009, p.242). Classification trees are mainly used when the predicted outcome variable is categorical and regression trees are used when it is numerical. Our predicted outcome is the return levels of Borsa Istanbul firms. Firstly, we classified the firms as those with negative returns and as those with positive returns. Since in this case the predicted outcome variable is categorical having two categories (positive-negative return), we applied classification trees to investigate which of the market performance rates play a crucial role in this distinction. Secondly, we used the numerically measured return levels of the firms and applied regression trees since in this case the predicted outcome is a numerical variable.

C&RT algorithm starts with gathering all the sampling units together in what is called as the root node. This root node is then partitioned into two branches according to the measure of diversity calculated using each of the explanatory variables (Crawford, 1989, p.199). The variable that provides the greatest reduction in the diversity is selected to construct the partitioning. After the selection of the variable at the first step, the process is repeated to construct more nodes for further branches of the tree. At each partition, the node that is split is called the parent node, and the nodes that result from partitioning are called the child nodes (Nisbet, et al., 2009, p.241; Friedl and Brodley, 1997, p.401). The algorithm continues to partition the tree until a stopping rule is met. These rules are categorized into two groups. The first group of rules are Bonsai techniques where the tree is grown according to one of the stopping rules. The algorithm stops either when a predefined minimum node size is reached, or when a predefined minimum reduction in diversity level is reached, or when all the sampling units are classified into one category.

The second type of stopping rules is called Pruning techniques (Crawford, 1989, p.201). These techniques do not apply any stopping rules at the beginning and the tree is grown to its full size. After that some of the branches of the tree are pruned checking if the branch resulted because of overfitting and that it cannot be generalised to other sample data sets. Pruning can be applied according to either using a hold-out sample or cross-validation (Wehrens, 2011, p.132-135). Once the hold-out sample rule is used, the sample data set is divided into two sets, training and hold-out sample. The classification method is trained on the training sample and tested on the hold-out sample. The pruning is applied according to the validity of the correct classification rate obtained from each of the sample data sets. With cross-validation cost pruning, the sample data set is divided into randomly chosen sub samples. At each step, one of the subsamples are excluded from the entire sample data set and this excluded subsample is regarded as an independent test dataset and the rest are combined to be the learning set. This technique is the most commonly used stopping rule for C&RT algorithm that helps avoiding the overfitting problem which creates results that cannot be generalized to other random samples.

This process is called recursive partitioning since the algorithm starts with considering all of the objects in one group and then partitioning the sampling units into more homogeneous sub-groups until the algorithm cannot find any further improvement in the partitioning process regarding the variables in consider. The child nodes that cannot be partitioned any further are called the terminal nodes which are interpreted at the end of the analysis creating classification rules (Mahjoobi and Etemad-Shahidi, 2008, p.173).

The results obtained with the C&RT algorithm are evaluated using the correct and misclassification rates for the overall tree as well as for each of the categories in the predicted outcome variable. In order to apply C&RT algorithm we will use **rpart** R package (Therneau et al. 2013).

DATA AND RESULTS

The data set includes 306 firms that are quoted in Borsa Istanbul in 2012. Data has been collected from Finnet commercial website that collects and arranges firm level data (Finnet URL). We have excluded the outliers from our dataset. The distribution of the firms according to their operating markets and their sectors can be found in Table 1 and Table 2, respectively. It is clearly evident from both of the tables that most of the companies are in the national market (71%) and most of them are operating in manufacturing and financial sectors (83%).

Table 1: Distribution of the firms according to the markets

Market	Number of firms	Percentages
Watchlist companies market	18	5.9%
Regional market	35	11.4%
Collective products market	36	11.8%
National market	217	70.9%

Table 2: Distribution of the firms according to the sectors

Name of the Sector	Number of Firms	Percentages
Mining	2	0.7%
Construction and Public Works	3	1.0%
Electricity, Gas and Water	4	1.3%
Education, Health, Sports And Other Social Services	4	1.3%
Transportation, Telecommunication And Storage	6	2.0%
Technology	11	3.6%
Wholesale And Retail Trade, Hotels And Restaurants	20	6.5%
Financial Institutions	99	32.4%
Manufacturing Industry	157	51.3%

In order to apply classification tree algorithm, the return response variable is recoded into a binary variable consisting of negative and positive return companies. The explanatory variables are also recoded into binary format as follows:

Table 3: Recoding of the variables used in the analysis

Variable	Categories		N
Return	>=0	(positive return)	216
	<0	(negative return)	90
Beta	>1	(risky)	15
	<=1	(less risky)	291
Market to book value	>2	(overvalued)	100
	<=2	(undervalued)	206
Price to earnings ratio	>25	(high growth)	138
	<=25	(low growth)	168
Earnings per share	>0	(high profit)	232
	<=0	(low profit)	74
Debt ratio	>1	(high debt)	136
	<=1	(low debt)	170
Profitability	>=0	(positive profit)	232
	<0	(negative profit)	74

The analysis is run using both the numerical values and the recoded categorical values of the explanatory variables and the response variable. When the response variable is numerical the algorithm estimates regression trees and when it is a categorical variable, it estimates classification trees. Firstly, we will review the results obtained with classification trees when the response variable is categorical. Secondly we will review the results obtained with regression trees. All of the analysis is applied using **rpart** R package (Therneau et al. 2013).

CLASSIFICATION TREE RESULTS

The following classification tree is obtained when all the variables are encoded into categorical variables as defined in Table 3. The result of the tree was tested against overfitting problems using the cross validation cost rule with a constraint of minimum split of ten observations. In around 20% of the simulations the classification algorithm provided nine splits whereas around 80% of the simulations suggested zero split. These results indicate that the classification tree result is not very reliable since the number of splits given with the following tree diagram only appears in 20% of the cases. Moreover, when we look at the frequency of the selected variables as the root node (starting point for the algorithm), we see that in most of the cases earnings per share and price to earnings variables are chosen as root nodes. This also indicates an unstable solution that must be interpreted with suspicion.

Figure 1: Classification Tree Result

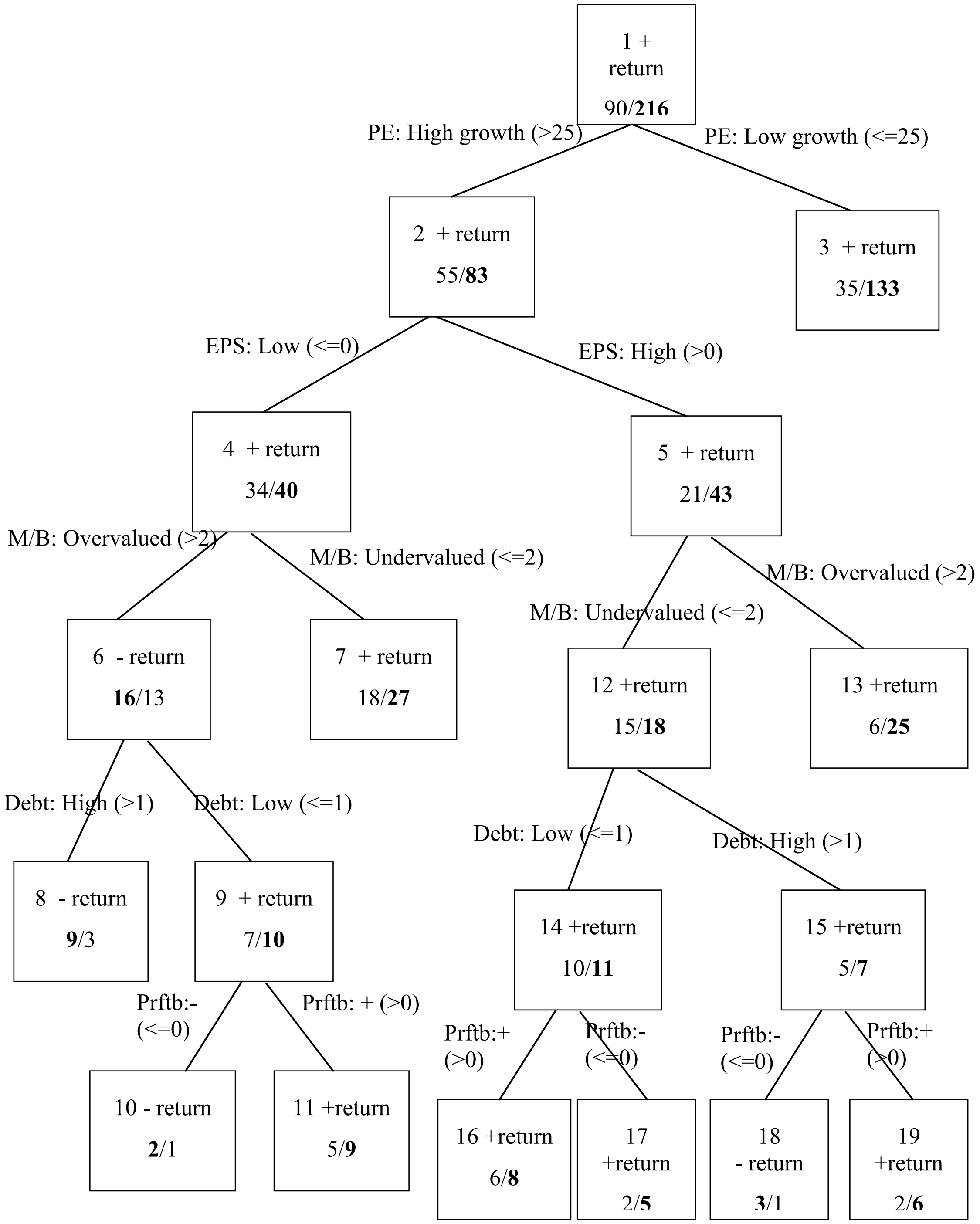


Table 4: Correct Classification Table with Classification Trees

		Observed Categories		
		+ Return	- Return	Total
Predicted Categories	+Return	**213**	74	287
	- Return	5	**14**	19
	Total	218	88	306

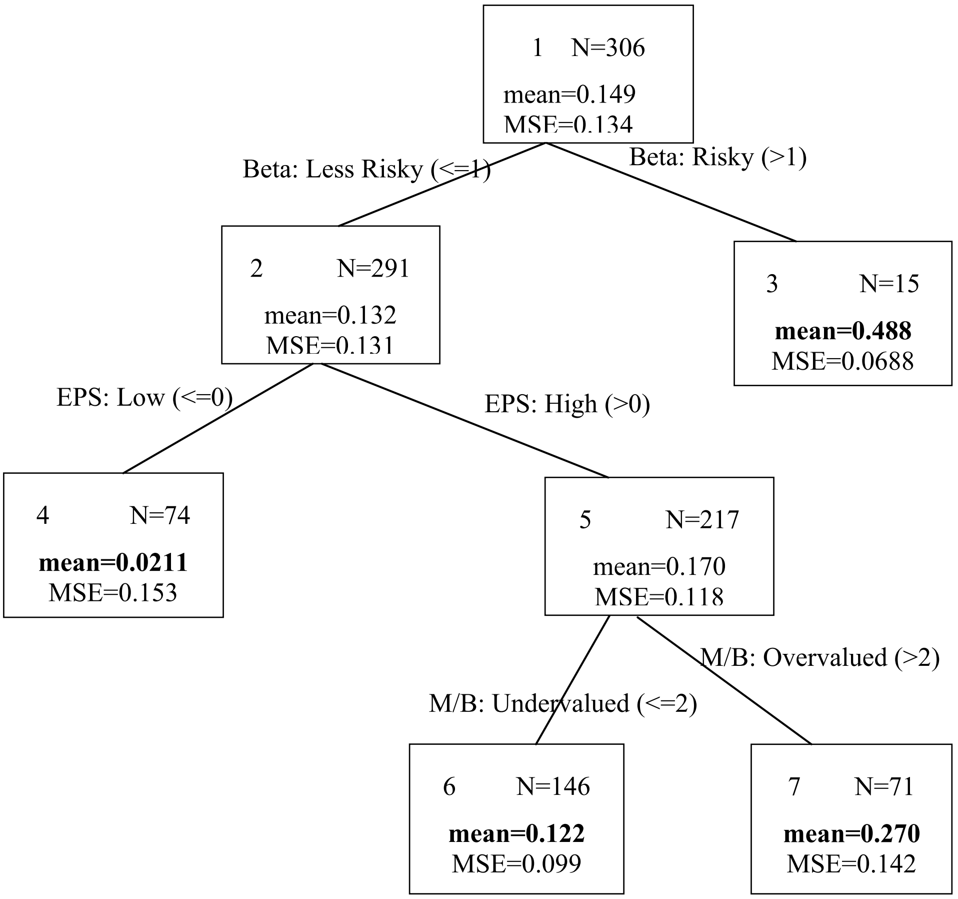
Keeping in mind that the cross validation results for the number of splits were very inconsistent, we will look at the correct classification and misclassification rates. The correct classification table shows that overall hit rate with this classification tree is 74.18% (=(213+14)/306) and it is clear that this rate is not very high. Moreover, the misclassification rate for the negative return category is very high with a rate of 84% (=74/88). The terminal nodes having very few observations, that the misclassification rate being very high for one of the categories of the response variable and inconsistent results of cross validation reveal that this output obtained with classification trees is not reliable. On the other hand, when the explanatory variables are used as numerical values, the tree is very much overfitted and there is no pruning alternative. Therefore we do not present the results here.

REGRESSION TREE RESULTS

The following regression tree is obtained when all the explanatory variables are categorical but the response variable return is numerical. The result of the tree was tested against overfitting problems using the cross validation cost rule. In around 60% of the simulations the regression tree algorithm provided 3 splits and around 40% of the simulations suggested five splits. Since the number of splits suggested by the initially generated algorithm is five, we provided these results. When we look at the frequency of the selected variables as the root node (starting point for the algorithm), we see the same picture with the classification trees with only an exception of the inclusion of the beta risk variable. This again indicates us an unstable solution that we will carefully interpret. However, here we have a more consistent output compared to classification trees.

Figure 2 provides the regression tree result. The very striking decision rule that we can extract from the tree is that if a stock’s beta value is greater than one indicating that it is a more risky stock (Node 3 in Figure 2), then the average return level in this case is 0.488. This average return value is significantly different than all the other averages provided with the rest of the terminal nodes. Furthermore, this terminal node has a very low mean square error (variance) value which means that the variation (with a coefficient of variation value of 54%) in this terminal node is very low. These reveal that this terminal node is a reliable source of decision. However, the other terminal nodes (nodes 4, 6 and 7) provide results with high variation especially the fourth node. The fourth node where the earnings per share is negative and the beta risk is less than one, we estimate that the average return levels will be very low with most of the negative returns. Even though the variation in this node is very high, we can conclude that this decision rule is consistent with the literature review. The sixth and seventh nodes have higher average returns compared to those companies that are less risky and with low EPS values. If a company is less risky but has a high EPS value, then the average return for those undervalued companies will be lower than the overvalued ones. This is not surprising to see that overvalued companies have a slightly larger return values than those that are undervalued.

Figure 2: Regression Tree Result



When the explanatory variables are used as numerical values, the tree is very much overfitted and when we check for cross validation in order to prune the tree, the results offer 1 split for nearly 70% of the simulations. As a result it is not very practical to generate a regression tree and interpret the results once all the variables are numerical. Therefore, we are not presenting these results here.

CONCLUSION

In this study, we aimed to classify Borsa Istanbul firms according to their return levels using market performance data with Classification and Regression Trees (C&RT). We also aimed to understand what variables or interactions of variables drive to the classification of 306 firms that are quoted in Borsa Istanbul in 2012. In order to apply classification tree algorithm we recoded the return response variable into a binary variable consisting of negative and positive return companies. The explanatory variables are also recoded into binary form. Using both the recoded and the numerical values of the response variable (return), we applied both of the classification trees and the regression trees with the aim of comparing the results obtained from both of the algorithms.

Both of the trees obtained from each of the algorithms were tested against overfitting problems using the cross validation cost rule. In around 20% of the simulations the classification algorithm provided nine splits whereas around 80% of the simulations suggested zero split. This indicated us that the results obtained with Classification Trees are not consistent. Moreover, the misclassification rate with Classification Trees for the negative return category is very high with a rate of 84% (=74/88).

On the other hand, Regression Tree algorithm provides more consistent results. Regression trees suggest that price to earnings and beta values play a crucial role. The most important decision rule that can be extracted from the regression tree is that if a stock's beta value is greater than one indicating that it is a more risky stock, then the average return level in this case is significantly different than all the other averages provided with the rest of the terminal nodes. It is also found that those companies that are less risky with high EPS values have higher average returns compared to companies with low EPS values. Moreover, if a company is less risky but has a high EPS value, then the average return for those undervalued companies will be lower than the overvalued ones. However, these findings should be interpreted with caution since terminal nodes provided with regression trees have high variation.

In both of the algorithms we used recoded explanatory variables. Once the numerical values of the explanatory variables are used, both of the trees are overfitted too much and when we check for cross validation in order to prune the tree, the results are not satisfying.

In summary, classification of Borsa Istanbul companies with classification trees did not provide satisfying results, however the regression trees were better in terms of pruning and overfitting problems. The analysis can be extended using the more advanced classification algorithms such as random forests, neural networks and support vector machines to find a more reliable classification method.

REFERENCES

Aydoğan, K., & A. Güney. (1997). P/E and Dividend Yield in Estimation of Stock Prices (In Turkish). *ISE Journal.* 1 (1), 83-96.

Basu, S. (1977). Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis. *The Journal of Finance.* 32 (3), 663-682.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. Wadworths Inc. USA.

Canbaş, S., S. Kandır, & A. Erişmiş. (2007). Test of the Some of the Firm Properties Affecting Stock Efficiency in ISE (In Turkish). *Finance Politics & Economic Comments.* 44 (512), 15-27.

Chan, L., Y. Hamao, & J. Lakonishok. (1991). Fundamentals and Stock Returns in Japan. *The Journal of Finance.* 46 (5), 1739-1764.

Chang, L. & Wang, H. (2006). Analysis of Traffic Injury Severity: An Application of Non-Parametric Classification Tree Techniques. *Accident Analysis and Prevention*. 38, 1019-1027.

Crawford, S.L. (1989). Extensions to the CART Algorithm. *International Journal of Man-Machine Studies.* 31,197-217.

Ege, İ., & A. Bayrakdaroğlu. (2007). Analysis of ISE Firm Stock Return Performance in Globalisation with Logistic Regression (In Turkish). *8th Turkey Econometrics and Statistics Conference*, 1-6.

Fama, E., & J. MacBeth. (1973). Risk, Return and Equilibrium: Empirical. *Journal of Political Economy.* 81, 607-636.

Fama, E., & K. French. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*. 4 (2), 427-465.

Finnet URL: http://www.finnet.com.tr/f2000/genel/Index.aspx

Friedl, M.A. & Brodley, C.E. (1997). Decision Tree Classification of Land Cover from Remotely Sensed Data. *Remote Sensing of Environment*. 61 (3), 399-409.

Mahjoobi, J. & Etemad-Shahidi, A. (2008). An Alternative Approach for the Prediction of Significant Wave Heights Based on Classification and Regression Trees. *Applied Ocean Research*. 30, 172-177.

Nisbet, R., Elder, J. & Miner, G. (2009). Handbook of Statistical Analysis and Data Mining. Elsevier.

Rosenberg, B., K. Reid, & R. Lanstein. (1985). Persuasive Evidence of Market Inefficiency. *Journal of Portfolio Management.* 11, 9-17.

Therneau, T., Atkinson, B., & Ripley, B. (2013). Recursive Partitioning and Regression Trees. (rpart R package), http://cran.r-project.org/web/packages/rpart/index.html.

Wehrens, R. (2011). Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences, Use R!. Springer-Verlag Berlin Heidelberg.