Comput Econ (2010) 36:133–151
DOI 10.1007/s10614-010-9205-3

Using Boosting for Financial Analysis and Performance Prediction: Application to S&P 500 Companies, Latin American ADRs and Banks

Germán Creamer · Yoav Freund

Received: 4 January 2009 / Accepted: 1 March 2010 / Published online: 5 May 2010

Abstract This paper demonstrates how the boosting approach can support the financial analysis functions in two ways: (1) As a predictive tool to forecast corporate performance, and rank accounting and corporate variables according to their impact on performance, and (2) As an interpretative tool to generate alternating decision trees that capture the non-linear relationship among accounting and corporate governance variables that determine performance. We compare our results using Adaboost with logistic regression, bagging, and random forests. We conduct 10-fold cross-validation experiments on one sample each of S&P 500 companies, American Depository Receipts (ADRs) of Latin American companies and Latin American banks. Adaboost results indicate that large companies perform better than small companies, especially when these companies have a limited long-term assets to sales ratio. Performance improves for large LAADR companies when the country of residence is characterized by a weak rule of law. In the case of S&P 500 companies, performance increases when the compensation for top officers is mostly variable.

Keywords Financial analysis · Machine learning · Adaboost · Data mining

This paper is based on an earlier work: Predicting Performance and Quantifying Corporate Governance Risk for Latin American ADRs and Banks. In Proceedings of the Financial Engineering and Applications conference, MIT-Cambridge, 2004.

G. Creamer (✉)
Stevens Institute of Technology, Castle Point on Hudson, Hoboken, NJ 07030, USA
e-mail: gcreamer@stevens.edu

Y. Freund
Computer Science Department, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0114, USA
e-mail: yfreund@cs.ucsd.edu

1 Introduction

A major task for financial analysts is how to merge accounting and corporate governance variables and show how they interact in order to justify the predictions of their financial models. They use accounting models demonstrated to be good predictors of corporate performance such as those proposed by Jegadeesh et al. (2004). Additionally, financial analysts complete the study with their own interpretation of additional factors such as the effects of new product development and changes in industry and corporate governance variables. Analysts must be highly specialized in each industry in order to interpret the interaction of all variables. The main drawback of this approach is that specialized training of financial personnel is a very expensive process. In the case of emerging markets the situation is even worse because of the limited number of companies, restricted information, and different corporate governance systems. Therefore, financial analysts will benefit from a method that employs accounting and non-accounting variables to predict corporate performance and discover relationships among these variables.

Previous studies on international and US securities (see pioneering work of Altman 1968) have employed linear discriminant analysis, regression analysis or logistic regression in evaluating financial distress, bankruptcy, credit risk, and bond and loan performance usually through accounting variables. These studies are based on estimating the parameters of an underlying stochastic system which is generally assumed to be a linear system. A major limitation of this methodology is that the stochastic system and the interactions among variables and non-linearities have to be incorporated manually.

In contrast, machine learning methods such as boosting and support vector machine avoid the question of modeling the underlying distribution and focus on making accurate predictions for some variables given other variables. Breiman (2001b) contrasts these two approaches respectively as the data modeling culture and the algorithmic modeling culture. According to Breiman (2001b), while most statisticians adhere to the data-modeling approach, people in other fields of science and engineering use algorithmic modeling to construct predictors with superior accuracy. For Breiman, the main drawback of algorithmic modeling is that generated representations are hard to *interpret*.

The objective of this paper is to demonstrate how financial analysis can be conducted using an adapted version of the boosting approach as a predictive and interpretative tool for corporate performance. We conduct our analysis with samples of diverse size and geographic regions: S&P 500 companies, American Depository Receipts (ADRs) of Latin American companies, and banks domiciled in Latin American countries.

We create a predictive model for evaluating whether a company's performance or a bank's efficiency is above or below par as a function of the main corporate governance factors and selected accounting ratios known to be important in evaluating corporate performance. We use Adaboost (Freund and Schapire 1997) as the learning meta-algorithm of our predictive model, and compare our results with logistic regression, random forest, and bagging.

The rest of the paper is organized as follows: Sect. 2 presents the predictive methods used in this paper; Sect. 3 describes the performance variables; Sect. 4 explains our

experiments in detail; Sect. 5 presents the results of our forecast, and Sect. 6 presents the conclusions.

2 Learning Methods

This section introduces the main learning methods used in this paper. The most important learning meta-algorithm in this research is Adaboost. We have also compared our results with other methods such as logistic regression (Hastie et al. 2003), random forest (Breiman 2001a), and bagging (Breiman 1996).

2.1 Boosting

Adaboost is a general discriminative meta-learning algorithm invented by Freund and Schapire (1997). The basic idea of Adaboost is to repeatedly apply a simple learning algorithm, called the *weak* or *base* learner[1], to different weightings of the same training set. In its simplest form, Adaboost is intended for binary prediction problems where the training set consists of pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$, x_i corresponds to the features of an example, and $y_i \in \{-1, +1\}$ is the binary label to be predicted. A *weighting* of the training examples is an assignment of a non-negative real value w_i to each example (x_i, y_i).

On iteration t of the boosting process, the weak learner is applied to the training sample with a set of weights $w_1^t, \ldots, w_m^t$ and produces a prediction rule h_t that maps x to $\{0, 1\}$.[2] The requirement on the weak learner is for $h_t(x)$ to have a small but significant correlation with the example labels y when measured using the *current weighting of the examples*. After the rule h_t is generated, the example weights are changed so that the weak predictions $h_t(x)$ and the labels y are decorrelated. The weak learner is then called with the new weights over the training examples, and the process repeats. Finally, all of the weak prediction rules are combined into a single *strong* rule using a weighted majority vote. One can prove that if the rules generated in the iterations are all slightly correlated with the label, the strong rule will have a very high correlation with the label and will predict it very accurately. For an introduction to boosting see Schapire (2002).

The whole process can be seen as a variational method in which an approximation $F(x)$ is repeatedly changed by adding small corrections given by the weak prediction functions. In Fig. 1, we describe Adaboost in these terms. We shall refer to $F(x)$ as the *prediction score* in the rest of the document. The strong prediction rule learned by Adaboost is $\text{sign}(F(x))$.

A surprising phenomenon associated with Adaboost is that the test error of the strong rule (percentage of mistakes made on new examples) often continues to decrease even after the training error (fraction of mistakes made on the training set) reaches zero. This behavior has been related to the concept of a "margin", which is simply the value

[1] Intuitively, weak learner is an algorithm with a performance at least slightly better than random guessing.

[2] Mapping x to $\{0, 1\}$ instead of $\{-1, +1\}$ increases the flexibility of the weak learner. Zero can be interpreted as "no prediction" (Freund and Schapire 1997).

$$
\begin{array}{l}
F_0(x) \equiv 0 \\
\text{for } t = 1 \ldots T \\
\qquad w_i^t = e^{-y_i F_{t-1}(x_i)} \\
\qquad \text{Get } h_t \text{ from } \textit{weak learner} \\
\qquad \alpha_t = \frac{1}{2} \ln \left(\frac{\sum_{i:h_t(x_i)=1, y_i=1} w_i^t}{\sum_{i:h_t(x_i)=1, y_i=-1} w_i^t} \right) \\
\qquad F_{t+1} = F_t + \alpha_t h_t
\end{array}
$$

Fig. 1 The Adaboost algorithm (Freund and Schapire 1997). y_i is the binary label to be predicted; x_i corresponds to the features of an instance i, w_i^t is the weight of instance i at time t, h_t and $F_t(x)$ are the prediction rule and the prediction score at time t respectively

$yF(x)$ (Schapire et al. 1998). While $yF(x) > 0$ corresponds to a correct prediction, $yF(x) > a > 0$ corresponds to a *confident* correct prediction, and the confidence increases monotonically with a.

2.2 Alternating Decision Trees

One successful and popular way of using boosting is to combine it with decision tree learning as the base learning algorithm (Friedman et al. 2000). We use boosting to learn the decision rules constituting the tree and to combine these rules through a weighted majority vote. The form of the generated decision rules is called an *alternating decision tree* (ADT) (Freund and Mason 1999). In ADTs each node can be understood in isolation.

We explain the structure of ADTs using Fig. 2. The problem domain is corporate performance prediction, and the goal is to separate stocks with high and low values based on 17 different variables. The tree consists of alternating levels of ovals (*prediction nodes*) and rectangles (*splitter nodes*). The first number within the ovals defines contributions to the prediction score, and the second number (between parentheses) indicates the number of instances. In this example, positive contributions are evidence of high performance, while negative contributions indicate corporate financial problems. To evaluate the prediction for a particular company we start at the top oval (0.04) and follow the arrows down. We follow *all* the dotted arrows that emanate from prediction nodes, but *only one* of the solid-line arrows emanating from a splitter node, corresponding to the answer (yes or no) to the condition stated in a rectangle. We sum the values in all the prediction nodes that we reach. This sum represents the prediction score $F(x)$ above, and its sign is the final, or probable, prediction. For example, suppose we had a company for which LnMarketCap $= 6$, KS $= 0.86$, RuleOfLaw $= 7.02$, and PartOutBOD $= 0.76$. In this case, the prediction nodes that we reach in the tree are 0.042, -0.7181, 0.583, and 1.027. Summing gives a score of 0.9339, i.e., a very confident indicator that the company has a high market value.

This example demonstrates how alternating decision trees combine the contributions of many indicators to generate a prediction. The ADT in the above figure was generated by Adaboost from training data. In Adaboost's terms, each prediction node represents a weak prediction rule, and at every boosting iteration, a new splitter node

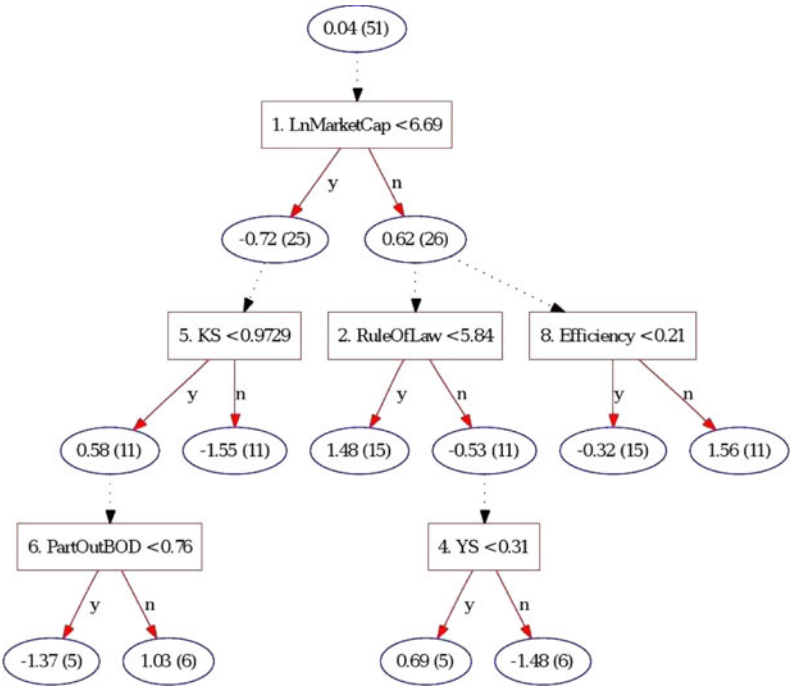

Fig. 2 Representative ADT of LAADR. The tree has ovals (*prediction nodes*) and rectangles (*splitter nodes*). The first number within the ovals defines contributions to the prediction score, and the second number (between parentheses) indicates the number of instances. The root node has the minimum prediction score for all cases. The sum of first values in all relevant prediction nodes, including the root, is the prediction score. The sign of prediction score is the prediction. The representative ADT is calculated selecting nodes present in at least 60% of the trees obtained from 10-fold cross-validation

together with its two prediction nodes is added to the model. The splitter node can be connected to any previous prediction node, not only leaf nodes, unless a splitter node is already attached. Each prediction node is associated with a weight α that contributes to the prediction score of every example reaching it. The weak hypothesis $h(x)$ is 1 for every example reaching the prediction node and 0 for all others. The value of α that corresponds to this weak rule is the value in the prediction node. The number in front of the conditions in the splitter nodes of Fig. 2 indicates the iteration number on which the node was added. In general, lower iteration numbers indicate that the decision rule is more important.

A drawback of Adaboost is its characteristic instability. A method that calculates the average of the results of Adaboost across different samples may reduce its variance. This is the foundation of bagging (Breiman 1996), algorithm that we also test in this paper.

We chose Adaboost as our discriminative learning algorithm because of its feature selection capability, its error bound proofs (Freund and Schapire 1997), its interpretability, and its capacity to combine quantitative, and qualitative variables.

3 Measures of Company Performance

We use Tobin's Q as a performance measure of the value of intangibles or the real value created by management for ADRs and S&P 500 companies. Tobin's Q is the ratio of the market value of assets to the replacement cost of assets.[3] A higher value of Tobin's Q indicates that more value has been added or there is an expectation of greater future cash flow. Hence, the impact of management quality on performance is captured by Tobin's Q. Any difference of Tobin's Q from one indicates that the market perceives the value of total assets to be different from the value required to replace their physical assets. The value of internal organization, management quality, or expected agency costs is assumed to explain the difference. Values of Tobin's Q above one indicate that the market perceives the firm's internal organization as effective in leveraging company assets, while a Tobin's Q below one shows that the market expects high agency costs. We use as a proxy for Tobin's Q the ratio of book value of debt plus market value of common stocks, and preferred stocks to total assets.[4]

For Latin American banks, we use an efficiency measure based on data envelopment analysis (DEA) instead of Tobin's Q because some of the banks under study are not public companies or participate in very illiquid markets. This efficiency measure is also an indicator of agency costs to the firm. Conflicts between managers and shareholders may arise when operating costs increase in relation to a fixed output. DEA, using linear programming, builds a frontier selecting the "best practice" firms, obtains an efficient score, and recognizes overuse of inputs or underproduction of outputs.

We calculate efficiency for the Latin American banking sector using the DEA with output-oriented constant returns to scale as our measure of banking efficiency (see Charnes et al. 1978; Lovell 1993). As input, we use interest-paying deposits, and non-interest expenses which may include personnel, administrative costs, commissions, and other non-interest operating costs. As output, we use total income which includes both interest and non-interest income. Banks are ranked according to this measure country by country.

4 Experiments

Adaboost is used (see Sect. 2.1) to classify stocks above and below the median of Tobin's Q for LAADR and S&P 500 companies, and the DEA technical efficiency indicator for LABANKS (see Sect. 3). As independent variables we used the accounting and corporate governance variables that we introduced in Appendix 1. The data we used in our experiments are from (1) Latin American ADRs (LAADR), (2) Latin American banks (LABANKS), and (3) S&P 500 companies. These data are described

[3] The intangibles can also refer to other factors such as intellectual capital or the value of information technology. In this research we control for differences among countries, and economic sectors where companies may have similar technology. We therefore assume that Tobin's Q reflects management quality. The discrimination between the contribution to performance of top management and other intangible assets such as intellectual capital requires a more detailed analysis.

[4] This proxy is empirically close to the well-known Lindenberg and Ross (1981) proxy. For international stocks, the information to calculate the Lindenberg and Ross proxy is very limited.

in Appendix 2. In the LAADR sample, the median of the Tobin's Q is very close to one. For LAADR, the results can be interpreted as the classification between those stocks with a market value of its assets above (Tobin's Q greater than one) or below (Tobin's Q smaller than one) its costs of replacement. For S&P 500 companies, the classification is between companies with positive scores that have high Tobin's Q, and companies with negative scores have low Tobin's Q. For LABANKS, the classification is between more efficient and less efficient banks. We calculated the efficiency indicators for each country because of the differences between their accounting systems. The banks' efficiency is calculated in relation to their peers in their respective countries.

The results of ADTs must be interpreted as companies with positive scores and high Tobin's Q and banks as efficient institutions, or companies with negative scores and low Tobin's Q and inefficient banks.

We eliminated variables that indicated multicollinearity using the variance inflation factor (VIF). In general, variables with large VIF (larger than 10) were removed. For LABANKS the eliminated variables were risk of contract repudiation, legal system, region, corruption, and debt ratio. For LAADR, we eliminated risk of expropriation, risk of contract repudiation, and region. For S&P 500 companies, we eliminated total compensation of officers, and CEOs.

We performed 10-fold cross-validation experiments to evaluate classification performance on held-out experiments using Adaboost. For LAADR and LABANKS we ran our experiments with 10 iterations. For S&P 500 companies, we run 300 iterations. We used the MLJAVA package, which implements the alternating decision tree algorithm described in Freund and Mason (1999).[5] The variables were ranked as an average of the iteration when each is selected and weighted by their frequency.

A drawback of Adaboost is that a small variation of the training set may lead to a different ADT, especially in the case of a small sample. As a result, our 10-fold cross-validation procedure generated ten different ADTs. To interpret the results of Adaboost, we had to have a representative ADT for each group of companies (ADRs, banks and S&P 500 companies). Considering that LAADR and LABANKS have only 10 iterations, we selected this representative ADT for having the lowest test error and the most important ranked variables located in the same nodes for at least 60% of the trees. For the case of the ADTs of the S&P 500 companies that have 300 nodes (one for each iteration), we obtained our representative ADT using the same algorithm described above and formalized by Creamer and Freund (2010).

To check for the possibility that the Adaboost results could be improved because of the characteristic instability of Adaboost, we ran bagging on top of Adaboost (bagged boosting). We created ten folds for testing and training and obtained 100 bootstrap replicates of each testing fold. The score of the bootstraps of each fold was averaged to get the estimated class. Finally, we averaged the test error of the ten folds. We also compared ADTs with a single decision tree classifier and a decision stumps classifier trained using boosting. We evaluated the differences between the average of the test error of Adaboost with the test errors of the rest of the learning algorithms using

[5] The functionality of MLJava is currently included in JBoost ⟨http://jboost.sourceforge.net⟩ which is an implementation of boosting in Java.

Table 1 Test errors and standard deviations of learning algorithms when all variables are included

	LAADR		LABANKS		S&P 500	
	Test error	St. dev.	Test error	St. dev.	Test error	St. dev.
Adaboost	14.0%	16.5%	17.8%	9.4%	16.1%	2.0%
Single tree	16.0%	12.7%	17.8%	11.9%	20.4% **	4.2%
Stumps	32.0%	19.3%	13.3%	11.5%	16.8%	2.1%
Bagged boosting	22.0%	23.9%	13.33%	8.66%	14.01% **	0.50%
Random forests	32.0%*	16.87%	16.67%	17.57%	11.50% **	4.56%
Logistic regression	23.3%	20.2%	20.1%	13.2%	16.9%	2.9%
Number observations	51		104		2278	

* 5%, ** 1% significance level of t-test difference between test errors of algorithms and Adabooost

the 10-fold cross-validated t-test with unequal variances as described by Dietterich (1998).

To evaluate the difficulty of the classification task, we compared our method, Adaboost, with multiple logistic regression and random forests (see Hastie et al. 2003; Breiman 2001a) using the softwares Weka (Witten and Frank 2005) and Random Forests V5.0 respectively.[6] We ran our random forests experiments with 1,000 trees. We also used four variables for LAADR and LABANKS, and eight for S&P 500 companies randomly selected at each node in order to reduce the test error. Our logistic regression analysis uses Tobin's Q as the dependent variable for LAADR and S&P 500 companies, and the DEA technical efficiency indicator as the dependent variable for LABANKS (see Sect. 3). Besides the independent variables introduced in Appendix 1, the multiple logistic regression also includes indicator variables for industrial sectors.

5 Results

The results of the test errors for the learning algorithms used are shown in Table 1. As both our LAADR and LABANKS datasets are very small (51 examples in LAADR and 104 examples in LABANKS), evaluating the statistical significance of the different models and the comparison of their test errors is difficult. Acknowledging these limitations, we present the results of the t-test: there is a significant difference between the test errors of Adaboost and random forests for LAADR, while there are no differences of the test errors for the rest of the tests in both samples. In the case of S&P 500 companies, the test errors of single tree, bagged boosting, and random forests show a significant difference with Adaboost. Random forests presents the lowest test error for S&P 500 companies, and it is followed by bagged boosting.

[6] A working version of Random Forests V5.0 can be obtained from ⟨http://stat-www.berkeley.edu/users/breiman/RandomForests/⟩.

Table 2 indicates the importance of each variable according to Adaboost and random forests. The results of both algorithms coincide in terms of what the most important variables are.[7] For the LAADR dataset the relevant variables are market capitalization, law and order tradition, outsider participation on the board of directors, and operating expenses to sales ratio. For the LABANKS dataset the relevant variables are long-term assets to deposits ratio, equity index, risk of confiscation, and number of directors. For the S&P 500 dataset the most important variables are operating expenses to sales ratio, operating income to sales ratio, capital expenditures to long-term assets, and long-term assets to sales ratio.

For some variables, there is an important discrepancy among boosting and random forests. In the case of the LAADR dataset, capital expenditures to sales ratio is considered the third and tenth most important variable according to random forests and boosting respectively, while for the LABANKS dataset the efficiency of the legal system is the variable that shows an important difference. For the S&P 500 companies, boosting and random forests have a very similar ranking.

The results of bagged boosting cannot be interpreted in terms of the impact of each variable on performance and efficiency because of the large number of trees generated.

In the case of LABANKS, four variables chosen by Adaboost are ranked among the top five variables according to random forests. Considering the similarity of the most important variables selected by random forests and Adaboost, we discuss the ADTs. The odds ratios of logistic regression also confirm the importance established by Adaboost and random forests of the following variables: long-term assets to sales ratio and corruption for LAADRs; long-term assets to deposits ratio, insider ownership, and risk of expropriation for LABANKS; and long-term assets to sales ratio, and debt ratio for S&P 500 companies.

In synthesis, in most of the cases Adaboost performed in a similar way to other learning algorithms such as bagging and random forests, and had the capacity to generate a score that evaluated the effect of corporate governance variables on performance. Additionally, Adaboost also allowed us to interpret the results because of the limited number of trees that were generated in contrast to the requirements of the other methods such as random forests.

As an illustration of how to use boosting as an interpretative tool, we discuss the implications of the representative ADT for LAADRs (Fig. 2) and for sectors 1 and 2 of S&P 500 companies (Fig. 3). According to Fig. 2, ADRs with market capitalization around or above the median perform better than the rest. Large companies in emerging markets are likely to be oligopolies or monopolies in their area of activity. However, the performance of LAADR improves in countries with a weak rule of law *(2. RuleOfLaw)*.[8] Large Latin American companies probably perform better in environments with a weak legal structure because of the close family relationships that help them to influence government decisions in their favor. The benefits of these government private sector connections are less important for small size companies *(1. LnMarketCap)*.

[7] We accept a difference of two in the ranking between both algorithms.

[8] We refer to a specific node of any ADT with its iteration number (first number of the node) and its variable name in italics.

Table 2 Results for LAADR, LABANKS and S&P 500 companies

	LAADR				LABanks				S & P 500			
	Logit	Boost	RF		Logit	Boost	RF		Logit	Boost	RF	
	Odds ratios	Rank	z-score	Rank	Odds ratios	Rank	z-score	Rank	Odds ratios	Rank	z-score	Rank
LnMarketCap (Nat. log market capitalization)	0.00	1	26	1					0.36	5	83.75	2
Equity index					0.00	2	35	1				
IK (Capital expenditures/ long-term assets)		10	6	3					0.03	3	69.85	6
Efficiency (Operating expenses / sales)		6	5	4					0.00	1	80.28	3
YS (Operating income/ sales)	0.00	3							0.00	2	89.22	1
Debtratio (Debt / total assets)			2	5					11.18	6	79.65	4
KS (L.T. assets/sales)	53157	4							4.26	4	70.43	5
KD (L.T. assets/deposits)					2E+12	1	33	2				
EfficiencyJudicialSystem (Effic. legal system)		*9*			*0.46*	*9*	*12*	*4*				
RuleOfLaw (Law and order tradition)	*0.03*	*2*	*6*	*2*	*0.63*	*7*	*8*	*6*				
Corruption (Level of government corruption)	*82222*											
RiskOfExpropiation (Risk confiscation)					*0.47*	*3*	*10*	*5*				
PartOutBOD (% outsiders as directors)	*0.00*	*5*	*2*	*6*	*2.53*	*7*	*3*	*9*				
Avg Participation					*0.16*	*6*	*4*	*8*				
LnDir (Natural log number directors)	*1E+11*	*7*			*0.41*	*5*	*12*	*3*				
InstPart (% institutional equity ownership)	*27.76*	*8*										
T_Insiders (% insider's equity ownership)			*0.02*	*7*	*0.53*	*4*	*4*	*7*	*1.15*		*32.34*	*10*
TotalMeetingPay (Total payment per meeting)									*1.03*		*22.87*	*17*
TotalCompExec (Total compensation)										*9*	*32.04*	*11*
OptionStockValueExec (Value stock option)									*1.00*		*25.12*	*15*
PayDirectors (Annual cash pay to directors)									*1.04*	*8*	*52.66*	*8*

Table 2 continued

	LAADR				LABanks				S & P 500			
	Logit	Boost		RF	Logit	Boost		RF	Logit	Boost		RF
	Odds ratios	Rank	z-score	Rank	Odds ratios	Rank	z-score	Rank	Odds ratios	Rank	z-score	Rank
OptionAllValExec (Total value options)									*1.00*		*31.45*	*13*
OptionsDirectors (Number options directors)									*1.00*		*32.54*	*9*
StockDirectors (Number of stocks directors)									*1.15*	*7*	*31.76*	*12*
TotalCompCEO (Total compensation CEO)											*30.06*	*14*
TotalValOptCEO (Total value options CEO)									*1*		*24.84*	*16*

This table reports statistics and results of predicting Tobin's Q for LAADR and for S&P 500 companies, and efficiency for LABANKS using logistic regression, Adaboost, and random forest. Country corporate governance variables are from La-Porta et al. (1998). *RF*: Random forests. z-score for random forests (Breiman 2001b) is the raw importance score divided by standard deviation. Logistic regression includes indicator variables to control for sector, although they are not included in the table. Variables that do not show any relevance are not included such as legal system, accounting, number of insiders in board of directors, and chairman as CEO. Corporate governance variables are in italics

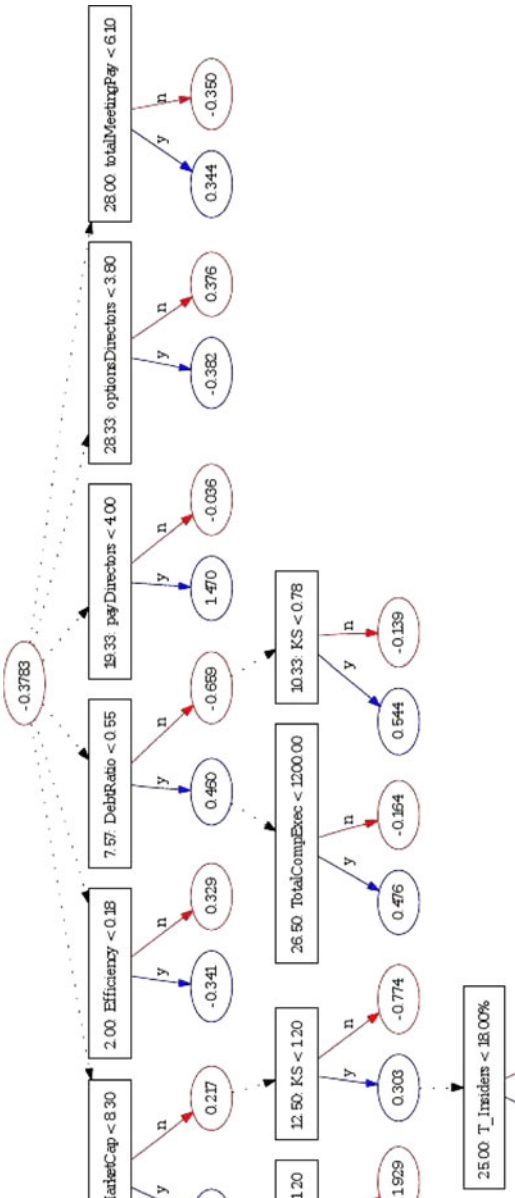

Fig. 3 S&P 500: representative ADTs for sectors 1 and 2. First number in each rectangle is average iteration. Creamer and Freund (2010) describes the algorithm to calculate this representative ADTs

In countries with a strong legal system, large companies may still have an important agency conflict that affects their performance if the cash available for operations is too high, as a large operating income to sales ratio *(4. YS)* indicates. An excessive amount of cash may allow managers to spend reserves on projects that benefit them directly instead of increasing the value of their companies. A very low operating expenses to sales ratio *(8. Efficiency)* may indicate future operational problems. Among the medium and large companies, 58% have a low efficiency ratio in relation to the threshold level found by Adaboost. The performance of small and medium size companies improves if the proportion of long-term assets to sales *(5. KS)* is low (below the median) for LAADR companies. These indicators are important for revealing agency problems. The long-term assets are easy to monitor, and can become collateral to finance new projects. However, if the level of long-term assets is too high, it may indicate inefficiency and overspending. The composition of the board of directors is important for smaller companies which have a long-term assets to sales ratio *(5. KS)* below 0.97. In these cases, the statistical model provides a hint that a small participation by insiders or a large participation by outsiders on the board of directors of LAADR companies *(6. PartOutBOD)* is connected (not necessarily causally) with better performance.

In relation to S&P 500 companies of sectors 1 and 2, those that have a maximum debt to total assets ratio *(7.57 DebtRatio)* of 0.55 (about the median) show a superior performance. Additionally, among these high performance companies, those that show a better performance have a limit of USD 1.2 million (below the first quartile) as the total compensation for top officers *(26.5 TotalCompExec)*. In the case of a company that is highly leveraged, performance may improve if the long-term assets to sales ratio *(10.33 KS)* is less than 0.78 (about the first quartile). In this segment of companies, size matters. Large companies have an advantage as long as the long-term assets to sales ratio *(12.5 KS)* is less than 1.2 (similar to the mean), and there is at least 18% of insiders *(25. T_Insiders)* in the board of directors. Smaller companies show an improvement in their performance when the number of stocks granted to directors *(12.5 stockDirectors)* is restricted to 1,200. Additionally, reducing the fixed compensation for directors *(19.33 payDirectors)* and increasing their variable compensation through options *(28.33 optionsDirectors)* have a positive effect in the company.

6 Final Comments and Conclusions

This research shows that Adaboost can facilitate the financial analysis functions in two ways:

1. As a predictive tool to forecast corporate performance and to rank accounting and corporate variables according to their impact on performance. In this respect, in most of the cases Adaboost performs better or as well as other learning algorithms with samples of different sizes and geographic regions (Latin America or US market).
2. As an interpretative tool, Adaboost can generate representative ADTs that simulate the non-linear relationship among the accounting and corporate governance variables that determine performance. These ADTs also segment the corporate

sample in subsets that are affected by similar factors. This capability has a great importance for risk studies that require segmentations of the companies under study in similar subsets, or for the exploration of new markets as in the case of emerging markets.

Adaboost results indicate that large companies perform better than small companies, especially when these companies have a limited long-term assets to sales ratio. Performance improves for large LAADR companies when the country of residence is characterized by a weak rule of law. In the case of S&P 500 companies, performance increases when the compensation for top officers is mostly variable.

The use of ADTs in finance requires large time-series or cross-sectional datasets in order to calculate meaningful nodes. Indicators that do not have enough information cannot be quantified using ADTs. As this research shows, Adaboost also works adequately with small datasets (LAADR and LABANKS) and is able to identify the most important variables that affect performance. However, the variance of the test error increases as the size of the dataset decreases. Considering this shortcoming, we extend our analysis of Latin American companies to S&P 500 companies and show how the boosting approach, with either small or large samples, select a combination of accounting and corporate governance variables that determine performance. We suggest that companies employing Adaboost as an interpretative tool use large datasets (industrial surveys or compensation surveys) or build their own internal dataset using the company's historical information.

Comparative regional studies always have a major problem in terms of how to integrate data coming from different sources, and generally with different standards. We saw that this problem was implicit in the LABANKS dataset. We believe the research of emerging markets can be improved by enlarging the dataset and running the learning algorithms in subsets aggregated by regions or corporate governance systems.

Acknowledgments Authors thank the editor Hans M. Amman, David Waltz, Tony Jebara, Sal Stolfo, Vasant Dhar, Salvatore Cantale, Paul Spindt, Tom Noe, Tom Reese, Kenneth Jameson, Greg Buchholz, John M. Trapani, and participants of the 2004 IASTED Financial Engineering and Applications conference, 2003 Latin American Studies Association meeting, 2001 Tulane University Latin American Research Consortium meeting, and 2000 Eastern Finance Association meeting for their helpful comments on partial versions of this paper, and to Patrick Jardine for proof-reading the article. GC also thanks the institutional support of Stevens Institute of Technology, Tulane University, the Board of Regents of the State of Louisiana, and the Center for Computational Learning Systems at Columbia University, and research support of Carolina Gomez, Monica Garcia, Leonardo Serrano, Marcelo Gonzalez, Juan Carlos Otalora, and Julian Benavides. The opinions presented are the exclusive responsibility of the authors.

Appendix 1: Independent Variables

See Table 3.

Table 3 Variables used for corporate governance experiments

Indicator	Definition	Type of companies
TobinQ	Tobin 's Q, which is the ratio of the market value to the replacement cost of assets. We use a proxy for Tobin's Q as the ratio of book value of debt plus market value of common stocks and preferred stocks to total assets	LAADR,S&P 500
PartOutBOD	% outsiders on the board of directors	LAADR,S&P 500, LABANKS
LnDIR	Natural logarithm of board size	LAADR LABANKS
InstPart	% institutional ownership	LAADR LABANKS
T_Insider	% insiders' ownership. In the case of LAADR and the Latin American banks, insider ownership is defined as ownership of a company by the CEO, managers, or relatives of the CEO, and members of the board of directors.	LAADR, S&P 500, LABANKS
ChairmanCEO	1 if CEO is chairman, 0 otherwise	LAADR LABANKS
TotalCompCEO	Total compensation for CEOs. It includes the same items as TotalCompExec (thousands of dollars).	S&P 500
TotalValOptCEO	Value of options for CEOs (thousands of dollars).	S&P 500
TotalCompExec	Total compensation for officers. It includes the following items: salary, bonus, other annual, total value of restricted stock granted, total value of stock options granted (using Black-Scholes), long-term incentive payouts, and all other total (thousands of dollars).	S&P 500
OptionStockValueExec	Value of stock options granted to the executive during the year as valued using S&P's Black Scholes methodology (thousands of dollars).	S&P 500
OptionAllValExec	The aggregate value of all options granted to the executive during the year as valued by the company (thousands of dollars).	
DexecDir	Dummy variable to indicate if officer was also a director for the reference year.	S&P 500
OptionsDirectors	Number of options and additional options granted to each non-employee director during the year (thousands).	S&P 500
StockDirectors	Stock shares (including restricted stock) granted to each non-employee director (thousands).	S&P 500
PayDirectors	Annual cash retained paid to each director (thousands of dollars).	S&P 500
TotalMeetingPay	Fees paid for attendance to board of directors meeting (thousands of dollars).	S&P 500
DcommFee	Dummy variable to indicate if directors are paid additional fees for attending board committee meetings.	S&P 500
SPindex	Standard and Poor's index membership. It indicates if companies are part of S&P500 (SP), S&P midcap index (MD), S&P smallcap index (SM), or is not part of a major US index (EX).	S&P 500

Table 3 continued

Indicator	Definition	Type of companies
LnMarketCap	Natural logarithm of market capitalization, used to measure firm size	LAADR, S&P 500
KS or KD	Ratio of long term assets (property, plant and equipment) to sales (KS) for LAADRs and S&P 500 companies, and to deposits (KS) for LABANKS. This ratio is considered for its effect in the reduction of the agency conflict because these assets can be monitored very easily and they can become collateral for the development of new projects.	LAADR, S&P 500, LABANKS
YS	The ratio of operating income to sales	LAADR, S&P 500, LABANKS
DebtRatio	The ratio of debt to total assets, used as a capital structure variable. Emerging markets are much less liquid than those of developed countries. Hence, firms may give more importance to debt, rather than equity, as a source of capital.	LAADR, S&P 500, LABANKS
Equity index	Index of equity according to country of residence. This is a measure of size applied to LABANKS.	LABANKS
Efficiency	The ratio of operating expenses to sales. This is the efficiency ratio and works as a proxy for market power. It also indicates cash flow available for management use. Similarly, this efficiency ratio may also reveal agency costs or agency conflicts. (This is different from the DEA technical efficiency indicator).	LAADR, S&P 500, LABANKS
IK	The ratio of capital expenditures to long term assets (stocks of property, plant and equipment)	LAADR, S&P 500, LABANKS
AvgParticipation	Measure of ownership concentration. This is calculated as the average of the participation of the three largest shareholders per firm	LAADR,LABANKS
English	If the firm is domiciled in a country whose legal regime is part of the common law or English law legal family according to La Porta et al. (1998)	LAADR,LABANKS
French	If the firm is domiciled in a country that is part of the Napoleonic or French legal family according to La Porta et al.	LAADR,LABANKS
RuleOfLaw	Law and order tradition according to the agency International Country Risk (ICR). Scores are from 0 to 10. Lower values indicate that a country is characterized by less tradition of law and order.	LAADR,LABANKS
Corruption	Indicator of level of government corruption according to ICR. Low levels indicate higher corruption, such as solicitation of bribery by government officials	LAADR,LABANKS
EfficiencyJudicialSystem	Index about the level of efficiency of the legal system according to the agency Business International Corp. Scale is from zero to ten. Lower values correspond to lower efficiency levels.	LAADR,LABANKS
RiskOfExpropriation	Risk of confiscation or nationalization according to ICR. Scale is from zero to ten. Lower values imply higher risks.	LAADR,LABANKS

Table 3 continued

Indicator	Definition	Type of companies
RiskOfContractRepudiation	Risk of modification of a contract by economic, social or political reasons as defined by ICR. Lower values correspond to higher risks.	LAADR,LABANKS
Accounting	Index based on 1990 annual reports according to their inclusion or omission of 90 items. These items are classified into the following categories: general information, income statements, balance sheets, fund flow statement, accounting standards, stock data and special items. For each country, a minimum of three companies were studied.	LAADR,LABANKS

Third column indicates the type of company or dataset where each variable is used: LAADR for Latin American ADRs, LABANKS for Latin American banks, and S&P 500 for S&P 500 companies. Corporate governance variables at the country level are from La-Porta et al. (1998). These variables are English, French, RuleOfLaw, Corruption, EfficiencyJudicialSystem, RiskOfExpropriation, RiskOfContractRepudiation, and Accounting

Appendix 2: Data

Our first dataset is called LAADR because it is a sample of 51 stocks domiciled in Latin America (LAADR) (Argentina, Brazil, Chile, Colombia, Peru, Mexico, and Venezuela) that have issued ADRs of level II, and III for the year 1998. Level I ADR are least restricted in their required compliance with US regulations, so we have not included them in our analysis. Level II ADRs correspond to foreign companies that list their shares on NASDAQ, AMEX, or NYSE. These companies must fully obey the registration requirements of the SEC, including compliance with US GAAP. Level III ADRs refer to foreign companies that issue new stocks directly in the United States. This means that they have the same compliance requirements as a US public company, and are therefore the most regulated. We chose ADRs from countries on the list of emerging markets database (EMDB) of the International Finance Corporation (IFC).[9]

We obtained the financial information from COMPUSTAT for the year 1998. The information on the value of market capitalization is from CRSP, and is compared with information from the NYSE. We extracted corporate governance information–such as list of directors, executives, and major shareholders–from the proxy statements published at Disclosure, Edgar, and companies' websites for the year 1998. In the case of LAADR, insider ownership is defined as ownership of a company by the CEO, managers, or relatives of the CEO, and members of the board of directors.

Our second dataset is called LABANKS because it is a list of 104 Latin American banks. LABANKS consists of banks headquartered in Argentina, Brazil, Chile, Colombia, Peru, Ecuador, and Bolivia representing about 80% of the total assets of the private sector in the major Latin American countries.[10] Corporate bank information was obtained from Internet Securities Inc., central bank, regulator, and company websites. We collected financial as well as corporate information similar to that collected for ADRs. Our sample of banks is restricted by the availability of corporate finance records. Most of the financial data is from 2000. A few companies that were merged or disappeared in 1998 were included using the financial statements of 1997. The corporate information is gathered from the period 1998–2000. Considering that the information about ownership structure is relatively stable, we do not foresee any major consistency problem.

Our third dataset is called S&P 500 because it includes the companies that are part of the S&P 500 index. The main sources of data for S&P 500 companies were ExecuComp for executive compensation information, and Compustat North America for accounting information. These two datasets are products of Standard & Poor's. We restricted our dataset to S&P 500 companies with available data from 1992 to 2004. We eliminated observations that did not have enough information to calculate Tobin's Q or incomplete executive compensation information.

[9] Standard and Poor's acquired this database in January 2000, and it became the Standard and Poor's EMDB.

[10] We were not able to include Venezuela's banks because the President of the Venezuelan Banking Association declined to supply any information to our research team, and asked member banks not to supply any corporate information to us due to the increased risk of kidnapping that its members would be subject to if this information were distributed.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance, 23*(4), 589–609.

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

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

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science, 16*, 199–231.

Charnes, A., Cooper, W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*, 429–444.

Creamer, G., & Freund, Y. (2010). Learning a board balanced scorecard to improve corporate performance. *Decision Support Systems*. doi:10.1016/j.dss.2010.04.004.

Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923.

Freund, Y., & Mason, L. (1999). The alternating decision tree learning algorithm. In *Machine learning: proceedings of the sixteenth international conference* (pp. 124–133). San Francisco: Morgan Kaufmann Publishers Inc.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics, 38*(2), 337–374.

Hastie, T., Tibishirani, R., & Friedman, J. (2003). *The elements of statistical learning*. New York: Springer.

Jegadeesh, N., Kim, J., Krische, S. D., & Lee, C. M. C. (2004). Analyzing the analysts: When do recommendations add value?. *Journal of Finance, 59*(3), 1083–1124.

La-Porta, R., de Silanes, F. L., Shleifer, A., & Vishny, R. (1998). Law and finance. *Journal of Political Economy, 196*, 1113–1155.

Lindenberg, E., & Ross, S. A. (1981). Tobin's Q ratio and industrial organization. *Journal of Business, 54*, 1–32.

Lovell, C. (1993). Production frontiers and productive efficiency. In H. Fried, C. Lovell, & S. Schmidt (Eds.), *The measurement of productive efficiency* (pp. 3–67). New York: Oxford University Press.

Schapire, R. E. (2002). The boosting approach to machine learning: An overview. In *MSRI workshop on nonlinear estimation and classification*.

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics, 26*(5), 1651–1686.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.