UNIVERSITY OF AGDER

Evaluating CEO Dismissal by the use of CART

Sven Høidal

Supervisor
Jochen Jungeilges

This Master's Thesis is carried out as a part of the education at the University of Agder and is therefore approved as a part of this education. However, this does not imply that the University answers for the methods that are used or the conclusions that are drawn.

University of Agder, 2012

Faculty of Economics and Social Sciences

Department of Economics and Business Administration

Table of Contents

Introduction

When evaluating factors that play a role in CEO dismissal, the statistical tool Classification and Regression Trees (CART) by Breiman, Friedman, Ohlsen and Stone (1984) will be used. It is a non-parametric method that uses binary splits in order to create a tree that classifies observed cases based on their information. The tree gives a clear view of how the different variables interact with each other in the classification process. Assumptions when using CART are much more relaxed than for traditional regression analysis, as argued by Morgan and Sonquist (1963) and Armstrong and Scott (1970). Assumptions such as homogeneity, non-linearity and normal distribution data are nonexistent.

The dataset used for the analysis contains information on the 250 largest firms in Europe in the time period 1998-2004. A total of 334 observations of dismissal cases were recorded with 27 variables. The dataset was used in studies of CEO dismissal both with and without dismissals due to mergers included.

This thesis has been exploratory in two ways. First I had to learn about CART which has not been a part of my master program or any other previous studies. To the best of my knowledge there are no studies on CEO dismissal that have used CART. Big parts of the thesis have then been to learn about CART and then apply it to see if CART can help give clarity to the subject of CEO dismissal.

CEO dismissals have been studied widely over the years, and I have used some of these studies in order to get an expectation of which variables that could be significant for CEO dismissals.

Company performance is a central aspect in many studies and Leker and Salomo (2000) found that poor performance increases likelihood for CEO dismissal. This was also supported by Malatesta and Parrino (2004) who in addition also found that the composition of the board proved significant. Jungeilges, Oxelheim and Randøy (2010) found that having an American citizen on the board increased the likelihood for CEO dismissal.

Allgood and Farrell (2000) found that CEOs who have been able to entrench themselves from the board are more guarded against dismissal due to poor performance. The entrenchment would according to them fade over time, making the likelihood of CEO dismissal higher for higher tenures. This is contrary to what Fredrickson, Hambrick and Baumrin (1988) found. Their results points to as tenure increases the likelihood of CEO dismissal would decrease.

Size of firm has also been proved to be significant. Pfeiffer and Moore (1980) suggest that bigger firms are more likely to dismiss their CEOs. Furthermore, Fredrickson, Hambrick and Baumrin (1988) found that the size of the industry also increases the likelihood for CEO dismissal. In both those research efforts it is argued that bigger size may increase the pool of suitable replacements.

In a study by Guan, Wright and Leikam (2005), income of the firm has proved to change in the year prior to a forced dismissal. The change tended to be positive, and it is argued that this is due to the CEO conducting earnings management in order to keep his position.

The effect on age of a CEO has to, my knowledge, not been studied that much. Some studies such as Shen and Cannella Jr. (2002) uses age around retirement age to discriminate between forced and voluntary dismissals. Jungeilges, Oxelheim and Randøy (2010) however found that age proved to be significant in the way that higher age increased the likelihood for CEO dismissal.

The results gotten in this analysis support some and deviate from other findings. When merger cases were included, CEOs of 58.5 years or older were directly classified as forced. Younger CEOs that achieved an average or poor performance in change in total stock return (TSR) right before a succession event was classified as voluntary dismissals. Younger CEOs that performed well would either be classified in the dismissals due to merger or the voluntary category depending on the country in which the firm was located.

When merger cases were excluded, the best sized tree would only use one split to decide between forced and voluntary dismissal cases. CEOs 58.5 years old or older were classified as forced dismissals while younger CEOs tended to leave their position voluntarily. To get a deeper understanding of how different variables play a role in CEO dismissals, the second best sized tree was analyzed. This was a much larger tree and variables that were significant were age, nationality of the firm, performance in the change in TSR right before a dismissal and change in industry adjusted TSR. Different countries would react differently to poor performance. In the tree poor performance would not always lead to dismissal and neither would good performance always lead to a voluntary leave of the CEO.

The content of the thesis starts with a section on previous studies on CEO dismissal before the dataset will be described. Then an introduction to CART will be given in the methodology part before an analysis on data and results will be conducted. Lastly is the discussion and conclusion.

Previous Studies on CEO Dismissal

It is natural to think that the performance of a firm plays a role in CEO dismissal cases. This is supported by Leker and Salomo (2000) who found that a decrease in firm performance, measured by RoA, increased the probability of a dismissal within the following year. A similar relationship between firm performance and likelihood of CEO dismissal is was also found by Huson, Parrino and Starks (2001). A study by Huson, Malatesta and Parrino (2004) also supports the negative relationship between firm performance and likelihood for CEO dismissal. Another finding in their studies was that the composition of the board also played a role.

Oxelheim and Randøy (2005) found that there is a connection between globalization of sales, financial markets and corporate control and increased pay of the CEO. With this connection established they argue that this is a premium the CEO receives for increased risk of dismissal. This is further backed up by Jungeilges, Oxelheim and Randøy (2010) where they found that having an American citizen as a board member increases the risk for CEO dismissal for European firms.

Jensen and Meckling (1976) applies agency theory to analyze the situation between the owner(s) (principals) and the CEO (agent), and how they both want to maximize their own utility. This may raise a conflict between their interests and goals, and Fama (1980) and Fama and Jensen (1983) suggest that the conflict should be covered through CEO incentives such as pay. Jensen and Murphy (1990) suggest that firms may benefit from a more incentive based pay (i.e. bonuses and stock options) for the CEO. Oxelheim and Randøy (2005) argue that the agency theory supports an underlying rational for increased CEO pay as compensation for increased dismissal risk. They also state that the same compensation suggested in agency theory is not as widespread in Scandinavia as in the US and that the influence from an American board member may reflect this.

The power play between CEOs and the board has been documented by Boyd (1994) and Zajac and Westphal (1996). They saw that CEOs have been able to interlock the board and enjoyed a high amount of power. Denis, Denis and Sarin (1997) have in their studies on ownership structure and CEO turnover found that entrenched executives are, to an extent, guarded against dismissal due to poor firm performance. Allgood and Farrell (2000) found that CEOs with high entrenchment are less likely to be dismissed due to poor performance. They found that in general the entrenchment fades over time, but for CEOs recruited from outside the firm it took some time to build up power. This entrenchment also faded off during later stages of the tenure.

The CEOs tenure is also discussed by Fredrickson, Hambrick and Baumrin (1988), they are of an opinion that as time passes the board’s allegiance moves closer to the current CEO, suggesting that as tenure increases the likelihood of dismissal decreases. A similar view is shared by Zhang (2008) who found that in cases where the predecessor had been dismissed there was an increased likelihood of early dismissal of the new CEO. This was especially in cases where the new CEO was recruited from outside the firm. Zhang meant that this was due to the board would not have sufficient time to evaluate all information about all candidates for the position sufficiently.

The size of the firm has also proved to play a role in the dismissal of top leader positions. Grusky (1961) and Pfeiffer and Moore (1980) found that bigger firms has a higher likelihood for forced dismissals. Although the dataset used in this thesis consists of the 250 largest firms in Europe, there will still be differences in sizes and market capitalization may still play a significant role in classifying succession events. Pfeiffer and Moore (1980) argued that one of the reasons for the significance of size was that they had a larger number of inside candidates to replace the current leader.

A similar approach can be used on industry size. Fredrickson, Hambrick and Baumrin (1988) argue that the number of firms in an industry determines the number of suitable CEO contenders. Firms in large industries may be more likely to dismiss a CEO as there are a greater number of replacements available.

The prior performance of the company is argued by Fredrickson, Hambrick and Baumrin (1988) to play a role in the likelihood of a dismissal as it may set standards that the new CEO is expected to achieve. They argue that this may affect the likelihood of dismissal in two ways: If the prior CEO produced better than average results, then a successor who performs below average may be dismissed as he is not living up to the expectations capital owners will hold. Another way is if the firm has had a period of very low results they may be eager to get their results boosted, so poor performance in this case may also trigger a dismissal.

An important topic in a study by Zhang (2008) has been the origin of the CEO. They divide between insider and outsider CEOs. Insiders are recruited from within the firm while outsiders are brought in externally. In general an outsider CEO will have to prove himself more in the beginning while an insider CEO will already be somewhat familiar with the board and face a smaller risk of dismissal. Fredrickson, Hambrick and Baumrin (1988) also argue that certain external 'star' CEOs will be paid more than others as they are expected to be better than the other candidates. This stresses the importance of the 'star' to deliver results as that is why the firm chose to hire him and pay the extra salary. Karaevli (2007) on the other hand found that outsiders proved to increase performance for previously poor performing firms. Karaevli points out that firms could receive a boost should the outside CEO have fresh knowledge and skills. Data on the previous CEO and origin of the new CEO is not included in the dataset evaluated in this thesis.

In a study by Guan, Wright and Leikam (2005) income has proven to change prior to a succession event. By looking at forced dismissal events for the CEO they found that in the year prior to a dismissal, the CEO had conducted earnings management. Their results supported their hypothesis that the CEO had tried to increase the earnings in order to delay or avoid dismissal. Pourciau (1992) did a similar study, but did not experience the same results. Guan, Wright and Leikam (2005) argue that this is due to the fact that Pourciau did not discriminate between forced and voluntary successions, but focused on all non-routine successions. Brickley, Linck and Coles (1999) and Reitenga and Tearney (2003) found that this was also the case when the CEO left due to mandatory retirement and want to retain as a member of the board.

There have not, to my knowledge, been many studies that have looked on how the age of the CEO might play a role in CEO dismissal. In some studies age been used to decide a case to be a forced dismissal or a voluntary leave of the CEO, Shen and Cannella Jr. (2002). The age has then typically been set at around retirement age. One study, however, found that age played a significant role in CEO dismissals. Jungeilges, Oxelheim and Randøy (2010) found that higher age increased the probability for a CEO to be dismissed.

To summarize and illustrate the assumed effects on the likelihood of CEO dismissal I have created the table below:

Table: Effects of independent variables on likelihood of CEO dismissal

Name of the Variable	Assumed effect on likelihood of CEO dismissal
CEO age	Higher likelihood for older CEOs, but closer to retirement age this is uncertain
CEO tenure	Higher likelihood for lower values, suspected low likelihood for middle values and uncertain about higher values
Market capitalization	Higher likelihood for bigger firms (higher values)
Total Stock Return (TSR)	Higher likelihood for bad performance (lower values)
Income	Higher likelihood for higher income (higher values), but higher values may also suggest planned succession
Market capitalization prior	Higher likelihood for higher values
TSR prior	Higher likelihood for lower values
US exchange listing	Higher likelihood if the firm is listed in the US
US board membership	Higher likelihood if present
Industry	Higher likelihood for bigger industries

Description of the Data

The Source of the Data

The dataset I will be using contains information on CEO succession events in the 250 largest European publicly traded firms. The time period the data was collected ranges from 1995 to 2004. It was in 2004 and by market capitalization the 250 largest firms were selected. The data concentrates around CEO successions, and information regarding the succession events was gathered through financial media such as Financial Times etc. Other data included is market value and changes in this prior to the succession event, the same with total stock return. Information regarding nationality of the firms, as well as the industry to which the firm belongs is also included. The dataset was collected by Booz Allen. Data on whether the firm had at least one American citizen on the board and the cross-listing of European firms on American markets was collected from sources such as Annual Reports, company web pages or through direct contact with the firm.

Since some of the information was collected from media sources, the firms had to be large in order to be covered sufficiently. Had smaller firms also been included the dataset would have been larger which would have been preferred for the CART procedure, but it is not a necessity.

According to Morgan and Sondquist (1963) and Armstrong and Andress (1970), assumptions for the CART procedure are not as strict as for the more traditional regression analysis. In fact, Breiman, Friedman, Ohlsen and Stone (1984) states that the only requirement concerning the data is that information regarding the dependent variable is available. Other than that the only thing is that a large dataset is preferred over a smaller dataset. This means that assumptions regarding variables such as normally-distributed data, homogeneity of variance and linear relationship between independent and dependent variables are non-existing. In addition, problems regarding missing values are solved in an elegant manner using the second or third etc. best split at a node should a variable be missing. It is of course not optimal, but may not necessarily be too far away either.

Descriptive Analysis

All variables described in this section are included in the dataset. Whether they show up in the resulting trees or not is decided entirely by the software analyzing the data.

The Dependent Variable

As mentioned in the section above this data was collected through the financial media. In some cases it might not have been 100 % clear what the reason for the dismissal may have been, so how the data collector reads and interprets the sources he is working on can be somewhat subjective. This is cause for uncertainties regarding the data mining for the dependent variable.

The number of succession events for the 250 firms over the period is 334 and the reason for the succession was recorded in all cases, fulfilling the requirement regarding the dependent variable. The CEO succession events were initially spread amongst 11 different reasons and further into 3 different categories. This reasons for CEO dismissal are shown in table 1 below:

Table 1: Reasons for successions

r	Reason	Number of cases	Percent of classes	Classification
1	Board/power struggle	30	8.98	Forced
2	Move to lesser position	3	0.90	Forced
3	Poor performance	88	26.35	Forced
4	Death or illness	9	2.69	Voluntary
5	Interim CEO	12	3.59	Voluntary
6	Job demands	3	0.90	Voluntary
7	Merger	83	24.85	Merger
8	Planned succession	78	23.35	Voluntary
9	Move to another company	19	5.69	Voluntary
10	Earlier tenure	0	0.00	Voluntary
11	Governance change	9	2.69	Voluntary
	Total	334	100	

Here we see that in around 26 % of the cases the reason for succession is due to poor performance and the CEO is fired from the firm. Planned succession, meaning retirement or contractual agreements, happened in about 23 % of the cases. The last main category for succession events is mergers, which happened in almost 25 % of the cases. This results in that these three reasons account for almost 75 % of the cases.

The categorized reasons are shown below:

Table 2: Distribution of categories

i	Classification	Reasons (*r* in table 1)	Number of cases	Percent of classes
1	Forced	1,2,3	130	38.92
2	Voluntary	4,5,6,8,9,10,11	121	36.23
3	Merger	7	83	24.85
	Total		334	100

This table shows that in almost 39 % of the cases the CEO was forced out of the company. The CEO left "voluntarily" in about 36 % of cases. Note that death or illness has been classified as voluntary due to the fact that it is not a result of an exogenous force.

The independent variables shown in table 3 below will not be grouped in the same fashion as the dependent variable above and the variables regarding industry and nationality. I have therefore put them in table 3 below and will discuss them separately.

Table 3: Basic descriptive statistics

Variable	*n*	Expected value	Standard error	$x_{(1)}$min	$x_{(n)}$max
Dismissal	334				
CEO age	266	56.241	7.156	28	75
CEO tenure	275	6.028	5.445	0.1	43
Market capitalization	332	9789.151	17134.160	895.957	165074.200
Total Stock Return (TSR)	248	-0.016	0.249	-0.676	1.657
Income	175	-0.103	0.851	-4.498	2.713
Market capitalization prior	145	0.148	0.326	-0.726	1.359
TSR prior	250	0.066	0.279	-0.990	1.358
UK exchange listing	270	0.637	0.482	0	1
US exchange listing	270	0.667	0.472	0	1
US board membership	230	0.187	0.391	0	1
US board membership x TSR	181	-0.006	0.106	-0.676	0.630
US exchange listing x TSR	200	-0.019	0.183	-0.672	1.027

Age

This is data in a discontinuously numerical form, ranging from 28 to 75 years. The two cases involving the 75 year olds were due to planned succession. The mean is 56 meaning that the average age of when a CEO is dismissed is when he is 56 years old. The age with the most succession events is 59 and 60 both with a relative frequency of 7.52 %. The number of missing data for this variable is 68.

Tenure

There are 59 cases missing data regarding tenure. A CEO would on average stay in a position for a period of about 6 years before a succession event occurred. The shortest period a CEO was in his position was 0.1 years while the longest was 43 years. In 50 % of the cases a succession would occur after a CEO had been at his position for 4,8 years. The data is of a discontinuous numerical nature.

Market Capitalization

Market capitalization of a firm is defined by Hunt, Moyer and Shelvin (1997) to be share price times the total amount of shares outstanding. The value of the equity is also, according to Penman (2007), decided by future dividends and cash flow. The higher expected future cash flow or dividend payout will result in a relatively high market capitalization.

The data for market capitalization is measured at the beginning of the relevant year, and there are 2 cases where there is no data for this variable. Market capitalization is represented by a continuously numerical variable ranging from 896.0 to 165 074.2, and it has a mean of 9 789.2.

Change in Industry Adjusted Total Stock Return

Total stock return is defined by Penman (2007) to be the change in the price added with dividends paid. So if the firm performs well and becomes more attractive to investors the stock price increases and triggers a positive value for this variable. When it is adjusted with respect to the industry it operates in, the performance of the industry as a whole is taken into consideration, making it easier to see how the firm performed in comparison to industry.

The TSR variable is of a continuously numerical nature describing the change in TSR at a succession event. Data is missing from 86 cases. The average return for all the succession events with data present is close to zero at -0.016. The smallest change in return on stock recorded was -0.676 while the largest change was 1.657.

Change in Industry Adjusted Income

This is a continuously numerical variable showing us the change in industry adjusted income at a succession event. When it has been adjusted with respect to the industry, it is adjusted on the basis of how the change in income for the whole industry was and adjusted accordingly to that.

Data was missing in 159 of the cases. The change in income at a succession event was on average a decrease of -0.103. The lowest change in income was at -4.498 while highest change at a succession event was 2.713.

Change in Market Capitalization prior to an event

A change in market capitalization will tell us something about how the share price for the relevant firm is being valued. As mentioned before the share price is dependent on the predictions of future dividends and cash flow. So a change in the market capitalization will tell us if the future for the relevant firm looks bright (positive change) or dark (negative change).

This variable is just as the original change in market capitalization of a continuously numerical nature. It misses data from a total of 189 cases (more than 56 %). The average change prior to a succession event was positive with 0.148. The change prior to an event ranges from a low at -0.729 to a maximum at 1.359.

Change in Total Stock Return prior to an event

For this variable the nature is continuously numerical and data is missing for 84 events. The variable describes the change in total stock return right before a succession event occurred. The average change in stock return prior to an event was 0,066, while the variable ranges from a low at -0.99 to a high at 1.358.

UK Exchange Listing

This is a variable of binominal nature, and it explains whether a firm is listed on the UK stock exchange. For this variable there were 64 cases missing data on the subject. For the cases with data almost 64 % were listed on the UK stock market.

US Exchange Listing

This is a variable of binominal nature and indicates whether a firm is listed on the stock exchange market in the US. Data was only retrieved for 270 of the cases, missing values in 64 cases. The mean is 66,67 % meaning that 2/3 of the cases were listed on the stock exchange in the US.

US Board Membership

This is a binominal variable describing if a firm has at least one US citizen on the board or not. Data was missing from 104 cases, and it was only in 43 of the cases that there was at least one US citizen on the board which is almost 19 % of the cases were data was present.

US Board Membership x Change in Industry Adjusted Total Stock Return

This variable maintains the continuously numerical nature of the IA TSR, but only counting the events where a US citizen is a member of the board. This happens in 181 cases, missing data in 153 cases. The average change in return with a US citizen board member present is closer to zero than the IA TSR was with a mean on -0.006. The lowest return on stock with a US citizen on the board was -0.676 while the largest return was 0.630.

US Exchange Listing x Industry Adjusted Total Stock Return

This variable also maintains the continuously numerical nature of the IA TSR, but is only counting the succession events where the firm is listed on the US stock market. This is the case for 200 events, lacking data from 134 events. The average change in return for a firm listed on the US stock market is -0.194. For firms being listed on the US stock market the change in industry adjusted total stock return ranges from a low at -0.672 to a high at 1.027.

The two variables above share that they are a result from the initial TSR variable being multiplied with either of the two categorical variables for US board members or US exchange listing. For the regression tree process it means that the tree can differentiate on more levels concerning the variables involved. For instance a possible question can now be; is the TSR of a firm with a US board member > 0?

Industry Variables

The Industry variable is a categorical variable. This variable describes in which industry the succession event happened. The distribution of industries is shown in table 4 below and there is no missing data. When looking at the table we see that the financial industry is clearly the biggest in respect to succession events with almost 25 % of the cases. The smallest industry is energy with around 3 % of the cases.

Table 4: Industry composition

Industry	Numerical code	Number of cases	Percent of classes
Energy	10	11	3.29
Materials	15	34	10.18
Industrials	20	48	14.37
Consumer discretionary	25	49	14.67
Consumer staples	30	23	6.89
Health care	35	13	3.89
Financial services	40	83	24.85
Information technology	45	20	5.99
Telecommunication services	50	24	7.19
Utilities	55	29	8.68
Total		334	100

The type of industries can be further categorized, as shown in table 5 below.

Table 5: Industry categories

Industry	Code(s)	id1	id2	id3	id4	id5
Financial	40	0	0	0	0	0
IT, Telecom	45,50	1	0	0	0	0
Materials, Ind.	15,20	0	1	0	0	0
Consumer	25,30	0	0	1	0	0
Energy, Utilities	10,55	0	0	0	1	0
Health Care	35	0	0	0	0	1

Now that we have categorized the different industries where a succession event may occur, we have another situation where the classifier has different ways to use the industry variable. The questions asked may be of the form is this case in the industry xx or does it belong to the industry category idx.

These dummy variables are normally used in traditional regression analysis and may seem irrelevant and superficial when using CART. They may, however, illustrate an easier understanding of which industry categories are important. It may be easier to see the connection when an industry category variable shows up instead of a list of industry codes. I have therefore chosen to include these dummy variables in the dataset for the analysis.

The Country Variables

This variable is a categorical variable and the data is present in all the cases. The table below shows the composition of the countries in which the succession events occurred, and we see that the countries from the European Union have the most cases with about 93 % of the cases. The UK is clearly the biggest contributor with almost 45 % of the cases. The two next big countries are Germany and France with respectively 13 % and 9 % of the cases. For these three that accumulates to be almost 67 % of the dismissal cases.

Table 6: Country composition

i	Country	Code	Number of cases	Percent of classes
1	Belgium	BEL	6	1.80
2	Czech Republic	CZE	3	0.90
3	Denmark	DEN	4	1.20
4	Finland	FIN	7	2.10
5	France	FRA	31	9.28
6	Germany	GER	43	12.87
7	Ireland	IRE	5	1.50
8	Italy	ITA	20	5.99
9	Luxembourg	LUX	2	0.60
10	Netherlands	NLD	19	5.69
11	Portugal	POR	1	0.30
12	Spain	SPA	10	2.99
13	Sweden	SWE	12	3.59
14	United Kingdom	UK	149	44.61
15	Norway	NOR	7	2.10
16	Russian Republic	RUS	2	0.60
17	Switzerland	SWI	13	3.89
	Total		334	100

Here it can be useful to note that there are some countries that only have been observed a few times. For such countries as the Czech Republic, Denmark, Luxembourg, Portugal and the Russian Republic that have less than five observations, a good generalization is unlikely to be made. This must be kept in mind should some of these countries be deciding factors in the classification process later on.

Furthermore the countries have been grouped up in the following way after region.

Table 7: Country grouping

Class	*i*	cg1	cg2	cg3	cg4	countg
Anglo-Saxon	7,14	1	0	0	0	1
Benelux	1,9,10	0	1	0	0	2
Mediterranean	8,11,12	0	0	1	0	3
Nordic	3,4,13,15	0	0	0	1	4
Rhine	5,6,17	0	0	0	0	0

With this grouping in place the classification tree will have the opportunity to classify succession with respect to country in different ways. It can for instance ask if a case has a XXX nationality or is in cgx country group.

As mentioned under the industry categories, these dummy variables may seem irrelevant when using CART. For these country groups, I deem it even more important to include those into the dataset as they may show a clearer picture of which areas, similar cultures and traditions that are important in the classification process.

Adapting the Data for use in R

I was given access to the dataset, but changes in the dataset were required in order to perform quick and accurate analysis. Even though R is the software used in the analysis, I am more comfortable altering data in Stata so this software was used for this purpose.

A variable not included in the dataset was US listing x TSR so this was created. In order to type commands quickly in R only data used in the analysis had to be in the dataset. The variables that were removed were variables concerning information regarding the class of the observation, which year the observation occurred and some variables regarding information on the nationality of the firm that were irrelevant. The variable with information regarding the class of the observation that was left as the dependent variable was REAS_CAT which had number 1, 2 and 3 for the classes. I further altered this variable in R to show FROCED, MERGER and VOLUNTARY as it makes the resulting trees a lot more understandable and direct. When dismissals due to mergers were excluded I named the dependent variable REASCAT.

A full list of variables with their variables names used in the dataset is listed in appendix A.

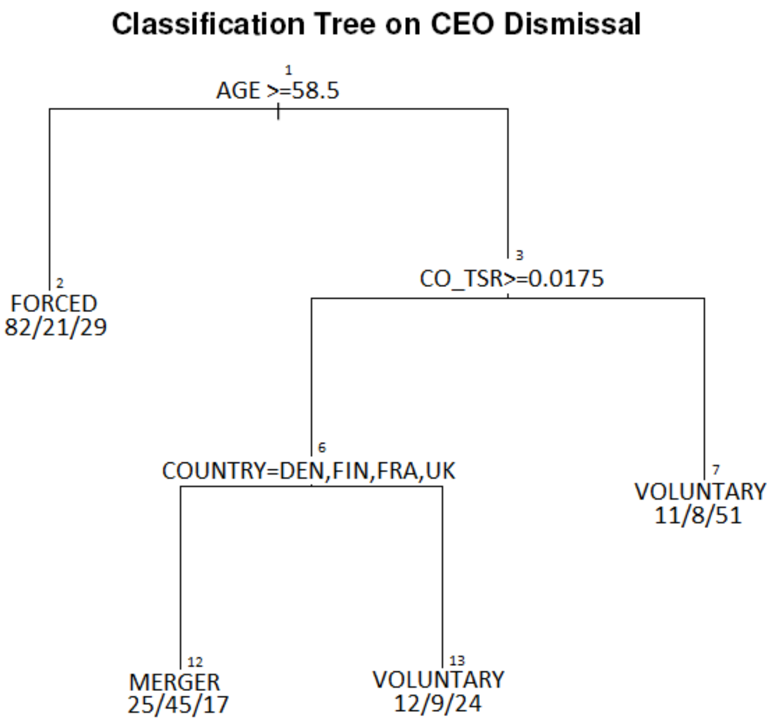
Methodology – Classification and Regression Trees

Introduction

The object of this thesis is to attack the CEO dismissal subject from another angle than the traditional regression analysis. The method that will be used for this purpose is Classification and Regression Trees (CART) by Breiman, Friedman, Olshen and Stone in 1984. This method is a strengthened and further developed version of the Automatic Interaction Detection (AID), developed by Morgan and Sonquist (1963), and THAID, developed by Morgan and Messenger (1973).

All the theory and figures used in this section has been gathered from the book by Breiman *et al.* unless stated otherwise.

The CART is a non-parametric method. It uses the relevant dataset to create a binary tree that classifies the data into groups. An example is shown in the figure below where one of the results is illustrated. This is for illustrational purposes only and analysis of the results will follow in the next section:



Note: This figure is an actual result of the analysis performed later on in the data analysis section on page 30.

The variable and its value used in the split are right above the node where the split happens. The number above the split variable is the node number. At the terminal nodes (nodes 2, 7, 12 and 13) the most frequent class determines the class of cases run through the tree that ends up there. Below the class label is the number of cases from each class. This is listed FORCED/MERGER/VOLUNTARY.

End nodes here represent dismissals of forced, merger or voluntary nature and the variable CO_TSR explains the change in total stock return right before a succession event. This is an example of how a classification tree can end up looking. In order to create a tree like this, one needs to perform a series of statistical procedures.

First, however, it would be natural to explain some of the advantages of CART. Both Morgan and Sonquist (1963) and Armstrong and Scott (1970) explain how the assumptions needed for tree structured analysis is much more relaxed than for the traditional regression analysis. Assumptions such as homogeneity, non-linearity and normally distributed data are non-existing. This relieves restrictions on datasets and allows for full focus on the analysis. The only requirement needed for tree analysis is naturally that all observations used in the learning sample have information on its classification, or dependent variable if you like. Preferred for tree analysis is large datasets, but it is not a necessity. Furthermore missing values on the independent variables are not a problem as second, third and so on splits are stored within the tree. This is also an advantage over regression analysis as it does not have such an elegant way of solving this problem. More on missing values are included later on in this section.

Basic Terminology

The dataset will consist of a certain number of variables: $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_{\mathrm{n},}$. The measurements $(\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{n})$ will be defined as the *measurement vector* $\mathbf{x}$, and the *measurement space* X will be defined to contained all possible measurement vectors. This will give X the property of being an n-dimensional space.

The cases will fall into different classes, the number of classes can range from 1 to J classes and C will denote the set of classes: $C = \{1, \dots, J\}$. To predict class membership we have the classifier $d(\mathbf{x})$ that will for every measurement vector $\mathbf{x}$, assign a class membership in C. We can also see the classifier by defining A_j as a subset of X for which $d(\mathbf{x}) = j$:

$$A_j = \{\boldsymbol{x};\ d(\mathbf{x}) = j\}$$

The sets $A_1, A_2, \dots, A_j$ are disjoint and $X = \cup_j A_j$. A_j will be a partition of X and for every $\mathbf{x} \in A_j$ the predicted class will be j.

In order to be able to construct a satisfactory classifier one must have a sufficient dataset or *learning sample,* $\mathcal{L}$. The learning sample will consist of data on *N* cases together with their classification: $\mathcal{L} = \{(\mathbf{x}_1, j_1), \dots, (\mathbf{x}_N, j_N)\}$, where $\mathbf{x}_n \in X$, $j_n \in C$ and $n = 1, \dots, N$. In the learning sample the variables in the measurement vector can have two different types; one is numerical (real numbers) while the other type is categorical (red, white, blue).

Estimation Accuracy

When one has collected the learning sample and created the classifier, $d(\mathbf{x})$, one will be interested in seeing how good a predictor the classifier is. To find the accuracy of the classifier we denote the *true misclassification rate*, $R^*(d)$.

The basic idea to test the accuracy of the classifier is to use the whole or part of the learning sample to create the classifier, and then draw a new learning sample from the same population as the previous learning sample was drawn from or use the remainder of the learning sample to test it. In the test procedure one looks at each case to see what class it really belongs to and then compare it to which class the classifier predicts it should belong to.

There are a few different ways of estimating $R^*(d)$, each good for its own uses. I will give a brief explanation on two of them and a more thorough explanation of a third one that is most suited to the dataset used.

A commonly used, but the least accurate estimate is the *resubstitution* estimate. Here one uses the whole learning sample to create $d(\mathbf{x})$ and then run the cases through the same classifier. This makes the resubstitution estimate biased, and for a larger number of terminal nodes the misclassification rate will go down using the estimate. The true misclassification rate however may decrease if the tree grows too large. The resubstitution estimate is:

$$R(d) = \frac{1}{N}\sum_{n=1}^{N} \chi[d(\mathbf{x}_n) \neq j_n]$$

Where χ is an indicator function which equals 1 for when $d(\mathbf{x}_n) \neq j_n$.

The second method is *test sample* estimation where one divides the learning sample, $\mathcal{L}$, into two sets, $\mathcal{L}_1$ and $\mathcal{L}_2$. Only cases from $\mathcal{L}_1$ are used to create the classifier, and then the cases from $\mathcal{L}_2$ are used to estimate $R^*(d)$. A split of 2/3 and 1/3 has become a normal split, and to help ensure independence of cases in $\mathcal{L}_2$ from $\mathcal{L}_1$ the cases should be drawn at random. The number of cases in $\mathcal{L}_2$ is denoted by N_2. The test sample estimate is:

$$R^{ts}(d) = \frac{1}{N_2}\sum_{(\mathbf{x}_n, j_n) \in \mathcal{L}_2} \chi[d(\mathbf{x}_n) \neq j_n]$$

The third method, preferred for smaller sample sizes according to a review by M. Stone (1977), is called *v-fold cross-validation*. As the dataset used in this thesis has 334 cases, it is considered quite small so the *v-fold cross-validation* is the approach I will use to create and test the classifier. Therefore, I will explain it in depth.

V-fold cross-validation

For this procedure the cases in $\mathcal{L}$ will randomly be divided into V subsets of close to or equal size. The subsets will be denoted $\mathcal{L}_1, \dots, \mathcal{L}_V$. Next one will create a classifier using the learning sample $\mathcal{L} - \mathcal{L}_v$ for every $v, v = 1, \dots, V$. Each learning sample will create its own classifier denoted $d^{(v)}(\mathbf{x})$, $\mathcal{L}_v$ will not have been involved in estimate of $d^{(v)}$and we can use the test sample estimate($R^{ts}(d)$) for $R^*\left(d^{(v)}\right)$:

$$R^{ts}\left(d^{(v)}\right) = \frac{1}{N_v} \sum_{(\mathbf{x}_n, j_n) \in \mathcal{L}_v} \chi[d^{(v)}(\mathbf{x}_n) \neq j_n],$$

where $N_v \simeq N/V$ is the number of cases in $\mathcal{L}_v$.

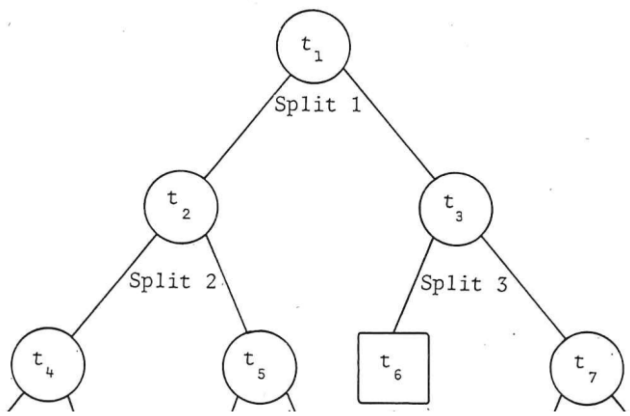
Next step is to create the classifier d using the whole learning sample $\mathcal{L}$. If the number of subsets, V, is large, all the V classifiers will be based on a learning sample nearly as large as $\mathcal{L}$. An assumption being made using the cross-validation procedure is that the procedure is stable; all the classifiers $d^{(v)}$ has been constructed using almost whole $\mathcal{L}$ will have misclassification rates $R^*\left(d^{(v)}\right)$ almost equal to $R^*(d)$. With this in mind we can define the V-fold cross-validation estimate $R^{CV}(d)$:

$$R^{CV}(d) = \frac{1}{V} \sum_{v=1}^{V} R^{ts}\left(d^{(v)}\right)$$

In the cross-validation procedure every case in $\mathcal{L}$ is used to create d, and every case is used once in a test sample. This is an important feature when one is working with a small learning sample and cannot afford to “loose” one third of the learning sample using the test sample procedure. Breiman *et. al.* also states that estimators resulting from the cross-validation procedure get satisfactorily close to $R^*(d)$ on simulated data.

Tree Growing Methodology

Here I will explain the tree growing procedure, that is what kind of questions are used for splits, goodness of the split, how to decide when to stop splitting and how to assign classes to the terminal nodes. First though, I will use a couple of figures to explain some terminology and to get a clearer view of the tree growing process:



(Cut from figure 2.4 from Breiman, Friedman, Ohlsen and Stone (1984), p. 23)

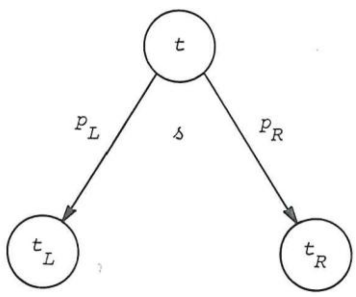
This is just a part of an existing tree, but it is enough to demonstrate what is needed. First we have the root node, $t_1 = X$. After split 1 there are subsets of X and $t_2 \cup t_3 = t_1 = X$. Square nodes are called terminal nodes, meaning the cases that end up there will be assigned a class.

Questions and the Splitting Rule

The object of the binary splits is to decrease the impurity of the descendant subsets. By purity I mean the mixture of classes in a node, it will be 100 % pure when a node contains only one class. The measure of impurity is denoted $i(t)$. In order to measure the goodness of a split, the decrease in impurity is used:

$$\Delta i(\mathcal{s}, t) = i(t) - p_L i(t_L) - p_R(i(t_R)$$

This function will have a greater meaning after looking at this figure:



(Figure 2.6 from Breiman, Friedman, Ohlsen and Stone (1984), p. 25)

For any node t, there is a split, $\mathcal{s}$, that divides the node into t_L and t_R. The proportion of the cases in node t that goes into t_L is p_L and for t_R it is the proportion p_R. At each node there will be a set, $\mathcal{S}$, of binary splits that has been generated from a set of questions, $\mathcal{Q}$. The Questions in $\mathcal{Q}$ will have the form:

Is $\mathbf{x} \in A$?, $A \subset X$

Each split will only depend on one variable, and an example of questions can for instance be; is $x_1 \leq 72$?, is $x_2 = \{red, blue\}$? All the cases answering 'yes' gets sent to the left and right if "no" is the answer.

As there will be a set of questions at each node, the question that decreases impurity by the most gives the best split, $\mathcal{s}^*$, for node t:

$$\Delta i(\mathcal{s}^*, t) = \max_{\mathcal{s} \in \mathcal{S}} \Delta i(\mathcal{s}, t)$$

When the tree has been created and one wants to classify new data all the questions are stored in case the new data is missing observations on some of the variables. Having ordered the set of questions according to the decrease in impurity, one can then use the second best question as a surrogate split, $\mathcal{s}'$. The surrogate splits can also be used in the creation of the tree so that data missing observations on some of the variables (observed class must be present) can still be part of the learning sample.

The Stop-Splitting Rule

Suppose now that we have created a tree, T, with a current set of terminal nodes, $\tilde{T}$. Set $I(t) = i(t)p(t)$ and define the tree impurity, $I(T)$:

$$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t)p(t)$$

In words this is the sum of the impurity of all the terminal nodes multiplied by the proportion of total cases in that terminal node. So to decrease the impurity further one can simply split any of the terminal nodes, $t \in \tilde{T}$, using split $\mathcal{s}$, into t_L and t_R. The impurity of the new tree, T', will be:

$$I(T') = \sum_{\tilde{T}-\{t\}} I(t) + I(t_L) + I(t_R)$$

The decrease in impurity will be:

$$I(T) - I(T') = I(t) - I(t_L) - I(t_R)$$

The decrease in impurity only the depends on the node t and the split $\mathcal{s}$, and we can then write the decrease in impurity as:

$$\Delta I(\mathcal{s}, t) = I(t) - I(t_L) - I(t_R)$$

By maximizing this expression the impurity of the whole tree will decrease. In theory you would then keep splitting nodes until you end up with all nodes being pure, containing only one class. With a learning sample of any size it would quickly result in a too large and complex tree to be of any use. Therefore it is important to stop the splitting at a point before the tree becomes too big, but still works as a classifier. The initial stop-splitting rule was to help decide when to stop splitting and make a node terminal:

$$\max_{\mathcal{s} \in \mathcal{S}} \Delta I(\mathcal{s}, t) < \beta$$

where β is a set threshold greater than zero. This means that if the decrease in impurity resulting from a potential split, $\mathcal{s}$, is lower than β, the decrease in impurity is not great enough to improve the classifier on cost of, for instance, increased complexity.

Stop-splitting rules have, however, proved to not compute the best trees. In some cases where β is set too low there may be too much splitting and the tree will become too large and complex. Should β be set too high the tree may be stopped too early. There can for instance be a node t where $\max_{\mathcal{s}} \Delta I(\mathcal{s}, t)$ is small, but the descendant nodes t_L and t_R may have splits with a significant decrease in impurity. It would be a shame to miss out on such splits. The solution for this problem is to grow a large tree and then "prune" it to a right sized tree. I will discuss this technique after the last tree growing rule.

Class Assignment Rule

Having reached a complete tree, T, only the class assignment of terminal nodes, $\tilde{T}$, remains. The class with the most cases in a terminal node will determine the class that the terminal node will be assigned. The class assignment rule, $j(t)$, is

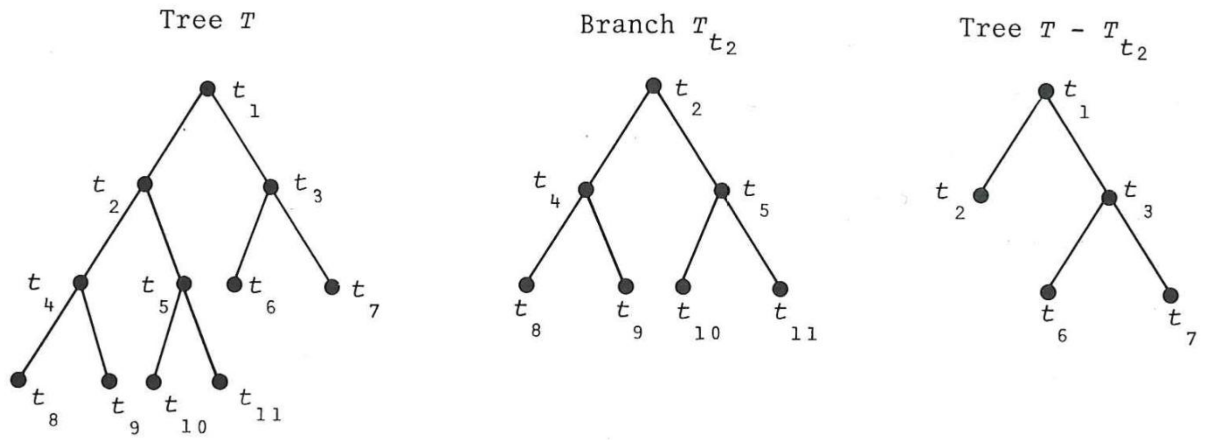
$$j(t) = \max_{j} p(j|t)$$

where $p(j|t)$ is the proportion of j classes in the terminal node.

Tree Pruning

As mentioned earlier the results using stop-splitting rules were unsatisfactory and the answer to this problem was to look at it from another angle. By letting the tree grow so large that it either had only one case in each terminal node or had terminal nodes consisting of only cases of one class each, a 100 % pure tree, and then starting to prune of branches in order to get a tree that would not be too big and still be a strong classifier. The large tree will be denoted T_{max}.

To illustrate how pruning works we can look at the figure below:



(Figure 3.1 from Breiman, Friedman, Ohlsen and Stone (1984), p. 64)

We start off with the initial tree T and decide to prune of the branch T_{t_2}. The resulting tree will be the tree $T - T_{t_2}$, where the branch T_{t_2} is removed, leaving only the node t_2 as a terminal node.

The new tree $T - T_{t_2}$ is a pruned subtree of the initial tree T. In general we denote a pruned subtree of T, T'. To say it is a subtree we write it like this: $T' \prec T$. The initial tree and the pruned subtree has the same root node, t_1.

With a large T_{max} there will be a great number of different ways to prune it into a smaller tree. This means that there will be several pruned subtrees with same number of terminal nodes. It is then important to have a method to seek out the best subtree for the given number of terminal nodes. The method developed for this purpose is called minimal cost-complexity pruning.

Minimal Cost-Complexity Pruning

In minimal cost-complexity pruning the complexity is the number of terminal nodes, $|\tilde{T}|$, in the subtree $T \prec T_{max}$. There is also a complexity parameter, $\alpha > 0$, working as complexity cost per terminal node. This together with the resubstitution misclassification rate we get a linear relationship for the cost-complexity measure $R_\alpha(T)$ for the subtree T:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

To find the best subtree for a value of α we need to minimize $R_\alpha(T)$ for the subtree $T(\alpha) \prec T_{max}$:

$$R_\alpha[T(\alpha)] = \min_{T \prec T_{max}} R_\alpha(T)$$

For small values of α, the penalty for larger number of terminal nodes is small and therefore allows the trees to grow large. Is the value of α high, the penalty for the number of terminal nodes is greater and smaller trees will now minimize the cost-complexity measure. This means that for different values of α, different sized trees will be optimal. When one has obtained a series of optimal trees, each of different size, having a different number of terminal nodes, one will have to select which optimal tree is the right sized subtree. The trees obtained will be $T_1 \succ T_2 \succ \cdots \succ T_k \succ \cdots \succ \{t_1\}$, with $T_k = T(\alpha_k)$ and $\alpha_1 = 0$. T_1 will be the largest of the pruned subtrees while $\{t_1\}$ is the smallest, containing just the root node.

To select which tree is the right sized subtree one cannot use the resubstitution estimate as it is biased and would always choose T_1. The cross-validation estimate (and test sample estimate) is less bias, but a probability model is needed to study the bias and standard error of an estimate:

The space $X \times C$ will be defined as a set of all couples $(\boldsymbol{x}, j)$, where $\boldsymbol{x} \in X$ and $j \in C$. On $X \times C$ we have the probability $P(A, j)$, with $A \subset X$ and $j \in C$. $P(A, j)$ is the probability that a case drawn at random from the relevant distribution will have its measurement vector in A and that its class is j.The learning sample $\mathcal{L}$ consists of N different cases randomly drawn from the distribution $P(A, j)$. From $\mathcal{L}$ we will create $d(\mathbf{x})$ and then define $R^*(d)$ to be the probability that d will misclassify a new case, $(\mathbf{X}, Y)$ drawn from the same distribution as the learning sample.

$$R^*(d) = P(d(\mathbf{X}) \neq Y)$$

This definition assumes that the cost of misclassifying a class i into a class j has the same cost as j into i. In order to account for this we define $Q^*(i|j)$ to be the probability that d will classify a class j as a class i.

$$Q^*(i|j) = P(d(\mathbf{X}) = i | Y = j)$$

Then we define $R^*(j)$ to be the expected cost of misclassifying class j cases.

$$R^*(j) = \sum_i C(i|j) Q^*(i|j)$$

Finally we will define $R^*(d)$ to be the expected misclassification cost for the classifier d.

$$R^*(d) = \sum_j R^*(j) \pi(j)$$

Where $\pi(j)$ is the class probabilities for $j = 1, \dots, J$, $\pi(j) = P(Y = j)$.

Finding the Best Pruned Subtree with Cross-Validation Estimates

With this established we can now look into how we use cross-validation as a tool in the tree pruning process. We start off by growing v large and pure trees $T_{max}^{(v)}, v = 1, \dots, V$, and also the normal tree T_{max}. The trees $T(\alpha)$ and $T^{(v)}(\alpha)$,$v = 1, \dots, V$, will for each value of α be the corresponding minimal cost-complexity subtrees of T_{max}, $T_{max}^{(v)}$. For each v, the trees $T_{max}^{(v)}$ and $T^{(v)}(\alpha)$ the cases in $\mathcal{L}_v$ has been excluded from the learning sample when creating those trees. This means that the cases in $\mathcal{L}_v$ can be used as an independent test sample for

$T^{(v)}(\alpha)$. The next step is to run $\mathcal{L}_v$ through $T_{max}^{(v)}$ for all values for v, α must be held constant. For each value of v, i, j, we define $N_{ij}^{(v)}$ to be the number of class j cases classified as class i in $\mathcal{L}_v$ by the tree $T^{(v)}\alpha$. The total number of class j cases classified as class i is denoted

$$N_{ij} = \sum\nolimits_v N_{ij}^{(v)}$$

As the cases in $\mathcal{L}$ is only used as a test sample once, the total number of class j cases in all the test samples will be N_j which is the number of class j cases in the learning sample.

For a large V the classification accuracy of the tree $T^{(v)}(\alpha)$ should be almost equal to the accuracy of the tree $T(\alpha)$. This allows us to estimate $Q^*(i|j)$ for the tree $T(\alpha)$:

$$Q^{CV}(i|j) = N_{ij}/N_j$$

This will give us;

$$R^{CV}(j) = \sum\nolimits_i C(i|j)Q^{CV}(i|j)$$

and;

$$R^{CV}(T(\alpha)) = \sum\nolimits_j R^{CV}(j)\pi(j)$$

If the class probabilities have been data estimated we can set $\pi(j) = N_j/N$ and rewrite $R^{CV}(T(\alpha))$ into

$$R^{CV}(T(\alpha)) = \frac{1}{N}\sum_{i,j} C(i|j)N_{ij}$$

This is the proportion of test set cases that have been misclassified.

As the complexity parameter α has continuous values the tree $T(\alpha)$ will be the best pruned subtree for an interval of α values until α increases until a jump point, α', is reached and a new tree, $T(\alpha')$, will be the new best pruned subtree, and so on.

This means that the minimal cost-complexity trees constructed by $\mathcal{L}$ are equal to T_k for $\alpha_k \leq \alpha < \alpha_{k+1}$

Next we set α'_k to be a geometrical midpoint so that $T(\alpha) = T_k$; $\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}$. With this we can alter the $R^{CV}(T(\alpha))$ expression;

$$R^{CV}(T_k) = R^{CV}(T(\alpha'_k))$$

This expression is the estimate received after running the test samples $\mathcal{L}_v$ through the trees $T^{(v)}(\alpha'_k)$.

We can now set the rule for selecting the right sized tree, T_{k0}: Select the tree T_{k0} so that

$$R^{CV}(T_{k0}) = \min_{k} R^{CV}(T_k)$$

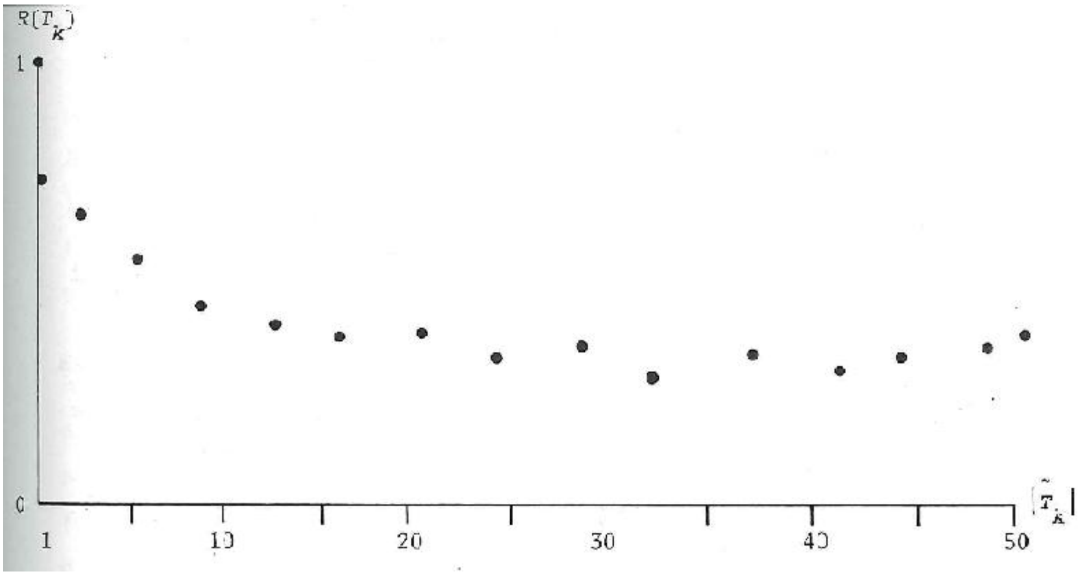
$R^{CV}(T_{k0})$ will also be used as the estimate of the misclassification cost of the right sized tree.

The 1 SE Rule

As $R^{CV}(T_{k0})$ is an estimate it may be helpful to evaluate the result by also looking at the standard error of the tree. $R^{CV}(T_{k0})$ is the misclassification ratio and works as the probability for cases to be misclassified by the tree T_{k0}. This probability was worked out by running the cases in $\mathcal{L}_v$ through the tree, and the number of cases in $\mathcal{L}_v$ is N_v. We can now use this to calculate the standard error of the tree T_{k0};

$$SE\big(R^{CV}(T)\big) = \big[R^{CV}(T)\big(1 - R^{CV}(T)\big)/N_v\big]^{1/2}$$

To see how we can use the standard error it is helpful to look at the cross-validation estimate, $R^{CV}(T_k)$, as a function of the number of terminal nodes, $|\widetilde{T_k}|$. The plot often looks similar to the one in the figure below;



(Figure 3.2 from Breiman, Friedman, Ohlsen and Stone (1984), p. 72)

We can see that the decrease in misclassification ratio is rapid at first, before it slows down and stays somewhat steady before it starts to increase slightly later on. When finding the right sized tree in the flat, steady part we see that there is not much difference to the next best trees and small changes in parameter values or even the random seed that selects the different sets used in cross-validation may change what number of terminal nodes that minimizes $\hat{R}$. The 1 SE rule was then created to reduce this instability. The 1 SE rule works so that you add one standard error to the misclassification rate estimated and choose the smallest tree;

$$R^{CV}(T_{k1}) \leq R^{CV}(T_{k0}) + SE\left(R^{CV}(T)\right)$$

So the tree to be chosen is now the smallest tree that has cross-validation estimate less than or equal to the cv estimate of the right sized tree added with its standard error. On the figure above, this would trigger a move to the left.

Analysis software

As mentioned in earlier parts, CART is not as common as traditional regression analysis and thereby it is not perfectly implemented in all types of statistical software. Stata for instance has a tool for tree structured problems, but it is hardly focused on survival analysis and not fitted for the problem discussed in this thesis. It was decided to use a collection of software packages called R[1]. The homepage says that "R is a language and environment for statistical computing and graphics."[2] By environment they mean that R supports a wide number of statistical and graphical techniques. R is a flexible environment in the way that it can be easily expanded via packages. A vast number of packages are available via the CRAN network[3].

A downloadable package called 'tree' was suggested by my supervisor to be used. In this package however it was not easy to include variables that had missing data. The result would have been a learning sample consisting of 59 cases and a good reflection on dismissal problems could not be assured. Another package called 'rpart'[4] was discovered by me and it proved to be a lot easier to get working with the entire dataset. In the recommended introduction file for the package, Therneau and Atkinson (1997)[5] explain how the software should be very close to the CART program. The only notable difference is concerning surrogate splits that are calculated in a different way from CART. This would only prove a problem should there be a high number of missing observations on some of the variables. As there are missing observations on some of the variables we must bear this in mind as it may affect the surrogate splits at some nodes. Normally the variables listed in sequence after the primary split will be used as surrogates, but should they miss observations rpart may suggest other variables as surrogates.

No commands used to grow the trees will be stated in the next section. If there is an interest in growing the trees yourselves, please go to appendix B on page 45, to get an example and overview on how to grow the trees.

---

1 http://www.r-project.org/
2 http://www.r-project.org/index.html
3 http://cran.uib.no/
4 http://www.statmethods.net/advstats/cart.html
5 http://www.mayo.edu/hsr/techrpt/61.pdf

Data Analysis

Creating the Tree

In this process the succession events due to mergers are include in order to see whether or not CART will be able to make a good classifier if dismissals due to merger activity is included explicitly as a reason for dismissal.

The Initial Tree

An initial tree was created using 10-fold cross-validation. The cost complexity parameter was set to be equal to minus infinity in order to grow a large tree. The resulting tree had 26 terminal nodes and a cross validation misclassification rate estimate of 0.5359. The resubstitution estimate is 0.3323 and therefore not pure as one would expect of a T_{max} tree. However, if the tree is a pruned subtree of T_{max}, there will be no difference in the results. The table given below contains a summary of the best pruned subtrees of T_{max}. Variables that were used in the actual tree were: AGE, CO_TSR , COUNTRY, IA_TRS, id3, IND_CAT, MKTVAL2, TENURE and USLIST. A table showing the misclassification errors, standard errors and complexity parameters of different sized best pruned trees follows:

Table 1: List of best pruned trees

Tree	CP	nsplit	R-error	X-error	X-std
1	0.2156863	0	1.00000	1.05392	0.042903
2	0.0686275	1	0.78431	0.82843	0.044790
3	0.0147059	3	0.64706	0.75980	0.044678
4	0.0122549	4	0.63235	0.82353	0.044792
5	0.0098039	6	0.60784	0.82843	0.044790
6	0.0016340	12	0.54902	0.87255	0.044696
7	0.0000000	15	0.54412	0.87745	0.044678
8	-Inf	25	0.54412	0.87745	0.044678

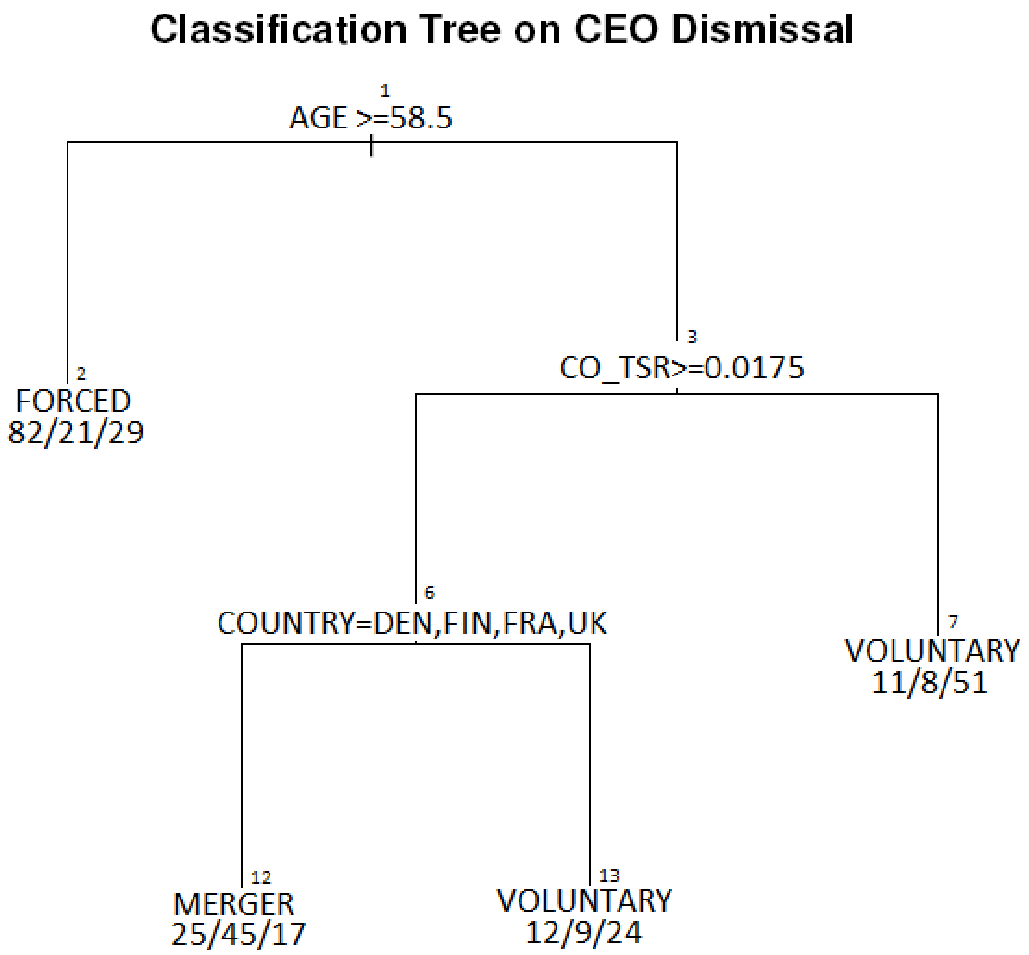
Note: The errors in the table are set so that they are relative to the root node and must be multiplied with the root node misclassification error = 204/334 = 0.61078. Also, the nsplit column shows how many splits are made. To get the number of terminal nodes in a tree one must add one, nsplit + 1.

The R-error, or resubstitution estimate is as mentioned in the methodology part biased in the way that the whole learning sample is used to create the tree and then run through the same tree in order to calculate the resubstitution estimate for misclassification error. This means that the bigger you grow the tree, the lower the resubstitution estimate will get, but the true misclassification ratio may increase should the tree grow too large. The X-error, or cross-validation estimate, is a better measure as the data run through the tree has not been part of the subset creating the tree. This error and the cross-validation standard error are the estimates we must look at to decide which tree is best.

By using the 1 SE rule, $R^{CV}(T_{k1}) \leq R^{CV}(T_{k0}) + SE\left(R^{CV}(T)\right)$, we get that the right sized tree in this case has an unadjusted cross-validation error of less than (0.75980 + 0.044678) 0.804478. Off the table we see that the smallest (and only) tree that satisfies the 1 SE rule is tree number 3 with 3 split,4 terminal nodes.

The Right Sized Tree

Even before creating the right sized tree we can get the resubstitution estimate and the cross-validation estimate for the table above. The resubstitution estimate is 0.3952, while the cross-validation estimate is 0.4641. As the resubstitution estimate is biased and lower than the true misclassification rate and the cross-validation estimate overrates the ratio, we can get a better picture by looking at both of them together (Breiman, *et. al.*, 1984). This means that from the right sized tree we can expect a misclassification ratio between 40 % and 46 %. This means that at best the tree will successfully predict the class of a case in six out of ten cases, while it at worst will misclassify almost half of the cases dropped through the tree. Let us now have a look at the tree:



Note: The numbers below the category of dismissal in each terminal node is arranged alphabetically: FORCED/MERGER/VOLUNTARY. Nodes are counted from left to right and downwards, and the node number is shown above the split variable or class of the node.

We see that the variables used in the classification of CEO dismissal are AGE, CO_TSR and COUNTRY. It is interesting to see that if a CEO is 58.5 years old or older, the case will be classified straight away as a forced dismissal. Age is also the only factor deciding whether a CEO will be dismissed or not.

In order for a voluntary dismissal to occur, the age of the CEO has to be below 58.5 and the change in total stock return right before the succession event had to be below 1.75 %. Should the change in TSR right before the event be higher the dismissal could still be classified as voluntary should the CEO not be from Denmark, Finland, France or the UK. Note that Denmark is only represented by a few observations and a good generalization is not certain.

So for a succession event to be classified as a merger case the CEO would have to be less than 58.5 years old, the company would need a change in TSR right before the event to be higher than 1.75 % and the company would have to be located in Denmark, Finland, France or the UK.

Composition of the Nodes

As these results might seem a bit odd, it may be helpful to look at the composition of the different nodes, and especially the terminal nodes, to see how accurate they really are:

Table 2: Composition of the nodes

Node	n-cases	n-wrong	Class	% forced	% merger	% voluntary
1)	334	204	Forced	38.92	24.85	36.22
2)*	132	50	Forced	62.12	15.91	21.97
3)	202	110	Voluntary	23.76	30.69	45.54
6)	132	78	Merger	28.03	40.91	31.06
12)*	87	42	Merger	28.74	51.72	19.54
13)*	45	21	Voluntary	26.67	20.00	53.33
7)*	70	19	Voluntary	15.71	11.43	72.86

Note: The terminal nodes are marked with * at the end of the node number.

Node 2 classifies 62 % of the cases right and has 22 % voluntary cases and 16 % merger cases misclassified. Node 12 correctly classifies 52 % of the cases as merger cases while as much as 29 % forced cases are misclassified. 20 % voluntary cases are also misclassified here. Node 13 has only slightly better stats by correctly classifying 53 % as voluntary cases while 27 % forced cases and 20 % merger cases are misclassified in this terminal node. Node 7 is the most successful terminal node, correctly classifying 73 % of the cases as voluntary, while 16 % forced and 11 % merger cases are being misclassified should they end up in this root node.

The impurity in the nodes 12 and 13 is rather high with the dominant class barely being above half of the cases in the nodes. Relatively, a lot of the misclassified cases come from these two terminal nodes.

Primary and Surrogate Splits

As noted in the description of the data section, there are missing observations on several variables. Another use for primary and surrogate splits is to see if other variables not used in the tree still can be significant. At a node where two variables have almost the same goodness in the split, the best one will mask the second one. Small changes in the dataset may cause the variables to switch place. This makes it useful to look at the primary and surrogate splits for the non-terminal nodes to see whether a variable may be masked or not:

Note: Even though only five surrogate splits are shown in these tables, the surrogate split for all variables is computed and remembered within the tree should there be use for them.

Node 1) Splits for tree, node 1

Primary splits			
AGE	< 58.5	to the right,	improve=20.167900, (68 missing)
CO_TSR	< 0.0975	to the right,	improve=12.094410, (84 missing)
IA_TRS	< -0.173	to the right,	improve= 8.430332, (86 missing)
IA_INCOM	< -1.8825	to the right,	improve= 6.670823, (159 missing)
TENURE	< 11.9	to the right,	improve= 6.407481, (59 missing)
Surrogate splits			
COUNTRY	splits as	LRRRLLLRLLRLRRRRR,	agree=0.680, adj=0.248, (68 split)
TENURE	< 6.9	to the right,	agree=0.650, adj=0.177, (0 split)
countg	< 0.5	to the left,	agree=0.609, adj=0.080, (0 split)
MKTVAL2	< 8056.325	to the right,	agree=0.602, adj=0.062, (0 split)
cg2	< 0.5	to the right,	agree=0.598, adj=0.053, (0 split)

Note: The improvement is *n* times the change in the impurity index. Agree is how much the surrogate and the primary split agrees, it is calculated using total number of agreements.

As you may notice the second primary split and first surrogate split is not the same, this is due to the way the 'agree' measure is calculated in R. While CO_TSR has 84 missing values, COUNTRY has no missing values and has in this case most cases that agree with the split of AGE. The agreement is 68 % which is not the best, but it includes the cases where the age of the CEO is missing. TENURE is the second best surrogate split, and is almost a good a surrogate as the COUNTRY variable. Further we see that countg and cg2 also are surrogate splits. Their meaning here represents that countries from the Rhine area (countg < 0.5) are more likely to dismiss a CEO than other countries.

The improvement in the purity multiplied with the number of cases is 20.17 which give that decrease in the impurity index for node 1 is 0.0604. The drop in the improvement in the purity is quite large and the fact that the COUNTRY variable is not even amongst the top five primary splits, the use of this as a surrogate split may not be the best. The reason it appears as the best surrogate split is, however, because of the way surrogate splits are created in R.

<u>Node 3)</u> <u>Splits for tree, node 3</u>

Primary splits			
CO_TSR	< 0.0175	to the right,	improve=7.800188, (62 missing)
COUNTRY	splits as	RRLLRR-L-RL-LRRRL,	improve=6.293009, (0 missing)
ncountry	splits as	RRLLRR-L-RL-LRRRL,	improve=6.293009, (0 missing)
IA_TRS	< -0.0835	to the right,	improve=4.742339, (63 missing)
countg	< 0.5	to the right,	improve=3.837950, (4 missing)
Surrogate splits			
IA_TRS	< -0.0955	to the right,	agree=0.843, adj=0.627, (0 split)
TENURE	< 2.35	to the right,	agree=0.707, adj=0.305, (21 split)
COUNTRY	splits as	RLLLRL-L-LL-LLLRL,	agree=0.643, adj=0.153, (41 split)
MKTVAL2	< 1720.617	to the right,	agree=0.643, adj=0.153, (0 split)
AGE	< 44.5	to the right,	agree=0.614, adj=0.085, (0 split)

Note: The "-" in COUNTRY variable is due to the fact that no cases coming to node 3 were a company located in any of the missing countries.

Here the improvement in impurity has dropped quite a lot, which may be due to there being fewer cases at the node. Looking at the decrease in impurity index it is only 0.0386. The goodness of this split is lower than the goodness of the split in node 1.

The best surrogate split for change in TSR right before a succession event is the industry adjusted change in TSR. The agreement between these two is quite high (84 %), making it a good surrogate and supporting the importance of performance being a good discriminator at this point. However, as it is missing much of the same data, tenure and country are the next ones making splits in this case. The agreement between tenure and CO_TSR is at 71 % while country agrees in 64 % of the cases. The country variable, however, is not too far down regarding the decrease in impurity.

Node 6)   Splits for tree, node 6

Primary splits			
COUNTRY	splits as	-RLLLR-R-RR-RRRRL,	improve=6.384639, (0 missing)
countg	< 0.5	to the right,	improve=3.133986, (3 missing)
cg1	< 0.5	to the right,	improve=2.840067, (0 missing)
IA_INCOM	< -0.311	to the right,	improve=2.301666, (70 missing)
AGE	< 48.5	to the right,	improve=2.093678, (45 missing)
Surrogate splits			
cg1	< 0.5	to the right,	agree=0.932, adj=0.800, (0 split)
countg	< 1.5	to the left,	agree=0.826, adj=0.489, (0 split)
cg3	< 0.5	to the left,	agree=0.765, adj=0.311, (0 split)
cg2	< 0.5	to the left,	agree=0.697, adj=0.111, (0 split)
cg4	< 0.5	to the left,	agree=0.682, adj=0.067, (0 split)

Note: Please have a look on page 16 where I motivated the reasons for including the dummy variables of countries.

The best split for node 6 is the variable COUNTRY. As it is easy to determine where a company is located there are no missing values, and all cases are run through this variable. The split decreased the impurity index by 0.048. What the variables countg and cg1 represent here is that countries located in the Rhine are less likely to dismiss a CEO due to a merger, while Anglo-Saxon countries are more likely to do so.

Should the information about the nationality of a firm be missing, we see that there is a rather big drop in purity improvement to the IA_INCOM variable. This is the best primary variable not using information regarding the nationality of the firm. This means that if a case is missing information the nationality of a firm, the risk of misclassification increases.

The two rather low decreases in the impurity index from splits 3 and 6 ads up to the large misclassification rate in terminal nodes 12 and 13. Had there been a variable making a better split, some of the cases might would have switched places and improved the misclassification rates. This, however, is not the case, and it is good that the COUNTRY variable is not missing any data as it is clearly the best split in node 6.

Excluding the Merger Cases

As the computed classification tree with dismissal cases due to merger activity was a quite poor classifier, almost misclassifying half of the cases run through the tree, excluding the merger cases may create a better tree. A tree that would with greater power explain which factors play a role in either forced or voluntary succession events than the previous tree.

The Initial Tree

Just as before, I start by growing a large tree to later prune it down to the right sized tree. Having removed the merger cases, the number of cases was decreased by 83. The learning sample now consists of 251 observations and the same variables as before. For the new learning sample the initial tree has 19 terminal nodes and a cross-validation estimate at 0.3623. We note that this is considerably lower than for the initial tree where mergers were included (0.5359). The variables used in this tree are; AGE, CO_TSR, COUNTRY, IA_TRS, IND_CAT, indcat and TENURE. We see that there is a slight difference in which variables are included in this tree. The only new variable is indcat while MKTVAL2, id3 and USLIST have been removed. Note that indcat is still included in the dataset even though it is connected to the dummy variables for industry. The reason for including the industry dummy variables was discussed on page 14. To see how much that needs to be pruned away I show the list of best pruned trees and their corresponding errors and complexity parameter:

Table 3: List of best pruned trees

Tree	CP	nsplit	R-error	X-error	X-std
1	0.347107	0	1.00000	1.00000	0.065425
2	0.033058	1	0.65289	0.67769	0.061408
3	0.028926	6	0.48760	0.79339	0.063633
4	0.022039	8	0.42975	0.72727	0.062476
5	0.000000	11	0.36364	0.75207	0.062945
6	-Inf	18	0.36364	0.75207	0.062945

Note: The errors here might not seem considerably lower than for the previous tree, but bear in mind that you have to multiplied the errors here with the root node error of 121/251 = 0.48207.

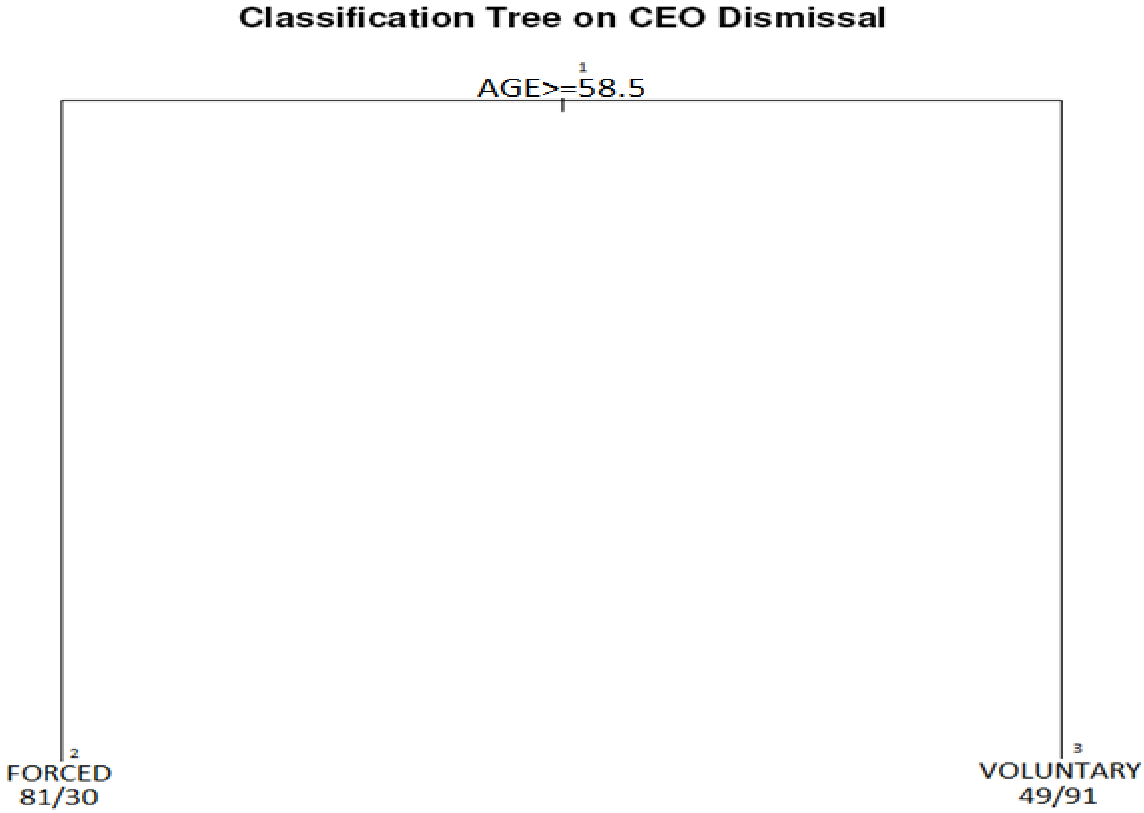
The tree that has the lowest cross-validation error is tree number 2 with one split and two terminal nodes. This is a very small and simple tree, and even though it is the best classifier it may not bring the biggest clarity on how different variables play a part in the dismissal of CEOs. I will address this further later on.

Again to choose the right sized tree we use the 1 SE rule which in this case gives us tree number 2 being the simples tree with an unadjusted cross-validation estimate $0.67769 \leq 0.67769 + 0.061408$.

Another notable feature that the revised learning sample brings is that there are bigger "jumps" in the sizes of the best pruned trees. This is due to that pruning cut of the weakest-link branch. For this tree the branches from node 2 or 3 must have been fairly weak as they were the weakest-links for several different sized trees.

The Right Sized Tree

Let us first have a look at the trees misclassification estimates. The resubstitution estimate is 0.3147 and the cross-validation estimate is 0.3267. The misclassification estimates are considerably lower than when merger cases were included. Instead of misclassifying 40 to 46 % of the cases run through the tree, the misclassification for the new tree is now less than a third of the cases run through the tree. Let us now look at the tree:



As the tree only contains two terminal nodes it only has to use one split, and the split purifying the tree the most uses the variable AGE. By comparing to the other tree we note that it is the same age that is chosen for the splitting. Just as before it turns out that if the CEO is above 58.5 years old he is likely to be dismissed. A younger CEO tends to leave his position voluntarily.

Composition of the Nodes

Although this tree is a lot less complicated, looking at the nodes may still prove helpful in deciding how good that one split is.

Table 4: Composition of the nodes

Node	n-cases	n-wrong	Class	% forced	% voluntary
1)	251	121	Forced	51,79	48.21
2)*	111	30	Forced	72.97	27.03
3)*	140	49	Voluntary	35.00	65.00

As stated in the table, node 2 classifies 73 % of the cases as forced, so forced dismissals clearly dominates this node. Node 3 is not as pure, classifying 65 % of the cases within the node as a voluntary dismissal. Both nodes however contribute considerably to improve the purity from the root node.

Primary and Surrogate Splits

As this tree only has one split, the surrogate splits are extra important. Should the surrogate splits' goodness be a lot lower than the age variable is, the risk of misclassifying cases may increase substantially. It is therefore very important to evaluate the goodness of the surrogate and primary splits for this tree.

Node 1) Splits for tree excluding mergers

Primary splits			
AGE	< 58.5	to the right,	improve=21.355680, (11 missing)
CO_TSR	< 0.0975	to the right,	improve=12.720770, (14 missing)
IA_TRS	< -0.173	to the right,	improve= 8.563315, (16 missing)
TENURE	< 11.9	to the right,	improve= 6.780272, (2 missing)
IA_INCOM	< -1.8825	to the right,	improve= 6.717279, (82 missing)
Surrogate splits			
TENURE	< 7.95	to the right,	agree=0.671, adj=0.262, (10 split)
COUNTRY	splits as	LRRRLLLRLLRRRRRR,	agree=0.667, adj=0.252, (1 split)
MKTVAL2	< 6217.992	to the right,	agree=0.608, adj=0.121, (0 split)
countg	< 0.5	to the left,	agree=0.588, adj=0.075, (0 split)
indcat	< 0.5	to the left,	agree=0.583, adj=0.065, (0 split)

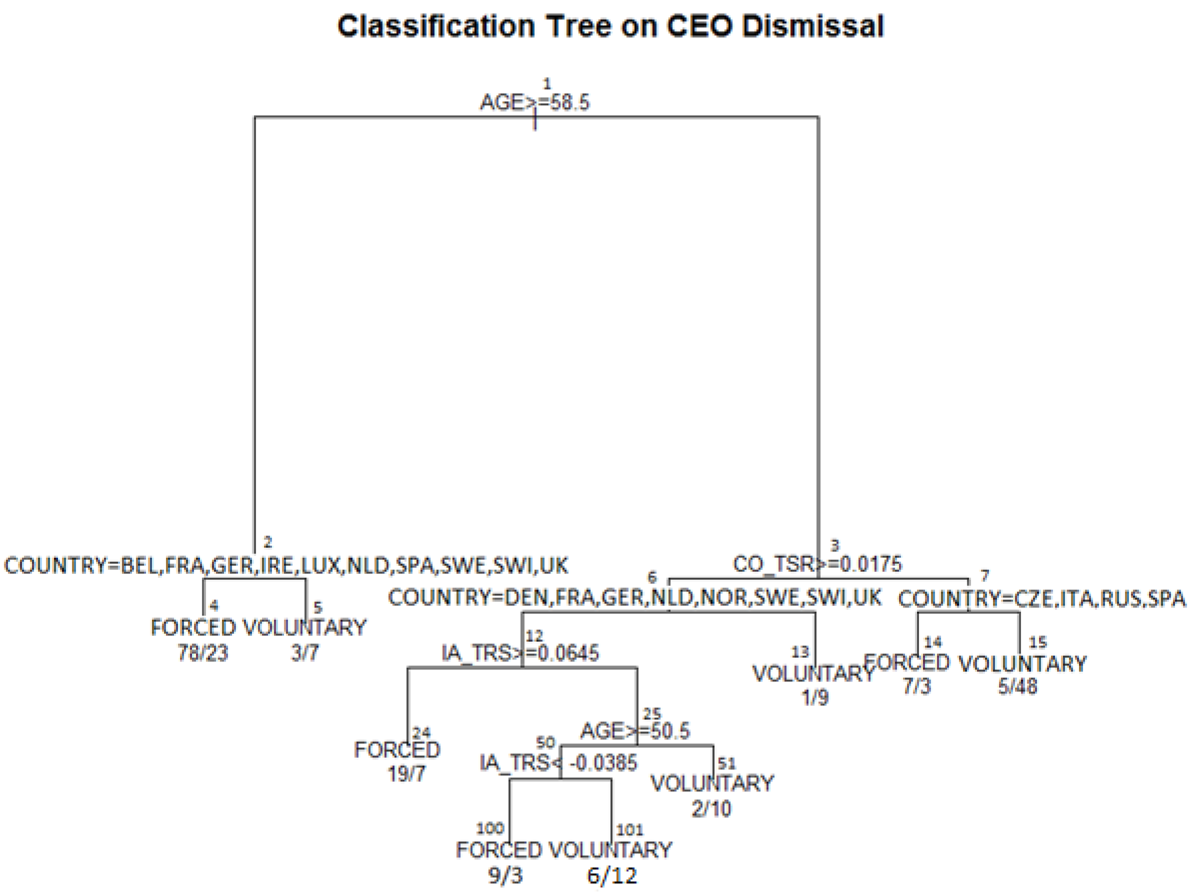
We can see that age is clearly a better discriminator than the other variables. The drop in the improve measure is more than a third down to CO_TSR as the next primary split, and it continues to drop further to the other primary splits. The best surrogate split tenure agrees 67 % of the times with age. This is not disastrous, but as this is the only split it would have been preferred had it been higher. The next surrogate split is almost as good which is a good thing should there be a case missing observations for several variables.

From this one split the impurity index is decreased by 0.0851, which is clearly higher than when the merger cases were included. This shows that excluding the merger cases removes a lot of the impurity and misclassification from the tree. The number of missing cases in the learning sample for the age variable is now 11, which is not too high. This is a good thing as the next primary and surrogate splits were not optimal. If both information regarding tenure of the CEO and nationality of the firm should be missing, MKTVAL2 only agrees with age in 61 % of the cases and that is starting to get low.

Growing the Second Best Sized Tree

As mentioned under the right sized tree section, I stated that the one node tree does not contribute to the understanding of how more of the variables can determine the nature of a dismissal. I therefore choose to include the tree with the second lowest cross-validation estimate, tree 4 in table. It is the second best sized tree both in regard to cross-validation estimate and the 1 SE rule. With this tree I hope to be able to map how other variables also can contribute to the classification of CEO dismissal cases, still bearing in mind that age is the most significant variable.

This tree has as stated before a higher cross-validation estimate at 0.3506 while the resubstitution estimate has decreased to 0.2072. The resubstitution estimate had a large decrease, but keep in mind that this estimate is biased and overly positive in comparison to what the true misclassification rate can be expected to be. We can, however, expect the true misclassification ratio to be at somewhere in between 21 % and 35 %. We note that this spread is a lot larger that for the right sized tree and extends in both ways. Let us now look at the tree in the figure below:



The eight splits in the tree use the following four variables; AGE, CO_TSR, COUNTRY and IA_TRS. There are now at least two splits and at most five splits that are involved in the classification process. Note that several of the countries that are deemed significant discriminators are observed only a few times and a generalization may not be valid. Please see page 15 for a list of countries and number of observations per country.

Old CEOs above 58.5 years are no longer automatically classified as forced. The class now depends on the nationality of the firm as well. In Belgium, France, Germany, Ireland, Luxembourg, Netherlands, Spain, Sweden, Switzerland and UK the firms tend to fire their old CEOs more than in the Czech Republic, Denmark, Finland, Italy, Portugal, Norway and Russia. The performance of a company is yet again not significant for CEOs over 58.5 years old.

For CEOs under 58.5 years old the performance of the firm plays a role. Is the change in TSR right before the succession event below 1.75 % and the firm is located in the Czech Republic, Italy, Russia or Spain, the performance will not be viewed as satisfactory and the CEO is most likely to be forced out of the firm. Is the firm located anywhere else the board tends to view the performance as sufficient and the CEO is most likely leaving the firm voluntarily.

Should the firm experience a positive change in TSR of at least 1.75 % before a dismissal, the nationality of the firm decides the further path. Firms located in Belgium, Czech Republic, Finland, Ireland, Italy, Luxembourg, Portugal, Spain and Russia will look at the performance as satisfactory and the leave of the CEO is voluntarily.

For firms located in Denmark, France, Germany, Netherland, Norway, Sweden, Switzerland and the UK, the industry adjusted TSR becomes a decisive variable in the classification process. Even if the firm performs above average of their industry it is more common to dismiss the CEO than him leaving voluntarily. Note that the CEO gets dismissed even though the firms change in TSR right before the event and the industry adjusted TSR are both positive, a firm in these countries will fire their CEO should he be younger than 58.5 years old. Should the firm in the countries above perform around the industry average or below, age again becomes the most significant variable. Should the CEO in a poor performing firm in Denmark, France, Germany, Netherland, Norway, Sweden, Switzerland or the UK be younger than 50.5 years old, he is most likely to leave the firm voluntarily.

Is the CEO older than 50.5 years old in a good performing firm compared to its industry, the extent to how bad the performance is becomes decisive. If the firm performs around average industry levels, the CEO is more likely to leave voluntarily. Is the firm experiencing a positive change in TSR right before the dismissal, but still performing poorly compared to the industry, a CEO in the age gap 50.5 – 58.5 in a firm located in Denmark, France, Germany, Netherland, Norway, Sweden, Switzerland or the UK is most likely to be forced out of the firm.

Further Analysis of Second Best Tree

As this tree is only included to shed light on how variables may be decisive in CEO dismissal and not to be used as a classifier, I will only analyze the tree briefly at the most critical places. The full list of composition of nodes and primary and secondary splits are listed in appendix C.

The composition of the terminal nodes is fairly good. The weakest nodes, 5 and 14, dominate the nodes at classification rate of 70 %. That is a lot stronger than the weakest node in the original tree excluding mergers.

Node 7) Masked variable at node 7

Primary splits			
COUNTRY	splits as	RLRLRR-LLR--RRRR,	improve=6.171968, (0 missing)
TENURE	< 0.9	to the left,	improve=6.171968, (0 missing)

In node 7, see table above, it is very interesting to look at the splits. The primary split at the node is COUNTRY with an improvement value of 6.172. The second primary split is TENURE and it also has an improvement value of 6.172. This means that TENURE and COUNTRY increases the purity equally. So even though TENURE is not directly used in the tree it is masked over by COUNTRY and it is just as significant at that node. TENURE would split so that cases with CEOs having tenure of less than 0.9 would go to the left into node 14 and be classified as forced dismissals.

Node 12) Masked variable at node 12

Primary splits			
IA_TRS	< 0.0645	to the right,	improve=3.448019, (1 missing)
MKTVAL2	< 16897.19	to the right,	improve=3.092380, (0 missing)

Node 50) Masked variable at node 50

Primary splits			
IA_TRS	< -0.0385	to the left,	improve=2.800000, (1 missing)
MKTVAL2	< 5557.015	to the right,	improve=2.440653, (0 missing)

At both nodes 12 and 50 the best variable is IA_TRS. The second primary split at both nodes is MKTVAL2 following closely behind. The improvement value for IA_TRS in node 12 is 3.448 and 3.092 for MKTVAL2. At node 50 the improvement value is 2.800 and 2.441 respectively. This indicates that the significance of MKTVAL2 is masked by IA_TRS. Small changes in the dataset may be enough to set MKTVAL2 as the most significant variable at one of the nodes. The way they discriminate the cases on the two nodes changes however. At node 12 they both send lower values to the right, while at node 50 IA_TRS sends low values to the left and MKTVAL2 sends lower values to the right.

Discussion and Conclusion

In this paper I have used classification and regression tree (CART) model to look at and evaluate CEO dismissal in 250 of the largest firms in Europe. The data was collected over the time period 1995 – 2004. CEO dismissals is an important topic as the CEO is the head of the company and having the right man may contribute to the company in a positive way, just as having the wrong man may put the firm in a disadvantageous position in relation to the competition. Contrary to most studies on CEO dismissal, merger cases were included in one part of the thesis. This was to see whether or not using CART could give some useful clarity into the topic of merger dismissals.

CART is often, and with success, used in medical research and for diagnostic purposes. To the best of my knowledge, however, there are no studies that have looked at CEO dismissals by using CART. Having studied CART over the course of this thesis I argue that it is an interesting statistical method that may look at different problems from new angles, shedding more light on how different factors and variables may be important and how they interact with each other.

When merger cases were included in the learning sample for the classification tree, the result was a tree with four terminal nodes. The misclassification ratio of the tree was estimated to be between 0.3952 and 0.4641. The only significant factor that decided a case to be a forced dismissal was that age of the CEO had to be 58.5 or more. I find this interesting as at that point many are starting to reach retirement age and one would expect there to be a high percentage of planned successions. The indication here might be that the knowledge of older CEOs, even though they may have vast experience, is starting to get outdated, and they are no longer able to compete with younger and more up to date CEOs. No performance variables were proved to have significance on forced dismissals.

For a dismissal due to merger the tree suggested that this happens most often when the age of CEO was less than 58.5, the change in total stock return right before the succession event is 1.75 % and if the company is located in Denmark, Finland, France or the UK. The fact that nationality of a firm has a decisive role may be due to different cultures and traditions as well as regulations within the countries.

Voluntary succession is triggered when the age of the CEO is less than 58.5 and if the change in TSR is less than 1.75 %. If the change in TSR were to be above 1.75 %, cases were also classified as voluntary should they be from other countries than Denmark, Finland, France and the UK. It is a bit interesting to see that if the performance of the firm is average or below, but the CEO is younger than 58.5 it tends to be a voluntary leave of the CEO and not a forced dismissal. It implies that the performance of the firm is not what is the most decisive when evaluating CEO dismissals.

To try to get a better understanding of how the relationship between forced and voluntary dismissals are, merger cases were excluded. The right sized tree in this scenario proved to be only a two node tree. This tree was estimated to have a misclassification ratio between 0.3147 and 0.3267, a much better classifier than when the merger cases were included. The only variable that decided between forced and voluntary dismissal was age. Yet again the CEOs being 58.5 years old or older tended to be forced out of their position while the younger ones left the firm voluntary.

This, however, did not give reasonable insight into the importance of several variables on CEO dismissals so I decided to include the second best tree in order to try and increase the understanding of CEO dismissals. This tree had nine terminal nodes and had 5 different variables included in the classification process. The misclassification ratio for this tree is somewhere between 0.2072 and 0.3506. That is a wider spread than the two node tree, but still better than when mergers were included. This time the classification of older CEOs also depended on the nationality of the firm. This may indicate that the culture of handling older CEOs may be different depending on the nationality of the firm. For younger CEOs the variables that were significant were the change in TSR right before a dismissal, nationality of the firm, the industry adjusted TSR, age yet again and also the tenure of the CEO. Tenure of the CEO was at a node masked by the nationality of a firm and did not show on the tree, yet it is just as significant.

A comparison can be made to the study by Jungeilges, Oxelheim and Randøy (2010) who used the same dataset, but focused only on forced and voluntary cases from 2000 up to 2004. The variables they found significant were age, country variables and US board member x return. Age was the most significant one, down to a 0.001 level, the other variables were significant on a 0.05 level. Age and the country variables were also found significant when using CART, but US board membership or any other variables suggesting an influence from the US on CEO dismissal were not in any of the best sized trees or in any of the top five primary splits at any node. Also note that other variables were not found significant. Another aspect that their regression does not show is how variables depend on the values of other variables. For instance if age is below 58.5, other variables will play a role in the classification process than if it had been above. This gives CART an extra dimension.

For further research on CEO dismissal by the use of CART, I will suggest using larger datasets that also are more complete. This may help create a better and more sophisticated tree. Also use of hypothesis' can be conducted, not just exploratory research as this study has been. CART gives a nice and clear view of how variables interact in the classification process and can help give clarity to the subject of CEO dismissal.

Appendices

Appendix A – Names of Variables

Variable explanation	Variable name in dataset
Dependent variable	REAS_CAT
Dependent variable exluding mergers	REASCAT
CEO age	AGE
CEO tenure	TENURE
Market capitalization	MKTVAL2
Change in industry adjusted Total Stock Return (TSR)	IA_TSR
Change in industry adjusted income	IA_INCOM
Change in market capitalization before an event	CO_MV
Change in TSR before an event	CO_TSR
US exchange listing	USLIST
UK exchange listing	UKLIST
US board membership	USIND
US board membership x TSR	US_IA
US exchange listing x TSR	USL_IA
Industry categories by numerical code	IND_CAT
Industry group 1	id1
Industry group 2	id2
Industry group 3	id3
Industry group 4	Id4
Industry group 5	sd5
Numerical values of industry groups	indcat
Nationality of the firm in letters	COUNTRY
Country group 1	cg1
Country group 2	cg2
Country group 3	cg3
Country group 4	cg4
Numerical values of country groups	countg

Appendix B – Example on Commands to Grow a Tree in R

To import a dataset used in Sata the package 'foreign' must be installed. After 'foreign' is installed it is easy to import the data:

```
Library(foreign)
```

```
ceotree<-read.dta("C:/Sven/Skole/Master/ceo_tree.dta")
```

This was the location of my file, your file should lie in another folder. Also ceo_tree is the name of the file used in Stata. May vary if you have renamed it.

When the data is imported we must change the values of the dependent variable to named labels:

```
ceotree$REAS_CAT[ceotree$REAS_CAT==1]<-"FORCED"

ceotree$REAS_CAT[ceotree$REAS_CAT==2]<-"VOLUNTARY"

ceotree$REAS_CAT[ceotree$REAS_CAT==3]<-"MERGER"
```

Now the dataset is ready to create the tree, provided that it has been altered in the same way as on page 16, and we only have to install the package 'rpart'. Now load the package:

```
library(rpart)
```

The intial tree can be created like this:

```
ceotr<-rpart(REASCAT~.,ceotree,method="class",cp=-Inf,xval=10)
```

and to get an overview over the tree and it's subtrees type

```
printcp(ceotr)
```

You can now see which tree will be the right sized tree according to the 1 SE rule and you must prune the tree using a value for the complexity parameter cp. To get the right sized tree, use any cp value from the wanted tree and up to the value of the tree smaller than yours. For the first best tree created I used cp=0.015, see table 1 on page 29 for the cp values there.

```
ceotrxbest<-prune(ceotr,cp=0.015)
```

To view it straight away with text type:

```
Plot(ceotrxbest,main="Classification Tree on CEO
Dismissal",uniform=T,compress=T);text(ceotrxbest,use.n=T,cex=.8)
```

To see the composition of the nodes, type:

```
ceotrxbest
```

For all primary and surrogate splits type:

```
summary(ceotrxbest)
```

Appendix C – Splits and Nodes for Second Best Sized Tree excl. Mergers

Composition of the Nodes

Node	n-cases	n-wrong	Class	% forced	% voluntary
1)	251	121	FORCED	51.79	48.21
2)	111	30	FORCED	72.97	27.03
4)*	101	23	FORCED	77.23	22.77
5)*	10	3	VOLUNTARY	30.00	70.
3)	140	49	VOLUNTARY	35.00	65.00
6)	77	37	VOLUNTARY	48.05	51.95
12)	67	31	FORCED	53.73	46.27
24)*	26	7	FORCED	73.08	26.92
25)	41	17	VOLUNTARY	41.46	58.54
50)	29	14	FORCED	51.72	48.28
100)*	11	2	FORCED	81.82	18.18
101)*	18	6	VOLUNTARY	33.33	66.67
51)*	12	2	VOLUNTARY	16.67	83.33
13)*	10	1	VOLUNTARY	10.00	90.00
7)	63	12	VOLUNTARY	19.05	80.95
14)*	10	3	FORCED	70.00	30.00
15)*	53	5	VOLUNTARY	9.43	90.57

Primary and Surrogate Splits

Node 1)

Primary splits			
AGE	< 58.5	to the right,	improve=21.355680, (11 missing)
CO_TSR	< 0.0975	to the right,	improve=12.720770, (14 missing)
IA_TRS	< -0.173	to the right,	improve= 8.563315, (16 missing)
TENURE	< 11.9	to the right,	improve= 6.780272, (2 missing)
IA_INCOM	< -1.8825	to the right,	improve= 6.717279, (82 missing)
Surrogate splits			
TENURE	< 7.95	to the right,	agree=0.671, adj=0.262, (10 split)
COUNTRY	splits as	LRRRLLLRLLRRRRRR,	agree=0.667, adj=0.252, (1 split)
MKTVAL2	< 6217.992	to the right,	agree=0.608, adj=0.121, (0 split)
countg	< 0.5	to the left,	agree=0.588, adj=0.075, (0 split)
indcat	< 0.5	to the left,	agree=0.583, adj=0.065, (0 split)

Node 2)

Primary splits			
COUNTRY	splits as	LR--LLLRLLR-LLLL,	improve=4.059031, (0 missing)
CO_TSR	< 0.065	to the right,	improve=3.692803, (4 missing)
IA_TRS	< -0.0125	to the right,	improve=2.763805, (5 missing)
IND_CAT	< 17.5	to the left,	improve=2.151131, (0 missing)
MKTVAL2	< 24014.85	to the left,	improve=1.839743, (2 missing)
Surrogate splits			
cg3	< 0.5	to the left,	agree=0.964, adj=0.6, (0 split)
countg	< 2.5	to the left,	agree=0.928, adj=0.2, (0 split)

Node 3)

Primary splits			
CO_TSR	< 0.0175	to the right,	improve=8.330873, (10 missing)
TENURE	< 0.95	to the left,	improve=5.037371, (0 missing)
IA_TRS	< -0.0835	to the right,	improve=4.052262, (11 missing)
AGE	< 48.5	to the right,	improve=3.314301, (7 missing)
COUNTRY	splits as	RLLLRL-LLLLRRLLL,	improve=2.986290, (0 missing)
Surrogate splits			
IA_TRS	< -0.0955	to the right,	agree=0.831, adj=0.607, (0 split)
TENURE	< 2.35	to the right,	agree=0.715, adj=0.339, (10 split)
COUNTRY	splits as	RLLRRL-R-LLLLLRL,	agree=0.646, adj=0.179, (0 split)
MKTVAL2	< 1720.617	to the right,	agree=0.638, adj=0.161, (0 split)
AGE	< 44.5	to the right,	agree=0.623, adj=0.125, (0 split)

Node 6)

Primary splits			
COUNTRY	splits as	-RL-LL-R-LLRRLLL,	improve=3.328126, (0 missing)
AGE	< 48.5	to the right,	improve=2.569204, (4 missing)
cg3	< 0.5	to the left,	improve=2.256776, (0 missing)
IA_INCOM	< -0.5995	to the right,	improve=2.175000, (19 missing)
TENURE	< 5.2	to the right,	improve=1.832035, (0 missing)
Surrogate splits			
cg3	< 0.5	to the left,	agree=0.974, adj=0.8, (0 split)

Node 7)

Primary splits			
COUNTRY	splits as	RLRLRR-LLR--RRRR,	improve=6.171968, (0 missing)
TENURE	< 0.9	to the left,	improve=6.171968, (0 missing)
IND_CAT	< 37.5	to the right,	improve=2.145827, (0 missing)
indcat	< 1.5	to the left,	improve=1.793789, (0 missing)
countg	< 1.5	to the right,	improve=1.443179, (1 missing)
Surrogate splits			
TENURE	< 0.35	to the left,	agree=0.889, adj=0.3, (0 split)
cg3	< 0.5	to the right,	agree=0.889, adj=0.3, (0 split)
countg	< 2.5	to the right,	agree=0.889, adj=0.3, (0 split)

Node 12)

Primary splits			
IA_TRS	< 0.0645	to the right,	improve=3.448019, (1 missing)
MKTVAL2	< 16897.19	to the right,	improve=3.092380, (0 missing)
IA_INCOM	< -0.5995	to the right,	improve=1.830476, (17 missing)
CO_TSR	< 0.108	to the right,	improve=1.819264, (1 missing)
AGE	< 48.5	to the right,	improve=1.803093, (3 missing)
Surrogate splits			
CO_TSR	< 0.2105	to the right,	agree=0.712, adj=0.269, (0 split)
MKTVAL2	< 40718.98	to the right,	agree=0.652, adj=0.115, (1 split)
COUNTRY	splits as	--L-LR---RR--RRR,	agree=0.636, adj=0.077, (0 split)
IND_CAT	< 42.5	to the right,	agree=0.636, adj=0.077, (0 split)
TENURE	< 20	to the right,	agree=0.636, adj=0.077, (0 split)

Node 25)

Primary splits			
AGE	< 50.5	to the right,	improve=2.643357, (2 missing)
MKTVAL2	< 9285.2	to the right,	improve=2.154052, (0 missing)
TENURE	< 4.9	to the right,	improve=1.706672, (0 missing)
CO_MV	< 0.132	to the left,	improve=1.542857, (11 missing)
COUNTRY	splits as	-----R---LL--RLL,	improve=1.246977, (0 missing)
Surrogate splits			
IA_TRS	< -0.1375	to the right,	agree=0.744, adj=0.091, (2 split)
id1	< 0.5	to the left,	agree=0.744, adj=0.091, (0 split)

Node 50)

Primary splits			
IA_TRS	< -0.0385	to the left,	improve=2.800000, (1 missing)
MKTVAL2	< 5557.015	to the right,	improve=2.440653, (0 missing)
IA_INCOM	< 0.064	to the left,	improve=1.937729, (8 missing)
COUNTRY	splits as	-----R---RR--RLL,	improve=1.577997, (0 missing)
TENURE	< 6.75	to the left,	improve=1.384719, (0 missing)
Surrogate splits			
IND_CAT	< 27.5	to the left,	agree=0.750, adj=0.3, (1 split)
CO_TSR	< 0.055	to the left,	agree=0.750, adj=0.3, (0 split)
MKTVAL2	< 1526.878	to the left,	agree=0.714, adj=0.2, (0 split)
COUNTRY	splits as	-----R---RL--RRR,	agree=0.679, adj=0.1, (0 split)
TENURE	< 2.25	to the left,	agree=0.679, adj=0.1, (0 split)

References

Alister, H., E. M. Susan, et al. (1997). Earnings Volatility, Earnings Management, and Equity Value.

Allgood, S. and K. A. Farrell (2000). "THE EFFECT OF CEO TENURE ON THE RELATION BETWEEN FIRM PERFORMANCE AND TURNOVER." Journal of Financial Research **23**(3): 373-390.

Armstrong, J. S. and J. G. Andress (1970). "Exploratory Analysis of Marketing Data: Trees vs. Regression." Journal of Marketing Research **7**(4): 487-492.

Boyd, B. K. (1994). BOARD CONTROL AND CEO COMPENSATION, John Wiley & Sons, Inc. **15:** 335-335.

Brickley, J. A., J. S. Linck, et al. (1999). "What happens to CEOs after they retire? New evidence on career concerns, horizon problems, and CEO incentives." Journal of Financial Economics **52**(3): 341-377.

David, J. D., K. D. Diane, et al. (1997). "Ownership structure and top executive turnover." Journal of Financial Economics **45**: 193-221.

Fama, E. F. (1980). "Agency Problems and the Theory of the Firm." Journal of Political Economy **88**(2): 288-307.

Fama, E. F. and M. C. Jensen (1983). "SEPARATION OF OWNERSHIP AND CONTROL." Journal of Law & Economics **26**(2): 301-326.

Fredrickson, J. W., D. C. Hambrick, et al. (1988). "A Model of CEO Dismissal." The Academy of Management Review **13**(2): 255-270.

Grusky, O. (1961). "Corporate Size, Bureaucratization, and Managerial Succession." American Journal of Sociology **67**(3): 261-269.

Guan, L., C. J. Wright, et al. (2005). "Earnings Management and Forced CEO Dismissal." Advances in Accounting **21**(0): 61-81.

Huson, M. R. M., Paul H.; Parrino, Robert (2004). "Managerial succession and firm performance." Journal of Financial Economics **74**: 237-275.

Huson, M. R. P., Robert; Starks, Laura T. (2001). "Internal Monitoring Mechanisms and CEO Turnover: A Long-Term Perspective." Journal of Finance **56**: 2265-2297.

Jensen, M. C. and W. H. Meckling (1976). "THEORY OF THE FIRM: MANAGERIAL BEHAVIOR, AGENCY COSTS AND OWNERSHIP STRUCTURE." Journal of Financial Economics **3**(4): 305-360.

Jensen, M. C. and K. J. Murphy (1990). "Performance Pay and Top-Management Incentives." Journal of Political Economy **98**(2): 225-264.

Jungeilges, J. A. O., Lars; Randøy, Trond (2010). "Globalization of Corporate Governance: The American Influence on Dismissal Performance Sensitivity of European CEOs." http://ssrn.com/abstract=1554406 **IFN Working Paper No. 823**

Karaevli, A. (2007). "Performance consequences of new CEO 'Outsiderness': Moderating effects of pre- and post-succession contexts." Strategic Management Journal **28**(7): 681-706.

L. Breiman, J. H. F., R. A. Olshen, C. J. Stone (1984). Classification And Regression Trees. Belmont, California, Wadsworth International Group.

Leker, J. S., Sören (2000). "CEO turnover and corporate performance." Scandinavian Journal of Management **16**: 287-303.

Morgan, J. N. and R. C. Messenger (1973). THAID: a sequential analysis program for the analysis of nominalscale dependent variables. Ann Arbor, Mich., Survey Research Center, Institute for Social Research, The University of Michigan.

Morgan, J. N. and J. A. Sonquist (1963). "Problems in the Analysis of Survey Data, and a Proposal." Journal of the American Statistical Association **58**(302): 415-434.

Oxelheim, L. and T. Randoy (2005). "The Anglo-American Financial Influence on CEO Compensation in Non-Anglo-American Firms." Journal of International Business Studies **36**(4): 470-483.

Penman, S. H. (2010). Financial statement analysis and security valuation. Boston, Mass., McGraw-Hill.

Pourciau, S. (1993). "Earnings management and nonroutine executive changes." Journal of Accounting and Economics **16**(1–3): 317-336.

Reitenga, A. L. and M. G. Tearney (2003). "Mandatory CEO Retirements, Discretionary Accruals, and Corporate Governance Mechanisms." Journal of Accounting, Auditing & Finance **18**(2): 255-280.

Shen, W. and A. A. Cannella Jr (2002). "POWER DYNAMICS WITHIN TOP MANAGEMENT AND THEIR IMPACTS ON CEO DISMISSAL FOLLOWED BY INSIDE SUCCESSION." Academy of Management Journal **45**(6): 1195-1206.

Stone, C. J. (1977). "Consistent Nonparametric Regression." The Annals of Statistics **5**(4): 595-620.

Zajac, E. J. and J. D. Westphal (1996). "Director Reputation, CEO-Board Power, and the Dynamics of Board Interlocks." Administrative Science Quarterly **41**(3): 507-529.

Zhang, Y. (2008). "Information Asymmetry and the Dismissal of Newly Appointed CEOs: An Empirical Investigation." Strategic Management Journal **29**(8): 859-872.