INTELLIGENT SYSTEMS IN ACCOUNTING, FINANCE AND MANAGEMENT
Intell. Sys. Acc. Fin. Mgmt. **20,** 111–139 (2013)
Published online 17 April 2013 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/isaf.1338

CORPORATE DIVIDEND POLICY DETERMINANTS: INTELLIGENT VERSUS A TRADITIONAL APPROACH

PANTELIS LONGINIDIS[a*] AND PANAGIOTIS SYMEONIDIS[b]

[a] *Department of Engineering Informatics & Telecommunications, University of Western Macedonia, Karamanli & Lygeris Street, 50100 Kozani, Greece*

[b] *Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece*

SUMMARY

Dividend is the return that an investor receives when purchasing a company's shares. The decision to pay these dividends to shareholders concerns several other groups of people, such as financial managers, consulting firms, individual and institutional investors, government and monitoring authorities, and creditors, just to name a few. The prediction and modelling of this decision has received a significant amount of attention in the corporate finance literature. However, the methods used to study the aforementioned question are limited to the logistic regression method without any implementation of the advanced and expert methods of data mining. These methods have proven their superiority in other business-related fields, such as marketing, production, accounting and auditing. In finance, bankruptcy prediction has the vast majority among data-mining implementations, but to the best of the authors' knowledge such an implementation does not exist in dividend payment prediction. This paper satisfies this gap in the literature and provides answers that help to understand the so-called 'dividend puzzle'. Specifically, this paper provides evidence supporting the hypothesis that data-mining methods perform better in accuracy measures against the traditional methods used. The prediction of dividend policy determinants provides valuable benefits to all related parties, as they can manage, invest, consult and monitor the dividend policy in a more effective way. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: dividend policy; data mining; decision tree; neural network; logistic regression; Athens stock exchange

1. INTRODUCTION

The primary goal of financial management (FM) is to maximize the current value per share of the existing stock. One substantial financial decision affecting this value maximization goal is the dividend policy (DP). According to Baker (2009), dividend decisions, as determined by a firm's DP, affect the amount of earnings that a firm distributes to shareholders versus the amount it retains and reinvests. DP refers to the payout policy that a firm follows in determining the size and pattern of cash distributions to shareholders over time.

Corporate DP has captured the interest of economists since the middle of the twentieth century and over the last six decades has been the subject of intensive theoretical modelling and empirical examination. A number of conflicting theoretical models, which are lacking in strong empirical support, define current attempts to explain the corporate dividend behaviour (Frankfurter and Wood, 2002). Brealey *et al*. (2004) described eloquently the reason for this conflict in the DP modelling landscape. The authors stated that the endearing feature of economics, where it can

* Correspondence to: Pantelis Longinidis, Department of Engineering Informatics & Telecommunications, University of Western Macedonia, Karamanli & Lygeris Street, 50100 Kozani, Greece. E-mail: logggas@gmail.com

always accommodate not just two but three opposing points of view, is applicable and induces the controversy about DP. On the one side there is a group which believes that an increase in dividend payout increases firm value. On the opposite side there is a group which believes that an increase in payout reduces value. And in the centre there is a party which claims that DP makes no difference. Black (1976) characterized this controversy as a puzzle by arguing that the harder we look at the dividend picture, the more it seems like a puzzle with pieces that just do not fit together.

Despite exhaustive theoretical and empirical analysis to explain their pervasive presence, dividends remain one of the thorniest puzzles in corporate finance (Baker *et al*., 2002). The inability to resolve the dividend puzzle is mainly due to financial economists' efforts to develop universal models, although it is proven that DP is sensitive to factors such as market frictions, firm characteristics, corporate governance and legal environments (Frankfurter and Wood, 1997; Baker *et al*., 2008).

Frankfurter and Wood (2002) conducted an extensive literature review in order to explore whether the puzzling reality of corporate dividend behaviour is caused by three factors; namely: (1) method of analysis employed; (2) sample period; and (3) data frequency. The authors analysed 150 empirical studies of corporate DP and came to the conclusion that no dividend model, either separately or jointly with other models, is supported invariably. However, a semantic part of Frankfurter and Wood's (2002) analysis is the presentation of the methods utilized for each model. The vast majority of these models utilized regression and event studies methods, and the synchronous methods of data mining (DM) were implemented by none of these models.

The DP decision, and more specifically the decision to pay or not dividends, can be regarded as a typical binary classification problem of assigning new observations to two predefined decisions as classes (e.g. 'yes' and 'no' dividend payment classes). Despite the fact that many DP methods have been applied in the financial area, an analogous model in the DP literature does not exist to the best of our knowledge.

This gap in the existing DP literature has stimulated the research interest of this work as it desires to fulfil the need to employ DM methods in order to model the decision to pay or not dividends and to explore whether these techniques are capable of predicting the dividend payment decision better and more precisely than the traditional regression approaches found scattered in the DP literature. However, by modelling the DP decision, this study aims further to provide a convenient and effective decision-support tool to investors. Investors will understand the financial and nonfinancial features that paying and nonpaying companies have and will take them into account when constructing and managing their investment portfolios. Our research effort is summarized in the following research questions:

RQ1. Are DM methods more accurate than the logistic regression in predicting the dividend payment decision of corporations?

RQ2. Which are the financial, managerial, and corporate governance features of corporations paying and not paying dividends to shareholders?

The rest of the paper is structured as follows. Section 2 reviews the previous and current body of literature for both DP and DM in the FM area. Section 3 provides insights into the research methodology employed, followed by the dataset generation process in Section 4. The available DM techniques are applied using this dataset and the results reported and commented on in Section 5. The paper ends with concluding remarks, managerial implications and further research directions.

DOI: 10.1002/isaf

2. RELATED LITERATURE

2.1. Dividend and Dividend Policy Determinants

The seminal papers of Linter (1956) and of Miller and Modigliani (1961) (MM) were the beginning of contemporary theoretical attempts to explain the role of DP. Since these pioneer works, the bulk of studies followed and either support or reject their validity. As Baker (2009) states, MM's unconventional and controversial conclusion about DP irrelevance stirred a heated debate that has reverberated throughout the finance community for decades. The DP theories developed are the following:

- The dividend irrelevance theory, where dividend payout policy does not affect overall firm value in perfect capital markets (Ang and Ciccone, 2009).
- The residual DP theory, where managers exhaust all available positive net present value investments and then pay the residual cash flow as dividend (Smith, 2009).
- The tax clientele effects theory, where investors prefer firms to retain cash instead of pay dividends because the tax rate on dividends is typically higher than on long-term capital gains (Saadi and Dutta, 2009).
- The cash flow signalling hypothesis, where the stock price moves in the same direction as the dividend because dividend changes convey information about the firm's future growth opportunities (Mukherjee, 2009).
- The free cash flow hypothesis, where price reacts favourably to the announcement of a dividend increase because this increase reduces the agency cost of free cash flow (funds available to managers for perquisite consumption) (Mukherjee, 2009).
- The signalling theory, where unexpected dividend increases (decreases) are associated with significant share-price increases (decreases) because dividend changes signal future prospects of the firm and thus reduce the information asymmetries existing between firm managers and the market (Filbeck, 2009).
- The firm life cycle theory, where the optimal DP of a firm is based on its life cycle. A firm will begin paying dividends when its growth rate and profitability are expected to decline in the future (Bulan and Subramanian, 2009).
- The catering theory, where managers cater to investor demand by paying dividends when investors prefer dividend-paying firms and by not paying dividends (or reducing the dividend) when investors prefer non-dividend-paying companies (De Rooij and Renneboog, 2009).
- The behavioural theory explains the impact of age, retirement status and income on the relationship between consumer expenditures and the preference for dividends, and a psychological approach to dividend theory explains the relationship between tolerance for risk and the preference for dividends (Shefrin, 2009).

Many studies across various countries and time periods investigated the validity of these theories. Denis and Stepanyan (2009) provided a synthesis of studies focused on DP determinants and concluded that dividends are associated with several firm characteristics, such as size, profitability, growth opportunities, maturity, leverage, equity ownership and incentive compensation. Additionally, the authors found an association between dividends and characteristics of the market in which the firm operates, such as tax laws, investor protection, product market competition, investor sentiment and public or private status, as well as the availability of substitute forms of corporate payout, primarily share repurchases. Dutta and Saadi (2009) discussed different external (such as shareholder rights and legal

DOI: 10.1002/isaf

environment) and internal corporate governance mechanisms (such as managerial and block-holder ownership, compensation and board structure) that may influence a firm's DP. The authors reported a significant impact of these factors on DP and that the majority of studies showed that better legal protection of minority shareholders led to a higher level of dividend payments.

Table I summarizes representative studies that highlight the state of knowledge in the field of DP determinants. The findings of these studies provide support to all theories, and this contributes to the 'dividend puzzle' as no model is able to express the DP invariably and under all circumstances. Researchers investigate a large number of financial, managerial and corporate governance features of firms aiming to evaluate their research hypothesis and employ a variety of methods and techniques. Logistic regression was the most popular technique employed in DP determinants studies; more importantly, DM methods were lacking.

2.2. Data Mining Applications in Financial Management

DM can be applied to many different economic and/or financial prediction problems (Seng and Chen, 2010). Statistical analysis has been available to businesses for years, but somehow DM has captured the interest of businesses in a way that classical statistical analysis never did. The main reason for this widespread popularity is the real financial benefit to businesses (Jessen and Paliouras, 2001; Lee and Siau, 2001; Bose, 2009). DM in FM is an emerging field with potential benefits for both academics and practitioners.

Forecasting stock market, currency exchange rate, bankruptcies, understanding and managing financial risk, trading futures, credit rating, loan management, bank customer profiling and money laundering analyses are core financial tasks for DM (Kovalerchuk and Vityaev, 2005; Tsai, 2008; Huang *et al*., 2012). Wong and Selvi (1998) examined the historical trend of published finance applications of neural networks (NNs). Their survey indicated that only a few NNs were developed for supporting the strategic planning of decision making in finance. Kirkos and Manolopoulos (2004) conducted an excellent review of the DM applications in finance and accounting and concluded that the most popular DM method in finance was NNs and the most popular finance task was bankruptcy prediction. Zhang and Zhou (2004) highlighted the potential of DM techniques in finance and review application studies existing in core financial areas. Financial fraud detection with application of DM techniques was the topic that Ngai *et al*. (2011) investigated in their review article and concluded that insurance fraud was the most popular topic and logistic models were the most widely used. Sharma and Panigrahi (2012) reviewed DM applications on detection of financial accounting fraud and found that logistic models are again the most popular in application. The literature of financial crisis prediction with machine-learning applications was surveyed by Lin *et al*. (2012), where they came to the conclusion that the development of models in this area has a long way to go.

Based on the above-cited reviews, it is evident that the field of DM in finance is growing rapidly in depth and width. However, there are some finance areas that research has not yet directed its interest onto, and DP is among these. The only endeavours that applied DM in the DP field are two studies by Kim and co-workers.

In their first study, Kim *et al*. (2010) developed a dividend forecasting model that outperformed the popular (in the finance community) Marsh and Merton (1987) dividend prediction model – an econometric error correction model that utilized only past stock price (P_t and P_{t-1}) and past dividends (D_t) in order to predict future dividends – in accuracy measures under different tolerance levels. The model of Kim *et al*. (2010) was based on the concept of knowledge integration (KI), where the rules derived by implementing the classification and regression trees (CART) algorithm in four different

DOI: 10.1002/isaf

Table I Empirical studies on DP determinants

Reference	Data	Method	DP determinants
Michel (1979)	168 firms listed in *Moody's Handbook of Industrials* over the period 1967–1976	*F*-test, Kruskal–Wallis test	Industry
Baker *et al.* (1985)	318 firms listed in New York Stock Exchange in 1983	Survey, chi-square test	Future earnings, pattern of past dividends, cash availability, stock price, industry
Schellenger *et al.* (1989)	526 firms listed in COMPUSTAT database in 1986	Pearson correlation	BoD composition
Alli *et al.* (1993)	105 firms listed in COMPUSTAT database in 1985	Factor analysis	Issuance cost, pecking order, investment, financial slack, dividend stability, tax and agency costs, capital structure flexibility
Agrawal and Jayaraman (1994)	71 firms with no long-term debt and 71 matched firms with debt all listed in COMPUSTAT database in1981	*t*-test, Wilcoxon test, linear regression	Debt, managerial ownership
Barclay *et al.* (1995)	6780 firms covered by COMPUSTAT database over the period 1963–1993	Tobit regression	Investment opportunities
Holder *et al.* (1998)	477 firms listed in COMPUSTAT database over the period 1983–1990	Linear regression	Corporate focus, size, insider ownership, number of shareholders, free cash flow
Chen and Steiner (1999)	785 firms listed in New York Stock Exchange in 1994 that had complete ownership data on proxy statements	Nonlinear two-stage regression	Managerial ownership
Baker *et al.* (2001)	188 US firms traded on Nasdaq that paid dividends over the period 1996–1997	Survey, chi-square test	Pattern of past dividends, stability of earnings, level of current and future earnings
Fama and French (2001)	All firms listed in New York Stock Exchange, in American Stock Exchange and in Nasdaq over the period 1926–1999	Descriptive statistics, logistic regression	Profitability, investment opportunities, size
Ooi (2001)	44 property firms quoted on London Stock Exchange that paid dividends over the period 1986–1998 (528 firm-year data)	Linear regression	Size, asset and capital structure
Dickens *et al.* (2002)	677 US banking firms listed in Morningstar Principia Pro over the period 1998–2000	Tobit regression	Investment opportunities, size, agency problems, dividend history, risk
Short *et al.* (2002)	211 firms listed in London Stock Exchange over the period 1988–1992	Generalized linear regression	Institutional ownership
Omran and Pointon (2004)	94 firms included in *Kompass Egypt Financial Yearbook 1999/2000*	Linear regression	Debt, size
Chen *et al.* (2005)	412 firms listed in Stock Exchange of Hong Kong over the period 1995–1998	Linear regression	Family ownership
Amidu and Abor (2006)	22 firms listed in Ghana Stock Exchange over the period 1998–2003	Linear regression	Profitability, cash flow, tax, risk, institutional ownership, future prospect, investment opportunities
Ben Naceur *et al.* (2006)	48 firms listed in Tunisian Stock Exchange over the period 1996–2002	Generalized method of moments, pooled least square, fixed effect, random effect	Profitability, stable earnings

(*Continues*)

DOI: 10.1002/isaf

Table I. (Continued)

Reference	Data	Method	DP determinants
DeAngelo *et al*. (2006)	4363 firms listed in New York Stock Exchange, in American Stock Exchange and in Nasdaq over the period 1973–2002	Logistic regression	Earned/contributed capital mix
Mancinelli and Ozkan (2006)	139 firms listed in Milan Stock Exchange in 2001	Tobit regression, logistic regression	Voting rights by the largest shareholder, agreements among large shareholders
Denis and Osobov (2008)	All firms listed in Worldscope database over the period 1989–2002 from USA, Canada, UK, Germany, France and Japan	Logistic regression	Size, growth opportunities, profitability, earned/contributed capital mix
Ahmed and Javid (2009)	320 nonfinancial firms listed in Karachi Stock Exchange over the period 2001–2006	Generalized method of moments, pooled least square, fixed effect, random effect	Profitability, stable earnings, inside share holdings, market liquidity
Chen and Dhiensiri (2009)	72 firms listed in the New Zealand Stock Exchange over the period 1991–1999	Linear regression	Managerial ownership, ownership concentration
Kim and Gu (2009)	69 hospitality firms traded in USA in 2005	Logistic regression	Size, profitability, investment opportunities
Al-Kuwari (2010)	191 firms listed in stock exchanges of the Gulf Cooperation Council over the period 1999–2003	Logistic regression	Government ownership, size, profitability, growth
Nam *et al*. (2010)	All firms listed in the Compustat Execucomp database that initiate dividends over the period 1993–2005	*t*-test, *z*-test, logistic regression	Managerial stock holdings
Brockman and Unlu (2011)	80,725 firm-year observations from 12,871 firms from 31 countries listed in the Compustat Global database over the period 1996–2007	Logistic regression	Retained earnings
Coulton and Ruddock (2011)	7838 firm-year observations from firms listed in Australian Stock Exchange over the period 1993–2004	Logistic regression	Maturity, size, profitability, growth options, retained earnings
Jiraporn *et al*. (2011)	9893 firm-year observations from firms listed in Institutional Shareholder Services over the period 1993–2004	Logistic regression, linear regression, two-stage linear regression	Governance quality
Renneboog and Trojanowski (2011)	985 firms listed in London Stock Exchange and in the Worldscope database over the period 1992–2004	Logistic regression	Size, profitability, leverage, investment opportunities
Shabibi and Ramesh (2011)	90 UK firms listed in the FAME database in 2007	Linear regression	Board independence, profitability, size, risk
He (2012)	35,462 firm-year observations from 2008 Japan firms listed in the Pacific-Basin Capital Markets over the period 1977–2004	Logistic regression, linear regression	Product market competition
He *et al*. (2012)	312 firms listed in Hong Kong Stock Exchange in 2007	Linear regression	Family ownership, state ownership
Manos *et al*. (2012)	8865 Indian firms listed in the PROWESS database over the period 2000–2006	Logistic regression, Tobit regression	Group affiliation

datasets – one with variable P_t missing, one with variable P_{t-1} missing, one with variable D_t missing and one without any missing variable – are combined and provide a meta model with 39 rules. Data from a sample frame of 137 companies listed in the Korea Exchange market and for a time window of 20 years, from 1980 to 1999, were used to conduct experiments aiming to compare the KI model with the Marsh and Merton model, a CART model and a back-propagation NN. The experiments showed that the proposed KI model, with its cumulating rules from missing datasets, improved prediction performance as it reduces the error term and increases R^2, and this results in an excessive overall accuracy that outperforms the other three benchmark models.

In a subsequent paper, Won *et al*. (2012) suggested a knowledge refinement model that refines the multiple rules extracted through rule-based algorithms from dividend datasets using genetic algorithms (GAs). Through a seven-step framework, their genetic algorithm knowledge refinement (GAKR) technique starts with rules induction from traditional algorithms (CHAID, CART, QUEST and C5.0) and after implementing a GA iterative process that searches for the most valuable decision rule or optimal rule set provides the DP prediction. Although the GAKR model utilized the same input variables and the same experiment data that the KI model did, its predicted target variable was not the dividend value but a binary variable – if $D_{t+1} \geq D_t$ then the class is +1 and if $D_{t+1} \leq D_t$ then the class is -1 – showing the DP. The experiments provided evidence supporting the prediction performance superiority of the GAKR model against the traditional rule induction algorithms (CHAID, CART, QUEST and C5.0), and these results were verified statistically via the nonparametric Wilcoxon signed-rank test.

3. RESEARCH METHODOLOGY

The DP decision can be modelled as a typical classification problem where the outcomes are 'pay dividends' or 'do not pay dividends'. A classification technique (or classifier) is a systematic approach to building classification models from an input dataset. Examples include decision tree (DT) classifiers, rule-based classifiers, NNs, support vector machines, and naive Bayes classifiers. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records it has never seen before. Therefore, a key objective of the learning algorithm is to build models with good generalization capability (Tan *et al*., 2006).

Among a plethora of available DM techniques, in this study we select the DT and back-propagation NN methods and compare their results with the logistic regression method. NNs are the most widely used DM method in finance applications (Zhang and Zhou, 2004), while the back-propagation learning algorithm is used most frequently in business applications (Vellido *et al*., 1999; Wong *et al*., 2000). However, NNs lack explanation facilities when applied to DM problems as their knowledge is buried in their structures and weights, making it difficult to extract rules (Li and Wang, 2004). This shortcoming is managed by utilizing DTs, whose interpretability is very high. Moreover, these two methods had a satisfactory applicability in the DP field (Kim *et al*., 2010).

3.1. Decision Trees

A DT is a collection of decision nodes, connected by branches, extending downward from the root node until terminating in leaf nodes. Beginning at the root node, which by convention is placed at the top of the DT diagram, attributes are tested at the decision nodes, with each possible outcome resulting in a branch. Each branch then leads either to another decision node or to a terminating leaf node. The

DOI: 10.1002/isaf

dataset is partitioned, or split, according to the values of this attribute (Larose, 2005). There are many measures that can be used to determine the best way to split the records and these are defined in terms of the class distribution of the records before and after splitting. The measures developed for selecting the best split are often based on the degree of impurity of the child nodes. The smaller the degree of impurity, the more skewed the class distribution is. Among others, a popular impurity measure is the entropy:

$$\text{Entropy}(t) = -\sum\nolimits_{i=0}^{c-1} p(i|t)\log_2 p(i|t) \tag{1}$$

where c is the number of classes and $p(i|t)$ denotes the fraction of records belonging to class i at a given node t. However, in order to determine how well a test condition performs, we need to compare the degree of impurity of the parent node (before splitting) with the degree of impurity of the child nodes (after splitting), where the larger the difference is, the better the test condition is. The gain is a criterion that can be used to determine the goodness of a split:

$$\Delta = I(\text{parent}) - \sum\nolimits_{j=1}^{k} \frac{N(u_j)}{N} I(u_j) \tag{2}$$

where $I(\cdots)$ is the impurity measure of a given node, N is the total number of records at the parent node, k is the number of attribute values and $N(u_j)$ is the number of records associated with the child node u_j (Tan *et al.*, 2006).

In this study, the C5.0 DT algorithm was used. C5.0 is an extension of C4.5 (Quinlan, 1993), which is the result of a series of improvements to the ID3 algorithm (Quinlan, 1986). These improvements include methods for dealing with numeric attributes, missing values, noisy data and generating rules from trees (Witten and Frank, 2005). The algorithm works by splitting the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest level splits are re-examined, and those that do not contribute significantly to the value of the model are removed or pruned (IBM Corporation, 2011).

3.2. Neural Networks

An NN is composed of a set of elementary computational units, called neurons, connected together through weighted connections. These units are organized in layers so that every neuron in a layer is exclusively connected to the neurons of the preceding layer and the subsequent layer. These layers can be of three types: input, output or hidden. The input layer receives information only from the external environment without performing any calculation and transmits information to the next level. The output layer produces the final results, which are sent by the network to the outside of the system. Between the output and the input layer there can be one or more intermediate levels, called hidden layers because they are not directly in contact with the external environment. These layers are exclusively for analysis; their function is to take the relationship between the input variables and the output variables and adapt it more closely to the data. Every neuron, also called a node, represents an autonomous computational unit and receives inputs as a series of signals that dictate its activation. Following activation, every neuron produces an output signal. All the input signals reach the neuron simultaneously, so the neuron receives more than one input signal, but it produces only one output signal. Every input signal is associated with a connection weight. The weight determines the relative

DOI: 10.1002/isaf

importance the input signal can have in producing the final impulse transmitted by the neuron. The weights are adaptive coefficients that are modified in response to the various signals that travel on the network according to a suitable learning algorithm. A threshold value, called bias, is usually introduced. A generic neuron j, with a threshold θ_j, receives n input signals $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ from the units to which it is connected in the previous layer. Each signal is attached with an importance weight $\mathbf{w}_j = [w_{1j}, w_{2j}, \ldots, w_{nj}]$. The same neuron elaborates the input signals, their importance weights and the threshold value through a combination function. The combination function produces a value called the potential or net input. An activation function transforms the potential into an output signal. The combination function is usually linear; therefore, the potential is a weighted sum of the input values multiplied by the weights of the respective connections. The sum is compared with the threshold value. The potential and the output signal of a neuron j is defined by the linear combination shown in equations (3) and (4) (Giudici and Figini, 2009):

$$\mathrm{IN}_j = \sum\nolimits_{i=1}^{n} \left(x_i w_{ij} - \theta_j \right) \tag{3}$$

$$\mathrm{OUT}_j = f\left[\sum\nolimits_{i=1}^{n} \left(x_i w_{ij} - \theta_j \right) \right] \tag{4}$$

The combination of topology, learning paradigm and learning algorithm defines an NN model. There is a wide selection of popular NN models. For DM, perhaps the back-propagation network and the Kohonen feature map are the most popular (Bigus, 1996). In this study, a feed-forward back-propagation NN with exhaustive prune training method was used.

4. DATASET GENERATION AND PROFILE

4.1. Population Frame, Time Frame and Data Sources

In this study, the companies listed in the Athens Exchanges (ATHEX) were our initial population frame and the reason for this selection is at least twofold. First, the institutional framework that regulates the operation of these companies is very strict and is governed by transparency and disclosure. The ATHEX rulebook defines explicitly the regular or periodic reporting obligations, the extraordinary reporting obligations and the special categories of reporting. All information that might have an effect on a company's stock price, such as resolutions of the general meeting, payment of main dividends/interim dividends, corporate actions, and so on should be announced to both the authorities and investors. Second, the financial statements prepared by issuers should be in accordance with legislation in force and should be audited by a certified auditor. Consequently, the reliability and validity of these statements is high because the certified auditor is charged with both civil and criminal liabilities in the case of false and misleading financial statements. Thus, the quantity and quality of available information were the dataset's selection criteria. The data that constitute the research dataset were collected through different resources and during a time frame of approximately 8 months, starting from 1 September 2010. September is the month that almost all companies listed in the ATHEX have conducted their annual ordinary shareholders' meetings. During these meetings the board of directors

(BoD) suggests or does not suggest a dividend payment, for the previous fiscal year, and the body of shareholders accepts or rejects this decision based on the majority vote.

Pure financial data, such as financial ratios and financial statement accounts, were drawn from iMENTOR, an online business information platform featured and updated by Hellastat S.A. (http://www.hellastat.com). Hellastat is a company that operates in the areas of business information and market research. Moreover, Hellastat is a strategic partner of Standard & Poor's and a member of Thomson–Reuters plc. Data related to companies' profiles and announcements, under the rulebook and resolutions of the ATHEX, were drawn from the official web site of the ATHEX S.A. (http://www.ase.gr/default_en.asp). Several data triangulations were made with daily and periodic business press releases and with data available at web sites of business information vendors. Time-series data, such as stock closing price and trading volume for a large time horizon and with a daily frequency, were drawn from the official site of Naftemporiki Publishing S.A. (http://www.naftemporiki.gr), a leading company in economic and business press in Greece. Finally, specialized data regarding the corporate governance of the companies, such as composition of the BoD and nationality of subsidiaries, were drawn from the annual reports and annual financial reports of the companies. According to the ATHEX rulebook, all listed companies are obliged to post on their official web sites the aforementioned reports; thus, these data were gathered from the companies' official web sites.

The time frame included three consecutive years from 2007 to 2009. On 1 September 2010 the total number of ATHEX-listed companies was 280. Various companies had been listed and delisted from the ATHEX during the 3 years prior to this date, but in order to have recent and non-missing data the 280 companies listed on 1 September 2010 constituted the initial research sample. The first screening process on the initial research sample concerned the elimination of companies which had specific features on their financial statements making impossible any comparison with the other listed companies. These were banks and insurance companies which use different generally accepted accounting principles. A total of 19 companies were excluded, 15 banks and 4 insurance companies, resulting in a research sample of 261 companies. The second screening process concerned the elimination of 15 companies whose stocks had been classified in the 'Under Suspension Segment' for a period longer than 3 months. Suspension of trading in a stock means the temporary cessation of trading therein. The suspension decision is made by the Chairman of the BoD of ATHEX based on the Markets in Financial Instruments Directive (MiFID) and with the intention to safeguard the market and protect the interests of investors. In such a case, time-series data and also financial reports are lacking from most of the suspended companies, and this results in a lot of missing data to the dataset.

After having finalized the dataset's objects, selection and justification of the variables followed. The criteria that drive the selection decision for these variables were either scientific (previous studies in the field) or subjective (authors' initiations). First and foremost, dividend payments, which is the decision variable being modelled by this study, were recorded and then various accounts from the financial statements (balance sheet, income statement and cash flow statement) and various financial ratios were selected and recorded. Second, various non-financial variables related to administration, ownership, auditing, operating industry and other interesting corporate governance characteristics were gathered and recorded. All variables' codes, explanations and sources are provided in Table AI in APPENDIX A.

4.2. Dataset Descriptive Statistics

The variables included in the research dataset are divided into qualitative and quantitative. The former group includes variables with nominal and ordinal values, while the latter includes variables with interval and ratio values. Table II provides the frequencies, percentages and cumulative percentages of

DOI: 10.1002/isaf

Table II Descriptive statistics for the dataset's qualitative variables

Variable	Attribute	Frequency	Percent	Cumulative percent
1. FSYEAR	2007	246	33.3	33.3
	2008	246	33.3	66.7
	2009	246	33.3	100.0
	Total	**738**	**100.0**	
2. INDUSTRY	Personal and Household Goods	123	16.7	16.7
	Food and Beverage	87	11.8	28.5
	Industrial Goods and Services	81	11.0	39.4
	Construction and Materials	78	10.6	50.0
	Technology	66	8.9	58.9
	Basic Resources	48	6.5	65.4
	Travel and Leisure	45	6.1	71.5
	Retail	39	5.3	76.8
	Media	36	4.9	81.7
	Financial Services	30	4.1	85.8
	Real Estate	30	4.1	89.8
	Chemicals	27	3.7	93.5
	Health Care	24	3.3	96.7
	Utilities	12	1.6	98.4
	Oil and Gas	9	1.2	99.6
	Telecommunications	3	0.4	100.0
	Total	**738**	**100.0**	
3. SECTOR	Medium and Small Capitalization	417	56.5	56.5
	Big Capitalization	173	23.4	79.9
	Low Dispersion and Specific Features	95	12.9	92.8
	Under Supervision	51	6.9	99.7
	Under Suspension	2	0.3	100.0
	Total	**738**	**100.0**	
6. HEADQR	Attiki	588	79.7	79.7
	Thessaloniki	48	6.5	86.2
	Herakleion	12	1.6	87.8
	Viotia	12	1.6	89.4
	Kilkis	9	1.2	90.7
	Evros	6	0.8	91.5
	Imathia	6	0.8	92.3
	Larisa	6	0.8	93.1
	Serres	6	0.8	93.9
	Achaia	3	0.4	94.3
	Aitoloakarnania	3	0.4	94.7
	Chania	3	0.4	95.1
	Drama	3	0.4	95.5
	Evia	3	0.4	95.9
	Fokida	3	0.4	96.3
	Fthiotida	3	0.4	96.7
	FYROM	3	0.4	97.2
	Ioannina	3	0.4	97.6
	Kalamata	3	0.4	98.0
	Kavala	3	0.4	98.4
	Korinthos	3	0.4	98.8
	Lesvos	3	0.4	99.2
	Patra	3	0.4	99.6
	Rethymno	3	0.4	100.0
	Total	**738**	**100.0**	
9. NSUBSD	Both	420	56.9	56.9
	Domestic	144	19.5	76.4
	None	125	16.9	93.4
	Foreign	49	6.6	100.0
	Total	**738**	**100.0**	

(*Continues*)

DOI: 10.1002/isaf

Table II. (Continued)

Variable	Attribute	Frequency	Percent	Cumulative percent
12. SPLIT	None	677	91.7	91.7
	Normal	24	3.3	95.0
	Both	20	2.7	97.7
	Reverse	17	2.3	100.0
	Total	**738**	**100.0**	
13. CAPCGE	None	586	79.4	79.4
	Increase	101	13.7	93.1
	Both	31	4.2	97.3
	Decrease	20	2.7	100.0
	Total	**738**	**100.0**	
24. CEODUAL	No	440	59.6	59.6
	Yes	298	40.4	100.0
	Total	**738**	**100.0**	
33. AUDITOR	Sol	254	34.4	34.4
	Thornton	108	14.6	49.1
	Bdo	103	14.0	63.0
	Pwc	69	9.3	72.4
	Tilly	57	7.7	80.1
	Kpmg	34	4.6	84.7
	Ernst	32	4.3	89.0
	Deloitte	22	3.0	92.0
	Pkf	20	2.7	94.7
	Stephens	16	2.2	96.9
	Independent	9	1.2	98.1
	Orion	4	0.5	98.6
	Monday	3	0.4	99.1
	Nexia	3	0.4	99.5
	Rps	2	0.3	99.7
	Enel	1	0.1	99.9
	Rsm	1	0.1	100.0
	Total	**738**	**100.0**	
34. AUDITOROP	Unqualified	551	74.7	74.7
	Qualified	187	25.3	100.0
	Total	**738**	**100.0**	

qualitative variables in descending frequency order and Table III provides the minimum and maximum values, the mean and the standard deviation of quantitative variables.

Regarding qualitative variables, it is evident that some of them have a lot of possible attributes, like the variables 'INDUSTRY', 'HEADQR', 'AUDITOR', to name a few. On the other hand, quantitative variables are too many. Both of the above situations affect the dimension of the dataset and cause problems of dimensionality in DM. In Section 4.3 this problem will be challenged and managed effectively by employing appropriate statistical and DM techniques.

4.3. Dataset Exploration, Transformation and Purification

In DM it is often possible to have a dataset with a large number of variables. In such situations it is very likely that subsets of variables are highly correlated with each other. Including highly correlated variables in a classification or prediction model (or including variables that are unrelated to the outcome of interest) can lead to overfitting, and accuracy and reliability can suffer (Shmueli *et al.*, 2007). Also, retaining too many variables may lead to overfitting, in which the generality of the findings is hindered because the new data do not behave the same as the training data for all variables (Larose, 2005).

DOI: 10.1002/isaf

Table III Descriptive statistics for the dataset's quantitative variables

Variable	Min.	Max.	Mean	Standard deviation
4. FDYEAR	1879	2001	1974.78	n/a
5. LDATE	22 Feb 1912	04 Jan 2008	12 Jul 1993	n/a
7. EMPL	0	34,602	610.84	2069.543
8. SUBSD	0	172	10.10	17.957
10. NSHARES	610000	1,961,200,440	51,989,635.05	1.143E8
11. NV	0.30	8.63	0.8874	0.937
14. BoDFEES	0	14,600,000	1,181,651.47	1,513,179.800
15. BoD	4	26	7.75	2.397
16. FBoD	0	7	0.37	1.015
17. DBoD	0	26	7.35	2.495
18. MBoD	0	25	6.79	2.649
19. WBoD	0	10	0.97	1.352
20. EXBoD	1	11	3.56	1.653
21. NEXBoD	0	21	4.18	2.092
22. INDPBoD	0	21	2.37	1.275
23. NINDBoD	1	13	5.36	2.181
25. NOWN	0	9	2.75	1.348
26. OWNPRC	0.00	98.40	62.62	17.889
27. NINSTOWN	0	5	0.41	0.754
28. INSTOWNPRC	0.00	96.62	7.26	18.033
29. NMANGOWN	0	5	1.21	1.127
30. MANGOWNPRC	0.00	90.61	30.18	27.775
31. NFAMLOWN	0	9	1.04	1.492
32. FAMLOWNPRC	0.00	86.31	20.71	29.511
35. DIVD	0.00	6.50	0.09	0.358
36. EPS	−3.75	10.31	0.11	0.717
37. MV	0.09	64.55	3.85	6.436
38. BV	−1.88	75.40	2.93	5.198
39. TGA	0.00	1.31E10	1.16E8	7.903E8
40. ITGA	0.00	4.05E8	6.17E6	3.280E7
41. INVSUB	0.00	4.73E9	8.57E7	3.811E8
42. INVASS	0.00	2.15E8	1.16E6	1.348E7
43. DTXASS	0.00	1.88E8	2.57E6	1.528E7
44. FA	940.28	1.33E10	2.44E8	9.773E8
45. INV	0.00	1.41E9	2.27E7	9.378E7
46. ARECV	0.00	1.23E9	4.55E7	1.071E8
47. CASH	0.00	1.19E9	2.12E7	8.387E7
48. CA	231,617.77	2.50E9	9.73E7	2.382E8
49. TA	1,863,333.00	1.58E10	3.41E8	1.170E9
50. CC	1,811,112.91	5.06E9	1.07E8	3.566E8
51. RSV	−9.93E8	4.34E9	4.30E7	2.546E8
52. TRSH	−856,000.00	5.26E8	1.87E6	2.117E7
53. RETEAR	−2.46E8	1.54E9	2.32E7	1.250E8
54. TEQ	−43,732,098.00	6.45E9	1.64E8	5.319E8
55. DTXL	−283,440.39	4.91E8	5.33E6	2.428E7
56. LTDBT	0.00	6.35E9	8.70E7	4.461E8
57. TXL	−271,228.40	3.96E8	3.82E6	2.429E7
58. CL	101,089.23	3.06E9	8.18E7	2.605E8
59. TDBT	116,647.23	9.32E9	1.69E8	6.717E8
60. WCPT	−1.33E9	1.23E9	1.54E7	1.147E8
61. TREV	−39,072,236.84	9.32E9	2.00E8	7.773E8
62. COGS	1500.00	9.33E9	1.64E8	6.959E8
63. GPRF	−40,428,561.39	1.37E9	3.78E7	1.326E8
64. OPRF	−1.70E8	1.13E9	1.30E7	7.773E7
65. DEPR	0.00	5.67E8	9.03E6	4.642E7
66. TX	−93,747,000.00	3.50E8	4.62E6	2.556E7
67. NOPAT	−2.33E8	7.22E8	9.17E6	6.131E7
68. CFOP	−4.12E8	1.83E9	1.45E7	9.813E7
69. CFINV	−4.47E9	1.83E9	−2.21E7	2.114E8
70. CFFIN	−1.94E9	5.75E9	9.27E6	2.418E8
71. CURRAT	0.01	236.56	4.79	18.691

(*Continues*)

DOI: 10.1002/isaf

Table III. (Continued)

Variable	Min.	Max.	Mean	Standard deviation
72. QRAT	0.01	236.56	4.31	18.556
73. RECTURN	−111.86	3974.26	11.43	154.517
74. INVTURN	0.04	576,004.35	2709.42	37,090.095
75. PAYTURN	−16.95	37.36	3.37	3.830
76. DBTEQTY	−15.87	32.56	1.37	2.555
77. TDBTRAT	0.00	2.56	0.50	0.260
78. CASHCRAT	−8272.00	8655.20	3.26	593.697
79. ROA	−3.60	0.53	−0.00	0.169
80. ROE	−9.23	3.35	−0.01	0.584
81. FASSTURN	−13,288.61	21,366.67	16.79	952.420
82. TASSTURN	−0.62	8.77	0.62	0.793
83. SGA_EXPSLS	−0.33	660.33	1.83	25.813
84. FIN_EXPSLS	−1.44	660.33	1.45	26.371
85. SGA_EXPGPR	−343.64	871.08	2.85	40.137
86. FIN_EXPGPR	−147.36	871.08	2.44	37.955
87. MKTBOOKVAL	−6.17	1805.84	4.76	68.391
88. LFCYCLE1	−40.63	47.60	0.09	3.268
89. LFCYCLE2	−15.13	0.70	−0.09	0.947

Reducing the dimensionality of the data by deleting unsuitable attributes improves the performance of learning algorithms, it speeds them up and, more importantly, yields a more compact and more interpretable representation of the target concept, focusing the user's attention on the most relevant variables (Witten and Frank, 2005). Selecting the most relevant variables is usually suboptimal for building a predictor, particularly if the variables are redundant. Conversely, a subset of useful variables may exclude many redundant, but relevant, variables (Guyon and Elisseeff, 2003). The dimensionality of a dataset is also affected by the categories included in a predictor categorical variable, as a variable with m categories will be transformed into $m-1$ dummy variables when used in the analysis, resulting in a further dimension increase. One way to handle this is to reduce the number of categories by binning close bins together. However, this requires incorporating expert knowledge and common sense (Shmueli *et al*., 2007).

Based on the above argumentation, it clear that prior to applying DM techniques to the research dataset a specific procedure called 'feature/variable selection' should be implemented in order to avoid negative results in the next stages of the analysis. Specifically, the feature selection process will be realized by employing the one-way analysis of variance (ANOVA) test for quantitative variables and the Pearson's chi-square test of independence for qualitative variables in order to find relevant and irrelevant variables.

However, before this selection process, some new variables will be constructed, based on the raw data presented in Tables II and III, which are necessary in predicting the target variable of our analysis; and some of them have the advantage of being qualitative with nominal values, as many classification algorithms deal only with these types of variables. Table IV presents the equations applied in order to constructs the new variables.

For the benefit of feature selection, Table V presents the results of the one-way ANOVA test, and the Pearson's chi-square test results for qualitative variables are presented in Table VI. In Table V the dataset is divided into those records having BVID = 'no' and those having BVID = 'yes'. For each variable (columns 1 and 6) the mean within each sample is presented (columns 2, 3, 7and 8) and also the F statistic (columns 4 and 9) along with the p-values (columns 5 and 10) are provided. Those variables where the significance is lower than 0.05 are included in the next stage of the analysis, as a variable is salient if it has a high variance compared with others (Guyon and Elisseeff, 2003).

DOI: 10.1002/isaf

Table IV Dataset's variables transformation[a]

Code	Explanation
90. $\text{BDIVD} = \begin{cases} \text{yes if DIVD} > 0 \\ \text{no if DIVD} = 0 \end{cases}$	Binarization of the amount of dividends paid to shareholders
91. $\text{BSUBSD} = \begin{cases} \text{yes if SUBSD} > 0 \\ \text{no if SUBSD} = 0 \end{cases}$	Binarization of the number of subsidiaries owned by the company
92. $\text{BoDFEESAVG} = \frac{\text{BoDFEES}}{\text{BoD}}$	Average amount of fees delivered to BoD and management staff
93. $\text{FBoDPRC} = \frac{\text{FBoD}}{\text{BoD}}$	Percentage of foreign members of the BoD
94. $\text{DBoDPRC} = \frac{\text{DBoD}}{\text{BoD}}$	Percentage of domestic members of the BoD
95. $\text{MBoDPRC} = \frac{\text{MBoD}}{\text{BoD}}$	Percentage of men members of the BoD
96. $\text{WBoDPRC} = \frac{\text{WBoD}}{\text{BoD}}$	Percentage of women members of the BoD
97. $\text{EXBoDPRC} = \frac{\text{EXBoD}}{\text{BoD}}$	Percentage of executive members of the BoD
98. $\text{NEXBoDPRC} = \frac{\text{NEXBoD}}{\text{BoD}}$	Percentage of nonexecutive members of the BoD
99. $\text{INDPBoDPRC} = \frac{\text{INDPBoD}}{\text{BoD}}$	Percentage of independent members of the BoD
100. $\text{NINDPBoDPRC} = \frac{\text{NINDPBoD}}{\text{BoD}}$	Percentage of non independent members of the BoD
101. $\text{BNOWN} = \begin{cases} \text{yes if NOWN} > 0 \\ \text{no if NOWN} = 0 \end{cases}$	Binarization of the number of persons/companies owning more than 5% of a company's stocks
102. $\text{BNINSTOWN} = \begin{cases} \text{yes if NINSTOWN} > 0 \\ \text{no if NINSTOWN} = 0 \end{cases}$	Binarization of the number of institutional investors owning more than 5% of a company's stocks
103. $\text{BNMANGOWN} = \begin{cases} \text{yes if NMANGOWN} > 0 \\ \text{no if NMANGOWN} = 0 \end{cases}$	Binarization of the number of the company's management staff owning more than 5% of a company's stocks
104. $\text{BNFAMLOWN} = \begin{cases} \text{yes if NFAMLOWN} > 0 \\ \text{no if NFAMLOWN} = 0 \end{cases}$	Binarization of the number of persons having family relationships (same surname) and owning more than 5% of a company's stocks
105. $\text{DHEADQR} = \begin{cases} \text{Attiki if HEADQR} = \text{Attiki} \\ \text{Thessaloniki if HEADQR} = \text{Thessaloniki} \\ \text{Other if HEADQR} \neq \text{Attiki or Thessaloniki} \end{cases}$	Discretization of the company's headquarters
106. $\text{BCAPCGE} = \begin{cases} \text{yes if CAPCGE} = \text{Increase or Both or Decrease} \\ \text{no if CAPCGE} = \text{None} \end{cases}$	Binarization of the company's contributed capital increase/decrease
107. $\text{BAUDITOR} = \begin{cases} \text{yes if AUDITOR} = \text{Kpmg or Pwc or Ernst or Deloitte} \\ \text{no if AUDITOR} \neq \text{Kpmg or Pwc or Ernst or Deloitte} \end{cases}$	Binarization of the company's independent auditor
108. $\text{DAUDITOR} = \begin{cases} \text{GLDR if AUDITOR} = \text{Sol} \\ B4 \text{ if AUDITOR} = \text{Kpmg or Pwc or Ernst or Deloitte} \\ \text{Other if AUDITOR} \neq \text{Sol or Kpmg or Pwc or Ernst or Deloitte} \end{cases}$	Discretization of the company's independent auditor
109. In variable SECTOR the attributes 'Under Supervison' and 'Under Suspension' were merged in to attribute 'Under Suspevision'	

[a]Atributes 90, 91, 101–109 are Nominal and attributes 92–100 are Ratio.

Moreover, Table VI shows for each qualitative variable (column 1) the calculated chi-squared statistic (column 2), the degrees of freedom (column 3) and the *p*-values (column 4), where, again, variables with significance lower than 0.0001 are included in the next stage of the analysis. The feature selection

DOI: 10.1002/isaf

Table V Feature selection results: quantitative variables

	BDIVD (mean)					BDIVD (mean)			
Variable[a]	No	Yes	F	Sig.	Variable[a]	No	Yes	F	Sig.
4. FDYEAR	1973.90	1975.94	1.7	0.18385	5. LDATE	02 Feb 1993	09 Feb 1994	0.8	0.34807
7. EMPL	394.78	897.78	10.8	0.00105	**8. SUBSD**	7.43	13.65	22.2	0.00000
10. NSHARES	42,758,571	64,249,186	6.4	0.01138	11. NV	0.84	0.94	1.8	0.17130
14. BoDFEES	911,354	1,540,626	32.6	0.00000	15. BoD	7.48	8.10	12.3	0.00047
16. FBoD	0.26	0.52	12.2	0.00049	17. DBoD	7.18	7.58	4.7	0.03005
18. MBoD	6.50	7.18	11.9	0.00058	19. WBoD	1.00	0.92	0.6	0.43758
20. EXBoD	3.56	3.57	0.0	0.96362	21. NEXBoD	3.92	4.51	14.6	0.00014
22. INDPBoD	2.21	2.60	16.9	0.00004	23. NINDBoD	5.29	5.46	1.0	0.29941
25. NOWN	2.81	2.67	1.9	0.16399	26. OWNPRC	63.40	61.58	1.8	0.17009
27. NINSTOWN	0.35	0.49	6.0	0.01411	28. INSTOWNPRC	7.58	6.84	0.2	0.58440
29. NMANGOWN	1.22	1.19	0.1	0.73426	30. MANGOWNPRC	29.75	30.75	0.2	0.62881
31. NFAMLOWN	1.09	0.97	1.1	0.28030	32. FAMLOWNPRC	19.85	21.86	0.8	0.35967
36. EPS	−0.11	0.41	113.0	0.00000	**37. MV**	2.03	6.27	87.7	0.00000
38. BV	2.0167	4.14	31.6	0.00000	39. TGA	6.79E7	1.81E8	3.7	0.05387
40. ITGA	5.45E6	7.11E6	0.4	0.49655	41. INVSUB	5.05E7	1.32E8	8.4	0.00380
42. INVASS	929,543	1.47E6	0.2	0.58536	43. DTXASS	1.92E6	3.45E6	1.8	0.17717
44. FA	1.57E8	3.59E8	7.8	0.00535	45. INV	1.37E7	3.47E7	9.1	0.00261
46. ARECV	3.14E7	6.42E7	17.3	0.00003	47. CASH	1.29E7	3.22E7	9.6	0.00192
48. CA	6.49E7	1.40E8	18.4	0.00002	49. TA	2.22E8	4.99E8	10.2	0.00141
50. CC	9.05E7	1.29E8	2.1	0.14239	51. RSV	2.12E7	7.19E7	7.2	0.00729
52. TRSH	1.56E6	2.28E6	0.2	0.65076	**53. RETEAR**	−1.23E6	5.57E7	39.5	0.00000
54. TEQ	1.09E8	2.36E8	10.4	0.00125	55. DTXL	4.27E6	6.73E6	1.8	0.17350
56. LTDBT	5.15E7	1.34E8	6.2	0.01262	57. TXL	1.67E6	6.68E6	7.7	0.00553
58. CL	6.08E7	1.09E8	6.3	0.01164	59. TDBT	1.12E8	2.43E8	6.9	0.00871
60. WCPT	4.11E6	3.05E7	9.7	0.00187	**61. TREV**	8.65E7	3.42E8	19.6	0.00001
62. COGS	7.36E7	2.76E8	15.1	0.00011	**63. GPRF**	1.40E7	6.75E7	29.3	0.00000
64. OPRF	−1.93E6	3.29E7	38.3	0.00000	65. DEPR	4.75E6	1.47E7	8.3	0.00388

DOI: 10.1002/isaf

66. TX	416111	9.94E6	25.1	0.00000	**67. NOPAT**	−3.89E6	2.65E7	47.2	0.00000
68. CFOP	156670	3.35E7	21.5	0.00000	69. CFINV	−1.43E7	−3.24E7	1.3	0.25119
70. CFFIN	1.42E7	2.70E6	0.4	0.52172	71. CURRAT	4.13	5.67	1.2	0.26745
72. QRAT	3.63	5.21	1.3	0.25463	73. RECTURN	2.64	22.49	2.8	0.08975
74. INVTURN	111.35	5961.75	3.8	0.05149	75. PAYTURN	3.14	3.65	3.0	0.08142
76. DBTEQTY	1.59	1.09	7.0	0.00808	**77. TDBTRAT**	0.54	0.44	28.0	0.00000
78. CASHCRAT	−5.45	15.57	0.2	0.64942	**79. ROA**	−0.04	0.05	64.7	0.00000
80. ROE	−0.09	0.10	22.1	0.00000	81. FASSTURN	−35.27	81.84	2.6	0.10500
82. TASSTURN	0.51	0.75	15.4	0.00009	83. SGA_EXPSLS	0.62	3.34	1.9	0.16690
84. FIN_EXPSLS	0.13	3.13	2.1	0.14598	85. SGA_EXPGPR	2.71	3.03	0.0	0.91524
86. FIN_EXPGPR	1.16	4.07	0.9	0.32654	**89. BoDFEESAVG**	119,692	179798	24.0	0.00000
90. FBoDPRC	3.14	6.80	15.0	0.00011	91. DBoDPRC	96.38	93.20	9.8	0.00177
92. MBoDPRC	86.64	87.50	0.4	0.50546	93. WBoDPRC	13.60	12.50	0.7	0.40300
94. EXBoDPRC	47.66	45.27	3.8	0.04923	95. NEXBoDPRC	52.30	54.47	3.2	0.07302
96. INDPBoDPRC	31.20	32.82	3.3	0.06814	97. NINDBoDPRC	68.90	66.97	4.6	0.03189
102. MKTBOOKVAL	1.20	9.47	2.6	0.10419	103. LIFECYCLE1	−0.21	0.30	4.5	0.03270
104. LIFECYCLE2	−0.24	0.11	27.9	0.00000					

[a]Variables in bold have different means based on the ANOVA test with 99.99% confidence interval.

Table VI Feature selection results: qualitative variables

Variable[a]	χ^2	Df	Sig.
2. INDUSTRY	56.26	15	0.00000
3. SECTOR	106.09	4	0.00000
6. HEADQR	48.03	23	0.00166
9. NSUBSD	16.38	3	0.00094
12. SPLIT	0.55	3	0.90601
13. CAPCGE	8.50	3	0.03674
24. CEODUAL	0.82	1	0.36295
33. AUDITOR	40.47	16	0.00066
34. AUDITOROP	5.99	1	0.01433
91. BSUBSD	2.32	1	0.12724
101. BNOWN	1.33	1	0.24883
102. BNINSTOWN	2.26	1	0.13260
103. BNMANGOWN	1.50	1	0.22026
104. BNFAMLOWN	0.04	1	0.82846
105. DHEADQR	7.11	2	0.02856
106. BCAPCGE	0.75	1	0.38600
107. BAUDITOR	13.96	1	0.00019
108. DAUDITOR	14.70	2	0.00064

[a]Variables in bold italic are independent based on the chi-square test with 99.99% confidence interval.

has eliminated 66 out of 87 quantitative variables and 16 out of 18 qualitative variables based on their statistical properties. In conclusion, 23 variables remain in order to conduct the basic stage of the analysis.

5. EXPERIMENTAL RESULTS

After having finalized the dataset's objects, through the feature selection process, the next phase included the implementation of DM algorithms in order to predict the dividend payment decision. All algorithms' runs were made with the use of IBM® SPSS® Modeler 14.2 software (formerly SPSS Clementine) which is a powerful, versatile DM workbench that helps to build accurate predictive models quickly and intuitively, without programming. Three experiments were conducted. In the first experiment the algorithms were trained and validated on the whole sample data. In the second experiment the data were randomly partitioned into a training sample and validating sample with a 75%–25% analogy. In the third experiment the algorithms were trained on the records from fiscal year 2007–2008 and were validated on the records from fiscal year 2009.

In each one of the three experiments the DM algorithms were implemented with the same build settings. The C5.0 DT algorithm parameters were 75% pruning severity, global pruning and minimum five records per child branch. The feed-forward back-propagation NN algorithm parameters were exhaustive prune training method (persistence: 200; overall persistence: 4; hidden persistence: 100; hidden rate: 0.02; input persistence: 100; input rate 0.01) with a stopping criterion of 90% accuracy, one input layer with 41 neurons, and two hidden layers, one with 30 neurons and the other with 20 neurons.

The accuracy prediction results of all experiments are presented in Table VII. In the first experiment (EXA), where all models were trained and evaluated in the same dataset of 738 records, the C5.0 algorithm constructed a DT that reached an overall prediction accuracy of 93%, with only 5.70% of non-dividend-paying companies predicted incorrectly as dividend-paying companies (type I error) and only 8.83% of dividend-paying companies predicted incorrectly as non-dividend-paying companies (type II error). The NN also reached a high prediction accuracy as it succeeded in classifying correctly approximately 90% of companies, with 7.84% of non-dividend-paying companies predicted incorrectly as

DOI: 10.1002/isaf

Table VII Prediction accuracy of each model in each experiment[a]

Model	BDIVD	Pay dividends		Not pay dividends		Total		Errors (%)	
		Count	%	Count	%	Count	%	Type I	Type II
EXA									
C5.0	Correct	289	**91.17**	397	**94.30**	686	**92.95**	5.70	8.83
	Wrong	28	8.83	24	5.70	52	7.05		
	Total	317	100.00	421	100.00	738	100.00		
NN	Correct	277	87.38	388	92.16	665	90.11	7.84	12.62
	Wrong	40	12.62	33	7.84	73	9.89		
	Total	317	100.00	421	100.00	738	100.00		
Logistic regression	Correct	238	75.08	358	85.04	596	80.76	14.96	24.92
	Wrong	79	24.92	63	14.96	142	19.24		
	Total	317	100.00	421	100.00	738	100.00		
EXB									
C5.0	Correct	61	**83.56**	99	**93.39**	160	**89.39**	6.61	16.44
	Wrong	12	16.44	7	6.61	19	10.61		
	Total	73	100.00	106	100.00	179	100.00		
NN	Correct	59	80.82	95	89.62	154	86.03	10.38	19.18
	Wrong	14	19.18	11	10.38	25	13.97		
	Total	73	100.00	106	100.00	179	100.00		
Logistic regression	Correct	55	75.34	78	73.58	133	74.30	26.42	24.66
	Wrong	18	24.66	28	26.42	46	25.70		
	Total	73	100.00	106	100.00	179	100.00		
EXC									
C5.0	Correct	66	85.71	146	**86.39**	212	**86.18**	13.61	14.29
	Wrong	11	14.29	23	13.61	34	13.82		
	Total	77	100.00	169	100.00	246	100.00		
NN	Correct	67	**87.01**	136	80.47	203	82.52	19.53	12.99
	Wrong	10	12.99	33	19.53	43	17.48		
	Total	77	100.00	169	100.00	246	100.00		
Logistic regression	Correct	58	75.32	134	79.29	192	78.05	20.71	24.68
	Wrong	19	24.68	35	20.71	54	21.95		
	Total	77	100.00	169	100.00	246	100.00		

[a]EXA: whole sample; EXB: random sample partition with 75% training and 25% validating; EXC: sample partition with FSYEAR = '2007, 2008' training and FSYEAR = '2009' validating.
[b]Numbers in bold show the most accurate method in each experiment and in each attribute prediction.

dividend-paying companies and 12.62% of dividend-paying companies predicted incorrectly as non-dividend-paying companies. However, the logistic regression model had an almost 12% prediction accuracy lag with the C5.0 DT and an almost 10% prediction accuracy lag with the NN. Specifically, it achieved an overall prediction accuracy of 81%, with 14.96% of non-dividend-paying companies predicted incorrectly as dividend-paying companies and 24.92% of dividend-paying companies predicted incorrectly as non-dividend-paying companies. The lower prediction accuracy of the logistic regression model is evident from the comparison of type I and type II errors, where that of the DM method is two to three times lower.

In order to minimize the overfitting to data, where models achieve high performance in the training sample but suffer in predicting out-of-the-training-sample records, outliers and unknown records, we conducted a second experiment (EXB). The algorithms, in this experiment, were trained on a randomly selected sample of 599 records and were validated on the remaining sample of 179 records. The C5.0 DT reached an overall prediction accuracy of 89%, with only 6.61% of non-dividend-paying companies predicted incorrectly as dividend-paying companies and 16.44% of dividend-paying companies predicted incorrectly as non-dividend-paying companies. The NN reached an analogous prediction accuracy as it succeeded in classifying correctly approximately 86% of companies, with 10.38% of non-dividend-paying companies predicted incorrectly as dividend-paying companies and 19.18% of dividend-paying companies predicted incorrectly as non-dividend-paying companies. The prediction

DOI: 10.1002/isaf

accuracy lag of the logistic regression model has been increased with the C5.0 DT to almost 15% and with the NN to almost 12%. Specifically, it achieved an overall prediction accuracy of 74%, with 26.42% of non-dividend-paying companies predicted incorrectly as dividend-paying companies and 24.66% of dividend-paying companies predicted incorrectly as non-dividend-paying companies. The lower prediction accuracy of the logistic regression model is evident from the type I and type II errors, wherein one out of four records was predicted incorrectly.

In our third experiment (EXC) the algorithms were trained on all records from the fiscal years 2007–2008 (492 records) and were validated on all records from the fiscal year 2009 (246 records). This experiment reveals financial knowledge, as it uses data from preceeding years to forecast future outcomes. The C5.0 DT reached an overall prediction accuracy of 86%, with 13.61% of non-dividend-paying companies predicted incorrectly as dividend-paying companies and 14.29% of dividend-paying companies predicted incorrectly as non-dividend-paying companies. The NN reached an analogous prediction accuracy as it succeeded in classifying correctly approximately 82% of companies, with 19.53% of non-dividend-paying companies predicted incorrectly as dividend-paying companies and 12.99% of dividend-paying companies predicted incorrectly as non-dividend-paying companies. In contrast to previous experiments, the prediction accuracy lag of the logistic regression model decreased with the C5.0 DT to almost 8% and with the NN to almost 5%. Specifically, it achieved an overall prediction accuracy of 78%, with 20.71% of non-dividend-paying companies predicted incorrectly as dividend-paying companies and 24.68% of dividend-paying companies predicted incorrectly as non-dividend-paying companies.

The major finding of these experiments is the prediction accuracy superiority of the DM approaches against logistic regression. However, a secondary finding is the fact that the accuracy divergence is increased in the case where models face new and unknown records and is decreased when models are trained with time-consecutive records.

The results presented so far allow us to conclude that the implementation of DM methods in modelling the debatable question of dividend payment has the potential to yield better results than those gained by the, so far, mainstream method of logistic regression. Consequently, the answer to RQ1 is positive. However, this research has another scope concerning the development of a convenient and effective decision-support tool to investors that want to construct and manage a portfolio of securities.

In line with that objective, Table VIII presents the five variables that contributed most in constructing each model. This variable importance ranking indicates the relative importance of each variable in estimating each model. Since the values are relative, their sum is equal to unity, for all variables included in each model. Variable importance is determined by computing the reduction in variance of the target attributable to each predictor, via a sensitivity analysis. It does not relate to model accuracy; rather, it only relates to the importance of each variable in making a prediction, not whether or not the prediction is accurate.

Regarding DTs constructed with the C5.0 algorithm: in all experiments, the two most important variables are ‘NOPAT’ and ‘RETEAR’. In the NNs, the three most important variables are ‘EPS’, ‘NOPAT’ and ‘INDUSTRY’, while logistic regression has ranked ‘ROA’ as the most important variable without having a consistency, among experiments, regarding the second most important variable. The three selected methods do not agree on which variables affect the dividend payment decision more. The results gained from DTs show that the features dictating the decision to ‘pay’ or ‘not pay’ dividends are a company’s fundamentals, while in NNs and in logistic regression some non-pure financial features, namely operating industry and sector classification in the stock market, play a catalytic role.

The variables presented in Table VIII provide some initial evidence on DP determinants. However, we need to define the sign and magnitude of each determinant variable so as to provide a scientific answer to RQ2. It is necessary to know which values (range or attribute) of these variables are dictating

DOI: 10.1002/isaf

Table VIII The top five significant variables of each model in each experiment

Model[c]		1st	2nd	3rd	4th	5th
EXA						
C5.0 (8 VARs)	VAR	NOPAT	RETEAR	ROA	OPRF	MV
	RI	0.383	0.259	0.162	0.126	0.059
NN (23 VARs)	VAR	EPS	NOPAT	INDUSTRY	MV	OPRF
	RI	0.093	0.075	0.073	0.067	0.067
Logistic regression (5 VARs)	VAR	ROA	EPS	LFCYCLE2	SECTOR	MV
	RI	0.347	0.251	0.174	0.163	0.065
EXB						
C5.0 (7 VARs)	VAR	NOPAT	RETEAR	ROE	MV	TDBTRAT
	RI	0.461	0.247	0.187	0.073	0.016
NN (23 VARs)	VAR	EPS	NOPAT	INDUSTRY	OPRF	ROA
	RI	0.088	0.076	0.075	0.066	0.063
Logistic regression (10 VARs)	VAR	ROA	OPRF	EPS	LFCYCLE2	SECTOR
	RI	0.323	0.208	0.180	0.111	0.085
EXC						
C5.0 (6 VARs)	VAR	NOPAT	RETEAR	ROE	MV	CA
	RI	0.471	0.303	0.189	0.018	0.016
NN (23 VARs)	VAR	EPS	INDUSTRY	NOPAT	ROA	TDBTRAT
	RI	0.077	0.074	0.068	0.062	0.061
Logistic regression (3 VARs)	VAR	ROA	LFCYCLE2	EPS	—	—
	RI	0.510	0.296	0.194	—	—

[a]EXA: whole sample; EXB: random sample partition with 75% training and 25% validating; EXC: sample partition with FSYEAR = ‘2007, 2008’ training and FSYEAR = ‘2009’ validating.
[b]VAR: variable; RI: relative importance.
[c]Text in parentheses provides the total number of variables included in each model.

the decision to ‘pay’ dividends and which are not. For this purpose, Table IX presents the rules derived from the DT generated by the C5.0 algorithm in the third experiment (an illustrative presentation of the DT can be found in APPENDIX A). Since the C5.0 algorithm proved to be the most accurate under all experiments, since an effective DP model should be capable of predicting the next year’s results based on the data from preceding years, and due to the complicated and hardly interpretable nature of NNs the DT of the third experiment was selected in order to answer the RQ2.

Focusing on the two rules that represent almost 80% of the instances, a better answer to RQ2 is provided. A company having more than 478,000 in net operating profits after taxes, more than 193,000 in

Table IX The rules derived from the DT of the EXC

Rule	Description	INS[a]	CFD[b]
Pay dividend			
1	If NOPAT < = 477793 and NOPAT > 92146.540 and CA > 12072901.280 and MV > 0.790 then yes	12 (2.44%)	0.750
2	If NOPAT > 477793 and RETEAR > 192899 and ROE < = 0.024 and SUBSD < = 11 then yes	16 (3.25%)	0.625
3	If NOPAT > 477793 and RETEAR > 192899 and ROE > 0.024 then yes	227 (46.14%)	0.903
Not pay dividend			
4	If NOPAT < = 477793 and NOPAT < = 92146.540 then no	162 (32.92%)	0.981
5	If NOPAT < = 477793 and NOPAT > 92146.540 and CA < = 12072901.280 then no	13 (2.64%)	1.000
6	If NOPAT < = 477793 and NOPAT > 92146.540 and CA > 12072901.280 and MV < = 0.790 then no	6 (1.22%)	1.000
7	If NOPAT > 477793 and RETEAR < = 192899 then no	45 (9.15%)	0.778
8	If NOPAT > 477793 and RETEAR > 192899 and ROE < = 0.024 and SUBSD > 11 then no	11 (2.24%)	0.727

[a]INS: Instances = the number of records to which the rule applies.
[b]CFD: Confidence = the proportion of those records for which the entire rule is true, (number of records where rule is correct)/(number of records for which the rule’s antecedents are true).

DOI: 10.1002/isaf

retained earnings, and a return on equity ratio greater than 2.4% has great possibilities to pay dividends (Rule 3). On the other hand, a company having less than 92,000 in net operating profits after taxes has great possibilities to not pay dividends (Rule 4).

The results are consistent with the life cycle theory of dividends, where, according to pioneers of this theory (DeAngelo *et al*., 2006), dividends tend to be paid by mature, established firms, plausibly reflecting a financial life cycle in which young firms face relative abundant investment opportunities with limited resources so that retention dominates distribution, whereas mature firms are better candidates to pay dividends because they have higher profitability and fewer attractive investment opportunities.

6. CONCLUSIONS

Understanding the issue of what determines the magnitude of dividend payout is very important as many corporations distribute a substantial amount of their resources to shareholders every year. Moreover, security analysts and consulting firms need to know the proposed DP of a firm as they make recommendations to their clients/potential investors.

During the last quarter of the twentieth century, advances in computer science (theoretical and applied) and in computer engineering enabled companies to gain semantic benefits via the utilization of DM models. These models have gained much popularity in the fields of marketing, production, accounting, auditing and finance. In finance, and more precisely in the DP field, the studies implementing DM methods are very limited.

This study investigated, via two research questions, whether DM methods are more effective than traditional logistic regression techniques in gaining insights into the DP issue and, aiming to assist decision making, moved a step further by providing those variables contributing most in the decision to pay or not pay dividends. The results show that DM methods are more accurate in predicting the dividend payment decision and that profitability is the most important factor in deciding to pay dividends.

The findings can be used by various parties. First, academics that instigate the DP of companies might gradually start utilizing DM methods in their studies as these have been proved to be more accurate. Second, individual investors or even more portfolio management companies could use the DT created by this study in order to select securities for their portfolio as dividend is a semantic criterion to select a security. However, as with any study, this research has limitations. These concern the domain of the dataset. The fact that the dataset refers to a 3 year time frame of listed companies in the ATHEX makes any generalization of the findings to other countries doubtful. Stock exchange markets in other countries may have other regulation directives and different corporate governance rules resulting in different DPs. One avenue for future research is driven by the limitation previously noted. A similar study in other countries examining and comparing the DP results of DM methods could serve to further extend and enhance these findings.

7. DATASET ACCESS

Our dataset is maintained under the Data Engineering Laboratory (DELAB) in the department of Informatics at Aristotle University of Thessaloniki. It can be found at http://delab.csd.auth.gr/~symeon/index.php (file name: Athens_Stock_Exchange_Dataset.xlsx).

ACKNOWLEDGEMENTS

Pantelis Longinidis gratefully acknowledges financial support from the ‘Alexander S. Onassis Public Benefit Foundation’ under the fellowship code G ZD 037-2/2011-2012.

DOI: 10.1002/isaf

APPENDIX A

Table A1. The dataset's variables[a]

Code [Attribute-Source b]	Explanation	Justification (literature support)[c]
1. FSYEAR[O-H]	The fiscal year of each company's data	Auxiliary
2. INDUSTRY[N-A]	The company's operating industry	Fundamental feature (1, 2, 28, 30)
3. SECTOR[N-A]	The company's classification into sectors according to ATHEX rulebook	Fundamental feature
4. FDYEAR[I-A]	The foundation year of each company	A proxy for age (32, 33)
5. LDATE[I-A]	The date when the company listed for first time in the ATHEX	A proxy for age
6. HEADQR[N-A]	The place where the company's headquarters are located	A proxy for location synergies
7. EMPL[I-C]	The number of employees	A proxy for the size (30)
8. SUBSD[I-C]	The number of subsidiaries owned by the company	A proxy for corporate governance
9. NSUBSD[N-C]	The nationality of subsidiaries	A proxy for corporate governance
10. NSHARES[I-C]	The number of each company's shares	A proxy for share attractiveness
11. NV[R-C]	The nominal value of the company's stocks at the end of each fiscal year	Fundamental feature
12. SPLIT[N-C]	If the company has changed its nominal value via split or reverse split	A proxy for corporate governance
13. CAPCGE[N-C]	If the company has increase/decrease its contributed capital	A proxy for corporate governance
14. BoDFEES[R-C]	The amount of fees delivered to BoD and management staff	A proxy for corporate governance
15. BoD[I-A]	The number of persons constituting the BoD	A proxy for corporate governance quality (15, 28, 30)
16. FBoD[I-A]	The number of foreign members of the BoD	A proxy for corporate governance
17. DBoD[I-A]	The number of domestic members of the BoD	A proxy for corporate governance
18. MBoD[I-A]	The number of male members of the BoD	A proxy for corporate governance
19. WBoD[I-A]	The number of female members of the BoD	A proxy for corporate governance
20. EXBoD[I-A]	The number of executive members of the BoD	A proxy for corporate governance
21. NEXBoD[I-A]	The number of nonexecutive members of the BoD	A proxy for corporate governance (15, 28, 30)
22. INDPBoD[I-A]	The number of independent members of the BoD	A proxy for corporate governance (3, 15, 28)
23. NINDBoD[I-A]	The number of nonindependent members of the BoD	A proxy for corporate governance
24. CEODUAL[N-A]	If CEO and president of the BoD is the same person	A proxy for ownership concentration (15)
25. NOWN[I-C]	The number of persons/companies owning more than 5% of a company's stocks	A proxy for ownership concentration (5, 8, 17, 19, 21, 22, 29)
26. OWNPRC[R-C]	The total ownership percentage of persons/companies owning more than 5% of a company's stocks	A proxy for ownership concentration (5, 8, 19, 21, 22, 29)
27. NINSTOWN[I-C]	The number of institutional investors owning more than 5% of a company's stocks	A proxy for institutional ownership concentration (5, 8, 13, 16)
28. INSTOWNPRC[R-C]	The total ownership percentage of institutional investors owning more than 5% of a company's stocks	A proxy for institutional ownership concentration (5, 8, 13, 16)
29. NMANGOWN[I-C]	The number of the company's management staff owning more than 5% of a company's stocks	A proxy for managerial ownership concentration (4, 5, 7, 8, 12, 13, 22, 25, 28)
30. MANGOWNPRC[R-C]	The total ownership percentage of the company's management staff owning more than 5% of a company's stocks	A proxy for managerial ownership concentration and/or managerial stock incentives (4, 5, 7, 8, 12, 13, 22, 25, 28)

(*Continues*)

DOI: 10.1002/isaf

Table A1. (Continued)

Code $^{\text{Attribute-Source b}}$	Explanation	Justification (literature support)$^{\text{c}}$
31. NFAMLOWN$^{\text{I-C}}$	The number of persons having family relationships (same surname) and owning more than 5% of a company's stocks	A proxy for family ownership concentration (15)
32. FAMLOWNPRC$^{\text{R-C}}$	The total ownership percentage of persons having family relationships (same surname) and owning more than 5% of a company's stocks	A proxy for family ownership concentration (15)
33. AUDITOR$^{\text{N-A,NF}}$	The company's independent auditor	A proxy for auditing quality (30)
34. AUDITOROP$^{\text{N-A,NF}}$	The opinion of the independent auditor for the financial statements of the company	A proxy for auditing quality
35. DIVD$^{\text{R-A,NF}}$	The amount of dividends paid to shareholders	Auxiliary
36. EPS$^{\text{R-H}}$	Earnings per share	A proxy for profitability (21, 30)
37. MV$^{\text{R-A,NF}}$	The market value of the company's stock at the end of each fiscal year	A proxy for the size (2, 17, 29, 31)
38. BV$^{\text{R-NF}}$	The book value of the company's stocks at the end of each fiscal year	Fundamental feature
39. TGA$^{\text{R-H}}$	Tangible assets	A proxy for the size (11)
40. ITGA$^{\text{R-H}}$	Intangible assets	Fundamental feature
41. INVSUB$^{\text{R-H}}$	Investment in subsidiaries companies	Fundamental feature
42. INVASS$^{\text{R-H}}$	Investment in associate companies	Fundamental feature
43. DTXASS$^{\text{R-H}}$	Deferred tax liabilities	Fundamental feature
44. FA$^{\text{R-H}}$	Fixed assets	Fundamental feature
45. INV$^{\text{R-H}}$	Inventory	Fundamental feature
46. ARECV$^{\text{R-H}}$	Accounts receivables	Fundamental feature
47. CASH$^{\text{R-H}}$	Cash	A proxy for liquidity (2, 9)
48. CA$^{\text{R-H}}$	Current assets	Fundamental feature
49. TA$^{\text{R-H}}$	Total assets	A proxy for the size (5, 8, 10, 14, 15, 18, 20, 21, 23, 25, 26, 28, 29, 32)
50. CC$^{\text{R-H}}$	Contributed capital	Fundamental feature
51. RSV$^{\text{R-H}}$	Reserves	Fundamental feature
52. TRSH$^{\text{R-H}}$	Treasury shares	Fundamental feature
53. RETEAR$^{\text{R-H}}$	Retain earnings	Fundamental feature
54. TEQ$^{\text{R-H}}$	Total equity	A proxy for the size (8, 18)
55. DTXL$^{\text{R-H}}$	Differed tax liabilities	Fundamental feature
56. LTDBT$^{\text{R-H}}$	Long-term debt	Fundamental feature
57. TXL$^{\text{R-H}}$	Tax liabilities	Fundamental feature
58. CL$^{\text{R-H}}$	Current liabilities	Fundamental feature
59. TDBT$^{\text{R-H}}$	Total debt	Fundamental feature
60. WCPT$^{\text{R-H}}$	Working capital	Fundamental feature
61. TREV$^{\text{R-H}}$	Total revenues	A proxy for the size (6–8, 12, 19)
62. COGS$^{\text{R-H}}$	Cost of goods sold	Fundamental feature
63. GPRF$^{\text{R-H}}$	Gross profit	Fundamental feature
64. OPRF$^{\text{R-H}}$	Operating profit	Fundamental feature

DOI: 10.1002/isaf

65. DEPR[R-H]	Depreciation	Fundamental feature
66. TX[R-H]	Taxes	Fundamental feature (16)
67. NOPAT[R-H]	Net operating profit after taxes	A proxy for profitability (9, 13, 21)
68. CFOP[R-H]	Cash flow from operating activities	A proxy for liquidity
69. CFINV[R-H]	Cash flow from investment activities	A proxy for liquidity
70. CFFIN[R-H]	Cash flow from financing activities	A proxy for liquidity
71. CURRAT[R-H]	Current ratio (CA/CL)	A proxy for liquidity (14, 23)
72. QRAT[R-H]	Quick ratio ((CA − INV)/CL)	A proxy for liquidity
73. RECTURN[R-H]	Receivables turnover (TREV/ARECV)	A proxy for assets utilization
74. INVTURN[R-H]	Inventory turnover (COGS/INV)	A proxy for assets utilization
75. PAYTURN[R-H]	Payables turnover (TREV/CL)	A proxy for assets utilization
76. DBTEQTY[R-H]	Debt/equity ratio (TDBT/TEQ)	A proxy for leverage (9, 17, 21, 23, 24, 30)
77. TDBTRAT[R-H]	Total debt ratio (TDBT/TA)	A proxy for leverage (11, 14, 15, 18, 19, 25, 26, 28, 29, 32)
78. CASHCRAT[R-H]	Cash coverage ratio ((EBIT + DEPR)/INTEREST)	A proxy for leverage
79. ROA[R-H]	Return on assets (NOPAT/TA)	A proxy for profitability (3, 8, 11, 15, 17, 18, 23, 26, 27, 29, 31)
80. ROE[R-H]	Return on equity (NOPAT/TEQ)	A proxy for profitability (3, 15, 24)
81. FASSTURN[R-H]	Fixed assets turnover (TREV/FA)	A proxy for assets utilization
82. TASSTURN[R-H]	Total assets turnover (TREV/TA)	A proxy for assets utilization
83. SGA_EXPSLS[R-H]	Selling general and administrative expenses to sales (SGA/TREV)	A proxy for cost effectiveness
84. FIN_EXPSLS[R-H]	Finance expenses to sales (FIN/TREV)	A proxy for cost effectiveness
85. SGA_EXPGPR[R-H]	Selling general and administrative expenses to gross profit (SGA/GPRF)	A proxy for cost effectiveness
86. FIN_EXPGPR[R-H]	Finance expenses to gross profit (FIN/GPRF)	A proxy for cost effectiveness
87. MKTBOOKVAL[R-AC]	Market value to book value (MV/BV)	A proxy for investment opportunities (6, 11, 12, 15–21, 23, 25, 29, 31)
88. LFCYCLE1[R-AC]	RETEAR/TEQ	A proxy for the lifecycle (18, 20, 31)
89. LFCYCLE2[R-AC]	RETEAR/TA	A proxy for the lifecycle (18, 21)

[a]Variables 10, 11 and 38–60 are found in the balance sheet. Variables 35, 36 and 61–67 are found in the income statement. Variables 68–70 are found in the statement of cash flows. Variables 12 and 13 are found in the statement of shareholders' equity. Variables 71–89 are financial ratios calculated from the accounts presented in the balance sheet, in the income statement, in the statement of cash flows and in the statement of shareholders' equity. Variables 1–9 and 14–37 are found in the annual report.
[b]O: ordinal; I: interval; N: nominal; R: ratio. A: ATHEX; H: Hellastat; NF: Naftemporiki; C: company's official web site; AC: authors' calculation based on A, H, NF and C.
[c]1: Michel (1979); 2: Baker *et al*. (1985); 3: Schellenger *et al*. (1989); 4: Agrawal and Jayaraman (1994); 5: Alli *et al*. (1993); 6: Barclay *et al*. (1995); 7: Holder *et al*. (1998); 8: Chen and Steiner (1999); 9: Baker *et al*. (2001); 10: Fama and French (2001); 11: Ooi (2001); 12: Dickens *et al*. (2002); 13: Short *et al*. (2002); 14: Omran and Pointon (2004); 15: Chen *et al*. (2005); 16: Amidu and Abor (2006); 17: Ben Naceur *et al*. (2006); 18: DeAngelo *et al*. (2006); 19: Mancinelli and Ozkan (2006); 20: Denis and Osobov (2008); 21: Ahmed and Javid (2009); 22: Chen and Dhiensiri (2009); 23: Kim and Gu (2009); 24: Al-Kuwari (2010); 25: Nam *et al*. (2010); 26: Brockman and Unlu (2011); 27: Coulton and Ruddock (2011); 28: Jiraporn *et al*. (2011); 29: Renneboog and Trojanowski (2011); 30: Shabibi and Ramesh (2011); 31: He (2012); 32: He *et al*. (2012); 33: Manos *et al*. (2012).

DOI: 10.1002/isaf

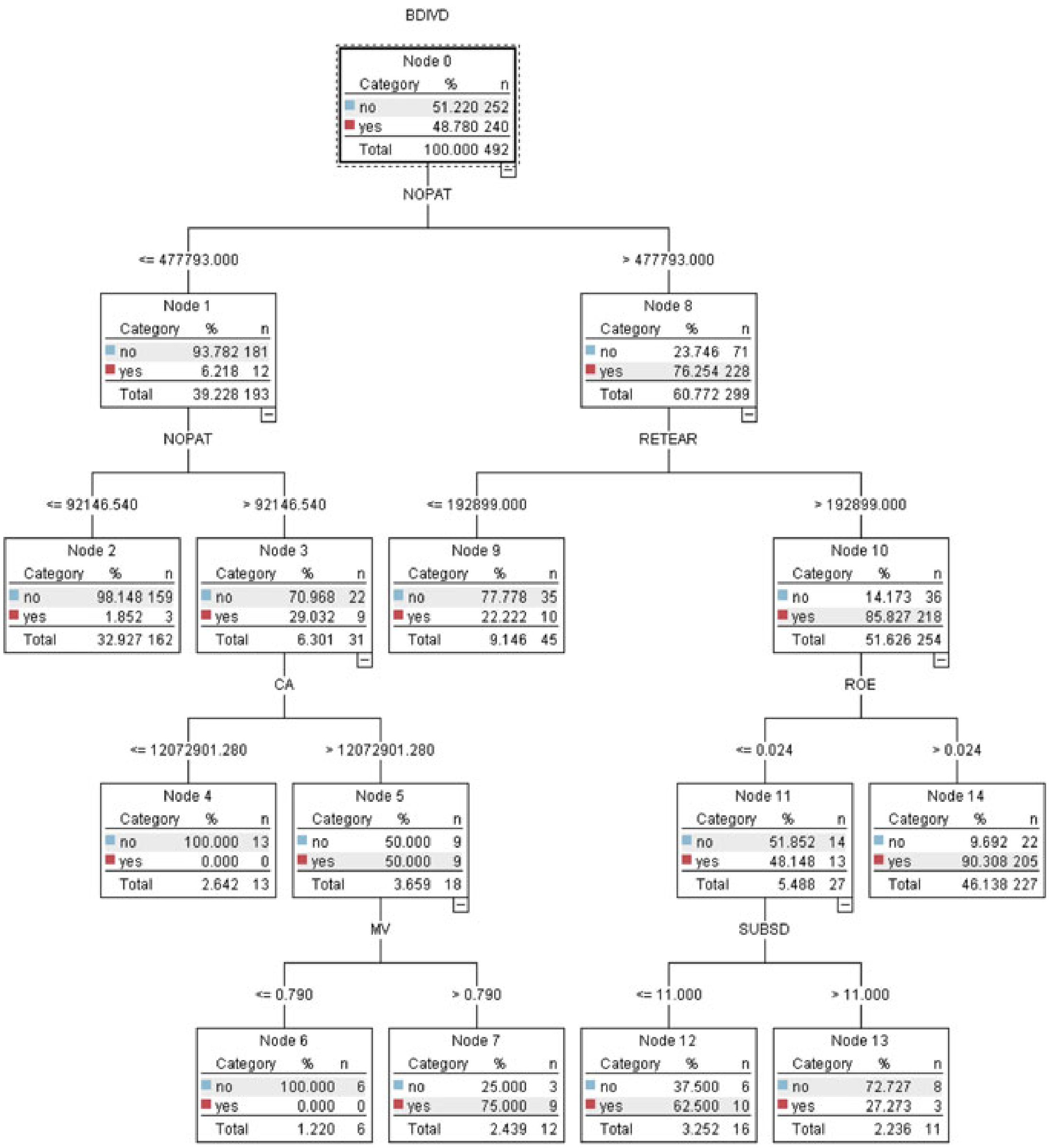

Figure A.1 The C5.0 DT.

REFERENCES

Agrawal A, Jayaraman N. 1994. The dividend policies of all-equity firms: a direct test of the free cash flow theory. *Managerial and Decision Economics* **15**(2): 139–148.

Ahmed H, Javid YA. 2009. The determinants of dividend policy in Pakistan. *International Research Journal of Finance and Economics* **29**: 110–125.

Al-Kuwari D. 2010. To pay or not to pay: using emerging panel data to identify factors influencing corporate dividend payout decisions. *International Research Journal of Finance and Economics* (**42**): 19–36.

Alli KL, Khan AQ, Ramirez GG. 1993. Determinants of corporate dividend policy: a factorial analysis. *Financial Review* **28**(4): 523–547.

Amidu M, Abor J. 2006. Determinants of dividend payout ratios in Ghana. *Journal of Risk Finance* **7**(2): 136–145.

Ang JS, Ciccone SJ. 2009. Dividend irrelevance theory. In Dividends and Dividend Policy, Baker KH (ed.). John Wiley & Sons, Inc.: Hoboken, NJ; 97–113.

Baker HK. 2009. Dividends and dividend policy: an overview. In Dividends and Dividend Policy, Baker KH (ed.). John Wiley & Sons, Inc.: Hoboken, NJ; 3–19.

Baker HK, Farrelly EG, Edelman BR. 1985. A survey of management views on dividend policy. *Financial Management* **14**(3): 78–84.

Baker HK, Veit ET, Powell GE. 2001. Factors influencing dividend policy decisions of Nasdaq firms. *Financial Review* **36**(3): 19–38.

Baker HK, Powell GE, Veit ET. 2002. Revisiting the dividend puzzle: do all of the pieces now fit? *Review of Financial Economics* **11**(4): 241–261.

Baker HK, Dutta S, Saadi S. 2008. Impact of financial and multinational operations on manager perceptions of dividends. *Global Finance Journal* **19**(2): 171–186.

Barclay MJ, Smith CW, Watts RL. 1995. The determinants of corporate leverage and dividend policies. *Journal of Applied Corporate Finance* **7**(4): 4–19.

Ben Naceur S, Goaied M, Belanes A. 2006. On the determinants and dynamics of dividend policy. *International Review of Finance* **6**(1–2): 1–23.

Bigus JP. 1996. Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support. McGraw-Hill: New York, NY.

Black F 1976. The dividend puzzle. *Journal of Portfolio Management* **2**(2): 5–8.

Bose R 2009. Advanced analytics: opportunities and challenges. *Industrial Management & Data Systems* **109**(2): 155–172.

Brealey RA, Myers SC, Marcus AJ. 2004. Fundamentals of Corporate Finance, 4th edn. Irwin/McGraw-Hill: New York, NY.

Brockman P, Unlu E. 2011. Earned/contributed capital, dividend policy, and disclosure quality: an international study. *Journal of Banking & Finance* **35**(7): 1610–1625.

Bulan LT, Subramanian N. 2009. The firm life cycle theory of dividends. In Dividends and Dividend Policy, Baker KH (ed.). John Wiley & Sons, Inc.: Hoboken, NJ; 201–213.

Chen CR, Steiner TL. 1999. Managerial ownership and agency conflicts: A nonlinear simultaneous equation analysis of managerial ownership, risk taking, debt policy, and dividend policy. *Financial Review* **34**(1): 119–136.

Chen J, Dhiensiri N. 2009. Determinants of dividend policy: the evidence from New Zealand. *International Research Journal of Finance and Economics* (**34**): 18–28.

Chen Z, Cheung Y-L, Stouraitis A, Wong AWS. 2005. Ownership concentration, firm performance, and dividend policy in Hong Kong. *Pacific-Basin Finance Journal* **13**(4): 431–449.

Coulton JJ, Ruddock C. 2011. Corporate payout policy in Australia and a test of the life-cycle theory. *Accounting & Finance* **51**(2): 381–407.

De Rooij M, Renneboog L. 2009. The catering theory of dividends. In Dividends and Dividend Policy, Baker KH (ed.). John Wiley & Sons, Inc.: Hoboken, NJ; 215–238.

DeAngelo H, DeAngelo L, Stulz RM. 2006. Dividend policy and the earned/contributed capital mix: a test of the life-cycle theory. *Journal of Financial Economics* **81**(2): 227–254.

Denis DJ, Osobov I. 2008. Why do firms pay dividends? International evidence on the determinants of dividend policy. *Journal of Financial Economics* **89**(1): 62–82.

Denis DJ, Stepanyan G. 2009. Factors influencing dividends. In Dividends and Dividend Policy, Baker KH (ed.). John Wiley & Sons, Inc.: Hoboken, NJ; 55–69.

Dickens NR, Casey KM, Newman AJ. 2002. Bank dividend policy: explanatory factors. *Quarterly Journal of Business & Economics* **41**(1–2): 3–12.

Dutta S, Saadi S. 2009. Dividend policy and corporate governance. In Dividends and Dividend Policy, Baker KH (ed.). John Wiley & Sons, Inc.: Hoboken, NJ; 447–462.

Fama EF, French KR. 2001. Disappearing dividends: changing firm characteristics or lower propensity to pay? *Journal of Financial Economics* **60**(1): 3–43.

DOI: 10.1002/isaf

Filbeck G. 2009. Asymmetric information and signaling theory. In Dividends and Dividend Policy, Baker KH (ed.). John Wiley & Sons, Inc.: Hoboken, NJ; 163–177.
Frankfurter GM, Wood BG. 1997. The evolution of corporate dividend policy. *Journal of Financial Education* **23**(1): 16–33.
Frankfurter GM, Wood BG. 2002. Dividend policy theories and their empirical tests. *International Review of Financial Analysis* **11**(2): 111–138.
Giudici P, Figini S. 2009. Applied Data Mining for Business and Industry, 2nd edn. John Wiley & Sons, Ltd: Chichester.
Guyon I, Elisseeff A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**(3): 1157–1182.
He T, Li W, Tang G. 2012. Dividends behavior in state- versus family-controlled firms: evidence from Hong Kong. *Journal of Business Ethics* **110**(1): 97–112.
He W. 2012. Agency problems, product market competition and dividend policies in Japan. *Accounting & Finance* **52**(3): 873–901.
Holder ME, Langrehr FW, Hexter JL. 1998. Dividend policy determinants: an investigation of the influences of stakeholder theory. *Financial Management* **27**(3): 73–82.
Huang S-Y, Tsaih R-H, Lin W-Y. 2012. Unsupervised neural networks approach for understanding fraudulent financial reporting. *Industrial Management & Data Systems* **112**(2): 224–244.
IBM Corporation. 2011. IBM SPSS Modeler 14.2 User's Guide, 2nd edn. IBM Corporation: Armonk, NY.
Jessen H, Paliouras G. 2001. Data mining in economics, finance, and marketing. In Machine Learning and Its Applications, Paliouras G, Karkaletsis V, Spyropoulos C (eds). Springer-Verlag: Berlin; 295–299.
Jiraporn P, Kim J-C, Kim YS. 2011. Dividend payouts and corporate governance quality: an empirical investigation. *Financial Review* **46**(2): 251–279.
Kim H, Gu Z. 2009. Financial features of dividend-paying firms in the hospitality industry: A logistic regression analysis. *International Journal of Hospitality Management* **28**(3): 359–366.
Kim J, Won C, Bae JK. 2010. A knowledge integration model for the prediction of corporate dividends. *Expert Systems with Applications* **37**(2): 1344–1350.
Kirkos E, Manolopoulos Y. 2004. Data mining in finance and accounting: a review of current research trends. In 1st International Conference on Enterprise Systems and Accounting (ICESAcc), Thessaloniki, Greece.
Kovalerchuk B, Vityaev E. 2005. Data mining for financial applications. In Data Mining and Knowledge Discovery Handbook, Maimon O, Rokach L (eds). Springer: New York, NY; 1203–1224.
Larose DT. 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Inc.: Hoboken, NJ.
Lee SJ, Siau K. 2001. A review of data mining techniques. *Industrial Management & Data Systems* **101**(1): 41–46.
Li R, Wang Z-o. 2004. Mining classification rules using rough sets and neural networks. *European Journal of Operational Research* **157**(2): 439–448.
Lin W-Y, Hu Y-H, Tsai C-F. 2012. Machine learning in financial crisis prediction: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* **42**(4): 421–436.
Lintner J. 1956. Distribution of incomes of corporations among dividends, retained earnings, and taxes. *American Economic Review* **46**(2): 97–113.
Mancinelli L, Ozkan A. 2006. Ownership structure and dividend policy: evidence from Italian firms. *The European Journal of Finance* **12**(3): 265–282.
Manos R, Murinde V, Green CJ. 2012. Dividend policy and business groups: evidence from Indian firms. *International Review of Economics and Finance* **21**(1): 42–56.
Marsh TA, Merton RC. 1987. Dividend behavior for the aggregate stock market. *Journal of Business* **60**(1): 1–40.
Michel A. 1979. Industry influence on dividend policy. *Financial Management* **8**(3): 22–26.
Miller MH, Modigliani F. 1961. Dividend policy, growth and the valuation of shares. *Journal of Business* **34**(4): 411–433.
Mukherjee T. 2009. Agency costs and the free cash flow hypothesis. In Dividends and Dividend Policy, Baker KH (ed.). John Wiley & Sons, Inc.: Hoboken, NJ; 145–161.
Nam J, Wang J, Zhang G. 2010. The impact of the dividend tax cut and managerial stock holdings on corporate dividend policy. *Global Finance Journal* **21**(3): 275–292.
Ngai EWT, Hu Y, Wong YH, Chen Y, Sun X. 2011. The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. *Decision Support Systems* **50**(3): 559–569.

DOI: 10.1002/isaf

Omran M, Pointon J. 2004. Dividend policy, trading characteristics and share prices: empirical evidence from Egyptian firms. *International Journal of Theoretical and Applied Finance* **7**(2): 121–133.
Ooi J 2001. Dividend payout characteristics of U.K. property companies. *Journal of Real Estate Portfolio Management* **7**(2): 133–142.
Quinlan JR. 1986. Induction of decision trees. *Machine Learning* **1**(1): 81–106.
Quinlan JR. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers: San Mateo, CA.
Renneboog L, Trojanowski G. 2011. Patterns in payout policy and payout channel choice. *Journal of Banking & Finance* **35**(6): 1477–1490.
Saadi S, Dutta S. 2009. Taxes and clientele effects. In Dividends and Dividend Policy, Baker KH (ed.). John Wiley & Sons, Inc.: Hoboken, NJ; 127–144.
Schellenger MH, Wood DD, Tashakori A. 1989. Board of director composition, shareholder wealth, and dividend policy. *Journal of Management* **15**(3): 457–467.
Seng J-L, Chen TC. 2010. An analytic approach to select data mining for business decision. *Expert Systems with Applications* **37**(12): 8042–8057.
Shabibi BKA, Ramesh G. 2011. An empirical study on the determinants of dividend policy in the UK. *International Research Journal of Finance and Economics* (**80**): 105–120.
Sharma A, Panigrahi PK. 2012. A review of financial accounting fraud detection based on data mining techniques. *International Journal of Computer Applications* **39**(1): 37–47.
Shefrin H. 2009. Behavioral explanations of dividends. In Dividends and Dividend Policy, Baker KH (ed.). John Wiley & Sons, Inc.: Hoboken, NJ; 179–199.
Shmueli G, Patel NR, Bruce PC. 2007. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel® with XLMiner®. John Wiley & Sons, Inc.: Hoboken, NJ.
Short H, Zhang H, Keasey K. 2002. The link between dividend policy and institutional ownership. *Journal of Corporate Finance* **8**(2): 105–122.
Smith DM. 2009. Residual dividend policy. In Dividends and Dividend Policy, Baker KH (ed.). John Wiley & Sons, Inc.: Hoboken, NJ; 115–126.
Tan P-N, Steinbach M, Kumar V. 2006. Introduction to Data Mining. Pearson/Addison-Wesley: Boston, MA.
Tsai C-F 2008. Financial decision support using neural networks and support vector machines. *Expert Systems* **25**(4): 380–393.
Vellido A, Lisboa PJG, Vaughan J. 1999. Neural networks in business: a survey of applications (1992–1998). *Expert Systems with Applications* **17**(1): 51–70.
Witten IH, Frank E. 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann Publishers: San Francisco, CA.
Won C, Kim J, Bae JK. 2012. Using genetic algorithm based knowledge refinement model for dividend policy forecasting. *Expert Systems with Applications* **39**(18): 13472–13479.
Wong BK, Selvi Y. 1998. Neural network applications in finance: A review and analysis of literature (1990–1996). *Information Management* **34**(3): 129–139.
Wong BK, Lai VS, Lam J. 2000. A bibliography of neural network business applications research: 1994–1998. *Computers and Operations Research* **27**(11–12): 1045–1076.
Zhang D, Zhou L. 2004. Discovering golden nuggets: data mining in financial application. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* **34**(4): 513–522.

DOI: 10.1002/isaf