Understanding and Predicting Sovereign Debt Rescheduling: A Comparison of the Areas under Receiver Operating Characteristic Curves

ARNULFO RODRIGUEZ[*] AND PEDRO N. RODRIGUEZ[**]

November 15, 2003

Abstract

This paper extends the existing literature on empirical research in the field of sovereign debt. To the authors' knowledge, only one study in the area of sovereign debt has used a variety of statistical methodologies to test the reliability of their predictions and to compare their performance against one another. However, those comparisons across models have been made in terms of different probability cutoff points and mean squared errors. Moreover, the issue of interpretability has not been addressed in terms of interactions among explanatory variables with their correspondent debt rescheduling threshold level. The areas under the Receiver Operating Characteristic (ROC) curves are used to compare the discrimination power of statistical models. This paper tests Logit, MARS, Tree-based and Neural Network models. Analyses of the relative importance of variables and deviance were done. All of the models rank the previous payment history as the most important explanatory variable.

KEY WORDS: Sovereign Debt Rescheduling; Data Mining; Classification Techniques; ROC Analysis

1. Introduction

The recent Argentinean default on its sovereign debt payments is one more piece of evidence that the intertemporal government budget constraint offers poor insights into the comprehension of government debt crises and defaults. Furthermore, to the authors' knowledge, empirical studies in the area of sovereign debt have not yet derived easy to

[*]Department of Economics, University of Texas at Austin. TX 78712, Ph. (512) 587-7812, Fax (512) 471-3510, E-mail: arnulfo@eco.utexas.edu

**Departamento de Estadística e Investigación Operativa II, Universidad Complutense de Madrid. E-mail: pedro_nahum@ucm.universia.es

follow rules to characterize the interactions among explanatory variables that give rise to a sovereign debt crisis.

In a recent paper, Barney and Alse (2001) address the issues of reliability and comparability among competing models only in regard to their debt rescheduling predictive power. They provide a test sample comparative statistics for OLS, Logit, and Neural Networks under two criteria: classification error and accuracy. It is known that the classification error criterion relies on implicit assumptions that play havoc with comparing performances between models and that the tradeoff between sensitivity and specificity due to changes in the cutoff value could come from a random classifier (see, e.g., Hand, 1997). The accuracy criterion is only indicative of how close a model's prediction is to the actual data – i.e. how good the approximation is. However, no information at all is extracted as to the discrimination power of a tested model.

This paper evaluates and compares the discrimination power of various traditional and modern statistical methodologies for both the training and the test debt-related data samples. This is done by a pair-wise statistical comparison of the areas under the Receiver Operating Characteristic (ROC) curves. A ROC Curve displays the tradeoff between sensitivity and specificity as a function of the cutoff value. The area under the ROC Curve is equal to the probability that a randomly selected observation from the rescheduling population scores higher than a randomly selected observation from the non-rescheduling population.[1]

[1] This is true when the rescheduling population is coded with ones and the non-rescheduling population with zeros, vice versa otherwise.

The principal problem that the researcher faces is the specification of the relationship between the response variable, Y, and the explanatory factors $x = \{x_1, ..., x_p\}$, in an equation like the following:

$$Y = f(x_1, ..., x_p) + \varepsilon \tag{1}$$

where ε usually reflects the dependence of Y on other variables different from x. The solution to the estimation of the function $f(x_1, ..., x_p)$ is generated by:

$$\hat{f}(x) = \underset{g(x)}{\arg\min} = \sum_{i=1}^{N} \left[y_i - g(x_i)\right]^2 \tag{2}$$

Eq. (2) does not generate a unique solution (Friedman, 1994). There exists a set of functions that can interpolate the data, so the researcher must restrict the solutions in Eq. (2) to a subset of functions. These restrictions are based on considerations outside the database. In general, they are done with the election of the function approximation method because each one assumes a relationship between the explanatory factors and the response variable.

The Logistic Regression (Logit) will produce accurate predictions if and only if the parameterized equation is similar to the true function – i.e., the boundary that separates the non-reschedulings from the debt reschedulings cases is linear. A Neural Network model, particularly a Multi-Layer Network, searches for a non-linear boundary in the explanatory factors space. Two of its known weaknesses are sensitivity to irrelevant variables and null degree of interpretability. Moreover, the only piece of information that can be extracted from a trained Neural Network is the Relative Importance Measure.

The need to analyze debt reschedulings with methods both robust to irrelevant variables and with a high degree of interpretability prompted the use of Classification Trees (Breiman *et al.*, 1984) and Multivariate Adaptive Regression Splines "MARS" (Friedman, 1991) in addition to Logit and Neural Networks. These techniques (Classification Trees & MARS) have an automatic selection of relevant explanatory factors and interactions among them.

Decision-tree models are used here to address the issue of how explanatory variables might interact with each other to give rise to a debt rescheduling vs. a non-rescheduling outcome. Such trees provide a sequence of IF-THEN rules for debt rescheduling where explanatory variables could trigger it as soon as they stop satisfying an inequality. In order to come up with this sequence, a Tree-based model partitions the explanatory factors space into rectangle-like regions and fits a simple model at each terminal node via an estimation of a constant. A Relative Importance Measure for each one of the variables is also obtained for the Tree-based models used here.

MARS algorithms tackle the main Classification Trees' weakness: the discontinuity inherent in each node. This contribution will enhance the reliability if the true function that discriminates the non-reschedulings from the debt reschedulings cases is continuous. Interestingly, Sephton (2001) uses MARS models to predict American recessions and compares its performance to a Probit model's only in terms of an accuracy criterion. He finds that the test sample evidence indicates that the MARS models in his paper are helpful, but not entirely accurate predictions of recessions.

Galindo and Tamayo (2000) analyze the performance of Probit, *k*-Nearest-Neighbors, Tree-based, and Neural Networks models to assess the credit risk of a mortgage loan portfolio. They find that the Tree-based models produce the most accurate results.

The issue of interpretability of models for debt rescheduling is not addressed by Barney and Alse (2001). In this paper, the interpretability is addressed in the four models used here. Having this in mind, the decision to determine which method is better will be a function of both its discriminative and interpretation abilities.

The paper is organized as follows. In section 2 the ROC curve concept and its applications are explained. Section 3 briefly discusses the methodologies and their results are presented. Section 4 contains the results from the comparison of the areas under the ROC curves derived from applying the methodologies in section 3 to approximate the function for debt rescheduling. Concluding remarks follow in section 5.

2. Measuring and comparing the discrimination power of models: the Area under the ROC curve

Receiver Operating Characteristic (ROC) curves assess the ability of a method to discriminate between two populations. In this paper, the area under the ROC curve represents the probability of correctly ranking a random (debt rescheduling, non-rescheduling) pair.[2] This area exhibits a number of desirable properties when compared to overall accuracy in the evaluation of predictive (machine) learning algorithms (Bradley, 1997). A ranking probability of one would simply indicate that the probability distributions

[2] See Hamber (1975) and Hanley and McNeil (1982) for more details on the probabilistic meaning of the area under the ROC curve.

of debt rescheduling and non-rescheduling did not overlap at all – i.e., there is no possibility of wrongly classifying a debt rescheduling nor a non-rescheduling outcome. As opposed to perfect discrimination, an area under the ROC curve equal to one half would mean that the model is not capable of distinguishing at all between classes – i.e., this occurs when there is a perfect overlap of probability distributions.[3]

The ROC curve is generated by sweeping the cutoff point probability for non-rescheduling from zero to one. The x-axis represents the difference between one and the ratio of correctly classified debt reschedulings to the total number of debt reschedulings. The y-axis represents the ratio of correctly classified non-reschedulings to the total number of non-reschedulings. The area under a ROC curve could vary from 0.5 (Random Classifier) to 1 (Perfect Classifier). Figure 1 illustrates a typical ROC curve.

[Figure 1 About Here]

Parametric and Non-Parametric procedures can be used to derive the area under the ROC curve. The parametric procedures assume a probability distribution form for the two populations. Typically, the probability distribution is a 'binormal' one. The area under the ROC curve and its variance are obtained with Maximum Likelihood Estimation (see, e.g., Metz *et al.* 1998). On the other hand, a non-parametric procedure may assume a mathematical form of a distribution for the two populations to calculate the variance of the estimate. Usually, the exponential distribution is used. The area under the ROC curve is

[3] See section 5.14 in the "Material docente de la Unidad de Bioestadística Clínica" to understand the implications on the ROC curve of the probability distributions' overlap degree.

obtained with the Mann-Whitney-U Statistic (see Lehman 1998 and Hanley and McNeil 1982).

For a wide range of distributions, choosing between parametric and non-parametric approaches should not be made on considerations of imprecision or bias of the estimates of the area under the ROC curve. The reason is because the bias/imprecision of the misspecification of the underlying distribution has been found to be very small (Hajian-Tilaki *et al.,* 1997).

In this research paper, a nonparametric approach was chosen because the Decision-Tree, one of the function approximation algorithms employed here, forecasts with discrete values of probabilities. The quantity will depend on the number of terminal nodes, which was found to be small. To the authors' knowledge, there is no implementation hitherto of a parametric procedure that obtains the area under the ROC curve and its variance under these conditions.[4]

A ROC curve has three important applications: (1) to measure the discrimination power of a model by finding the area under the curve, (2) to compare the discrimination power between models and (3) to compare two probability cutoff points on the same curve. Application 2 is of interest here because numerically different areas under two ROC curves will not be a result of random sampling if their difference is statistically significant.[5]

---

[4] Although the PROPROC Software developed by Metz from the Department of Radiology at the University of Chicago Medical Center obtains the area under the ROC curve under these conditions.

[5] See Hanley and McNeil (1983) to understand how to compare the areas under different ROC curves derived from the same data.

3. Methodologies

The annual data used here for the explanatory variables spanned the 1986-1994 period. Fifty-two middle-income countries from all around the world were sampled. The chosen countries were the same as those used by Barney and Alse (2001) with the exceptions of Portugal and Myanmar. The World Debt Tables (1996, 1991, 1990) and the World Development Indicators (2002) were the data sources.

The data sample was randomly divided into a training sample (almost three quarters of the data sample) and a test sample. Roughly 42% (148/349) of the data in the training sample were observations of countries that rescheduled their debt payments. As for the test sample, approximately 41% (49/119) of the data were also reschedulings outcomes. The discrete dependent variable takes on 1 when there is debt rescheduling, 0 otherwise.

All of the explanatory factors used by Barney and Alse (2001) were employed. All of them were normalized to a mean equal to zero and a standard deviation equal to one.[6] The notation for the variables is shown in Table 1.

Table 1. Explanatory variables' notation.

EDT/GNP(%): the percentage of Total Debt Stock to Gross National Product
TDS/XGS(%): the percentage of Total Debt Service to Exports of Gods and Services
PCGNPG(%): the growth rate of the per capita Gross National Product
RES/MGS(months): the ratio of International Reserves to Imports of Goods and Services
EGR(%): the growth rate of Exports of Goods and Services
IR(%): the Consumer Price Index growth rate
MGS/GNP(%): the percentage of Imports of Goods and Services to Gross National Product
PPH: the previous payment history or the lag of the dependent variable

[6] The mean and the standard deviation obtained from the training data were applied to normalize the test data.

3.1 The Logistic Regression

The Logit model arises from the desire to model posterior probabilities of the classes via linear functions. It relies on Maximum Likelihood Estimation (MLE) to find the model that most accurately approximates its outcomes to the actual data. Its weaknesses consist of assuming a functional form ex-ante and a probability distribution for the error term. The regression results along with its summary statistics are shown in Table 2.

Table 2. Results of the Logit regression.

	Coefficient	Std. Error	Z-Value
(Intercept)	-1.68	0.234	-7.19
EDT/GNP(%)	1.72	0.470	3.66
TDS/XGS(%)	0.30	0.175	1.71
PCGNPG(%)	0.19	0.187	1.01
RES/MGS(months)	0.10	0.183	0.56
EGR(%)	0.01	0.178	0.06
IR(%)	0.13	0.119	1.09
MGS/GNP(%)	0.16	0.236	0.70
PPH	3.37	0.335	10.07
Log-likelihood	-125.54		
Likelihood Ratio	224.66		
A.I.C.	269.08		

It can be seen from Table 2 that only the coefficients corresponding to the previous payment history and the ratio of total debt stock to GNP are significant and with the expected sign.[7] The model yields a highly significant likelihood ratio statistic suggesting that the explanatory factors contain substantial explanatory power.

[7] In a Logit regression, a 100% change in an explanatory variable x_i, ceteris paribus, brings about a $(e^{\beta_i x}-1)100\%$ change in $\frac{P_i}{1-P_i}$ - i.e. the odds in favor of debt rescheduling.

Due to the existence of correlation among the explanatory factors, there could be a factor statistically insignificant because of the presence of an irrelevant factor. However, this could be mitigated with a model selection strategy.

A model selection strategy tries to find a subset of the explanatory factors that are sufficient for explaining the response variable. In this article a backward subset selection was implemented. This was done by dropping the least significant coefficient and by refitting the model after. This was done repeatedly until a list from the most important to least important factor was obtained. Then explanatory factors were added one by one beginning with the most important and ending with the least important factor. The model was estimated every time a new factor was included. An analysis of deviance was done to decide which variables to exclude.[8] The result of this strategy is shown in Table 3.

Table 3. Analysis of Deviance**.**

Terms	AIC	Residual Dev.	LRT	p-value
PPH	275.31	271.31		
EDT/GNP(%)	264.10	258.10	13.21	0.00
TDS/XGS(%)	262.99	254.99	3.11	0.08
PCGNPG(%)	262.94	252.94	2.05	0.15
IR(%)	263.88	251.88	1.06	0.30
MGS/GNP(%)	265.39	251.39	0.49	0.48
RES/MGS(months)	267.09	251.09	0.31	0.58
EGR(%)	269.08	251.08	0.00	0.95

[8] The residual deviance of a fitted model is minus twice its log-likelihood and the Akaike Information Criterion (AIC) is the residual deviance plus twice the number of parameters to the number of observations ratio.

In Table 3, the column labeled "Terms" lists explanatory factors in order of importance. The AIC acronym stands for Akaike Information Criterion while the LRT for Likelihood Ratio Test. Both measures are obtained when a "term" is added to the model.

It can be seen from the Table 3 that the Likelihood Ratio Test determines that two explanatory factors are enough to explain the phenomena at hand. In contrast, the AIC determines that four explanatory factors are enough to understand the sovereign debt rescheduling problem. In other words, the Likelihood Ratio Test determines that the Previous Payment History and the Total Debt Stock to Gross National Product ratio are the relevant factors to understand the debt rescheduling phenomena. The Akaike Information Criterion, in addition to the aforementioned factors, indicates that the Total Debt Service to Exports of Goods and Services ratio and the growth rate of Gross National Product per capita must be included in the final model.

To determine if the reduction in the explanatory factor space increases the predictive power, the prediction capabilities on the test sample are analyzed. The full model has the lowest classification error rate. Consequently, the model with all the explanatory factors was selected as the final Logit model.[9]

In order to compare the actual data versus a model's outcomes, performance matrices are obtained for both the training and test samples for a given cutoff value of 40 %. Performance matrices for the final Logit model are shown in Table 4.

[9] A ROC analysis was also carried out. The area under the ROC curve of the full model was not statistically superior to the area of the four-factor model at 90%. However, the full model had higher lower and upper confidence bounds on top of a lower mean squared error in the test sample.

Table 4. Performance Matrices for the Logit model.

	Predicted 0	Predicted 1
Actual 0	177	24
Actual 1	22	126
Training Sample		

	Predicted 0	Predicted 1
Actual 0	55	15
Actual 1	8	41
Test Sample		

In section 4 the performance matrices of every single methodology will be compared against one another in terms of classification errors for both debt reschedulings and non-reschedulings populations.

3.2 Multivariate Adaptive Regression Splines (MARS) models

MARS is a non-parametric procedure used to specify the functional form that best fits the model to the data. Such functional form consists of a sum of basis functions. They can be highly non-linear transformations of the explanatory variables. Nevertheless, the dependent variable is still a linear function of the basis functions.[10] Eq. (3) provides an example of a typical MARS model with one single explanatory variable.

$$f(x) = 1 + \beta_1 x_1 + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - k_1)_+^3 + \beta_5 (x - k_2)_+^3 \qquad (3)$$

where k_1 and k_2 are the knots or activation points for the explanatory variable that capture the shifts in the relationships between variables. The expression $(x - k_1)_+^3$ is equal to the $(\max(0, x - k_1))^3$.

[10] For a simple explanation of the MARS methodology and its application to economic recessions see Sephton (2001)

MARS finds the best model that fits the data by choosing the knots as well as the additive and interactive effects among explanatory variables that minimize the sum of squared errors.

This is done by searching for the knot of each explanatory factor that minimizes the sum of squared errors. Then the explanatory factors along with their knot that minimized the sum of squared errors will be incorporated into the model. Finally, interactions among variables and knots already in the model are looked for. The ultimate selection of the model is based on the generalized-cross-validation criterion of Craven and Wahba (1979). The MARS models used here allow for one, two and three interaction levels (IL) among the explanatory variables.[11] See Friedman (1991) for a detailed explanation of the estimation procedure.

As for the M.A.R.S. estimation, up to thirty basis functions in the forward model-building procedure were allowed. The cost per degree of freedom (i.e., the price of selecting a knot in a piecewise linear regression) was fixed to two and three in the backward deletion procedure for the additive and the interactive models, respectively. The results of the additive model (IL=1) are shown in Table 5.

[11]The computational algorithm used to solve those models was MARS™ -- Decision Support System -- by Salford Systems.

Table 5. Results of MARS additive model**.**

MARS Debt Reschedulings Estimates
Linear GCV = 0.1116
Cubic GCV = 0.1125
Cost per degree of freedom = 2

Explanatory variables	**Coefficient**	**Variable**
Constant	0.938	
Basis function 1 (BF1)	-0.653	1 if PPH = 0, 0 otherwise
Basis function 4 (BF4)	-0.303	Max(0, 0.266-EDT/GDP(%))

Just like in the case of the Logit model, the two most important factors are the Previous Payment History and the percentage of Total Debt Stock to Gross National Product. Even though the interpretation derived from this estimation is different from the Logit's, a country must have a good Previous Payment History and a Total Debt Stock to Gross National Product ratio lower that .266 (normalized) or 110.58% to reduce its rescheduling probabilities.

In order to see if there are interactions among the explanatory factors, a two-level interaction (IL=2) model was estimated. By limiting the interaction level to two the interpretation of the final model increases as opposed to higher level of interactions. The results of the two-level interactions are shown in Table 6.

Table 6. Results of MARS (IL=2) model.

MARS Debt Reschedulings Estimates
Linear GCV = 0.1052
Cubic GCV = 0.1074
Cost per degree of freedom = 3

Explanatory variables	**Coefficient**	**Variable**
Constant	0.598	
Basis Function 1 (BF1)		1 if PPH = 0, 0 otherwise
Basis Function 4 (BF4)	-0.396	Max(0, 0.45-EDT/GDP(%))*BF1
Basis Function 8 (BF8)	0.084	Max(0, 2.712-RES/MGS(months))
Basis Function 9 (BF9)	-0.163	BF1*BF8
Basis Function 14 (BF14)		Max(0, PCGNPG(%)-0.738)
Basis Function 16 (BF16)	0.106	Max(0, TDS/XGS(%)+1.789)*BF14

The results from Table 6 indicate that the likelihood of a debt rescheduling decreases when there are a good previous payment history along with a low percentage of Total Debt Stock to Gross National Product combined with a good payment history with a low ratio of reserves to imports. On the contrary, the likelihood of a debt rescheduling increases when there are a low ratio of reserves to imports plus a high percentage of the total debt service to exports ratio with a high rate of Gross National Product per capita.

Higher-order products or interactions may increase the prediction power if the true function that determines the relationship between the explanatory factors and the response variable has higher-order products. The possibility of three way products or interactions was looked into, the results are shown in Table 7.

Table 7. Results of MARS (IL=3) model.

MARS Debt Reschedulings Estimates
Linear GCV=.1020
Cubic GCV=.1071
Cost per degree of freedom =3

Explanatory variables	Coefficient	Variable
Constant	0.966	
Basis function 1 (BF1)		1 if PPH=0, 0 otherwise
Basis function 2 (BF2)		1 if PPH=1, 0 otherwise
Basis function 3 (BF3)	0.941	Max(0,.45-EDT/GNP(%))*BF1
Basis function 4 (BF4)	7.943	Max(0,EDT/GNP(%)-.45)*BF1
Basis function 5 (BF5)		Max(0,RES/MGS(months)-1.644)*BF1
Basis function 6 (BF6)		Max(0,1.644-RES/MGS(months))*BF1
Basis function 7 (BF7)	-0.388	Max(0,EDT/GNP(%)+.246)*BF6
Basis function 8 (BF8)	0.549	Max(0,-.243-EDT/GNP(%))*BF6
Basis function 9 (BF9)	-0.36	Max(0,RES/MGS(months)+1.229)*BF2
Basis function 10 (BF10)	-0.332	Max(0, TDS/XGS(%)-.928)*BF9
Basis function 13 (BF13)	0.984	Max(0,-.005-IR(%))*BF9
Basis function 14 (BF14)	0.326	Max(0, TDS/XGS(%)+1.789)*BF5
Basis function 17 (BF17)	-0.897	Max(0, MGS/GNP(%)+10.171)*BF4
Basis function 18 (BF18)	0.163	Max(0, TDS/XGS(%)+.883)*BF9

It is more difficult to interpret the final MARS model with three-level of interactions than the MARS models with lower levels of interaction. The Relative Important Measure helps to determine which variables are more relevant to understanding the rescheduling phenomena. Figure 2 contains the relative importance of variables for the MARS models with one, two, and three variable interactions, respectively.[12] It is worth mentioning that the previous payment history ended up being the most important explanatory variable for the three MARS models considered here.

[12] Each number multiplied by a hundred indicates what happens to the explanatory power of the model when the corresponding explanatory factor is omitted. The higher the percentage number is, the more explanatory power the variable has.

[Figure 2 About Here]

The performance matrices with a cutoff value of 40% for each level of interaction are shown in Table 8.

Table 8. Performance Matrices for MARS additive and interaction models.

	Predicted 0	Predicted 1
Actual 0	179	22
Actual 1	24	124
Training Sample		

	Predicted 0	Predicted 1
Actual 0	56	14
Actual 1	9	40
Test Sample		

a) Performance Matrices for MARS (IL=1)

	Predicted 0	Predicted 1
Actual 0	179	22
Actual 1	17	131
Training Sample		

	Predicted 0	Predicted 1
Actual 0	56	14
Actual 1	9	40
Test Sample		

b) Performance Matrices for MARS (IL=2)

	Predicted 0	Predicted 1
Actual 0	180	21
Actual 1	15	133
Training Sample		

	Predicted 0	Predicted 1
Actual 0	53	17
Actual 1	16	33
Test Sample		

c) Performance Matrices for MARS (IL=3)

It can be seen from Table 8 that as the level of interaction is incremented the performance in the training sample increases. However, this is not the case in the test sample where the additive and the two-level of interaction models are equally better than the three-level of interactions model, clearly indicating a sign of over-fitting.

3.3 Tree-based models

Tree-based models (Breiman *et al.*, 1984) are powerful non-parametric methods that deliver accurate predictions and, most importantly, easy to interpret rules that characterize a phenomena. Decision-tree models are used here to address the issue of how explanatory variables interact with each other through a sequence of IF-THEN rules for debt rescheduling. Some explanatory variables will be located at different nodes of the tree and climbing down the right part of a tree is conditional on them not satisfying their inequality.[13]

The impurity functions used here were the Gini and the information or cross entropy indices. Such impurity functions derive their names from the fact that it is practically impossible to make all observations from one class go right and the rest of the observations from the other class go left – i.e., there is no purity. In other words, the Tree-based models try to separate the rescheduling from the non-rescheduling population via a split.

The procedure to generate a Tree-based model can be broken down into two stages. The first stage represents the growing of the tree. In this stage, the algorithm tries to find the split that maximizes the decrement in the impurity function in order to make the tree grow. This is done iteratively until a certain amount of observations is reached or until no further decrements in impurity functions are found.

[13] Notice from the trees in Figures 3 and 4 that going right when climbing down the tree represents situations of debt being rescheduled.

Due to the fact that the tree generated by the first stage will generally over-fit the training data, a second stage, the pruning, is needed. In this article, the Cost-Complexity Pruning was implemented. For a discussion of the different pruning methods see Esposito, Malerba, and Semeraro (1997). In summary, the pruning tries to find the best ratio derived from changes in the impurity function produced by changes in the number of terminal nodes (complexity parameter). Usually, the criterion that defines the best ratio is the error rate in an independent or test sample or the k-Fold Cross-Validation error rate.[14]

The Total Error Rate on the test sample as a function of the number of splits is shown in Table 9 for the Tree-based models with Gini and Information impurity functions.

Table 9. Tree-based model Total Error Rate in the test data as a function of the number of splits.

Complexity Parameter	Number of Splits	Total Error Rate
0.7283	0	58.82%
0.0186	1	19.33%
0.0100	3	19.33%

a) Gini Impurity Function

Complexity Parameter	Number of Splits	Total Error Rate
0.7283	0	58.82%
0.0170	1	19.33%
0.0100	3	25.21%

b) Information Impurity Function

It can be seen from Table 9 that the number of splits for the estimated trees is small. Moreover, according to Table 9, the tree grown under the information impurity function,

[14] The statistical software used to estimate the tree based models was R, available at http://www.r-project.org , under R-part package develop by Terry M. Therneau and Beth Atkinson, R-port by Brian Ripley. Unfortunately, such implementation does not provide a pruning code with cross-validation on the test data.

given a cutoff point, could increase its discrimination power in the test sample if it were pruned.

For several reasons it was decided not to prune the tree grown under the information impurity function. First, if a tree is pruned there will be a bias on the expected error rate on a future database. Second, the tree is highly interpretable. Third, after performing a ROC analysis the tree with three splits had a statistically significant higher area under the ROC curve.

The tree grown under the Gini impurity function is shown in Figure 3.

[Figure 3 About Here]

In order to explain how a rescheduling situation occurs, the tree from above is read in the following way:

Case a)

If PPH ≠ 0 then rescheduling. That is, there will always be rescheduling if there is a bad previous payment history.

Case b)

If PPH=0 and EDT/GNP (%) >= 69.9% and PCGNPG (%) >= 14.5% then rescheduling. That is, there will be rescheduling if the ratio of total debt stock to GNP is at least 69.94% and the per capita growth rate is at least 14.50% despite a good previous payment history.

The Tree-based model grown with Information Impurity Function is shown in Figure 4.

[Figure 4 About Here]

The same number of conditional rules for rescheduling outcomes occurs again. In order to explain how a rescheduling outcome is brought about, the tree from above is read in the following way:

Case a)

If PPH ≠ 0 then rescheduling. That is, there will be rescheduling if there is a bad previous payment history.

Case b)

If PPH=0 and EDT/GNP (%) >= 58.1% and RES/MGS (months) >= 4.55 then rescheduling. That is, there will be rescheduling if the ratio of total debt stock to GNP is at least 58.10% and the ratio of international reserves to imports is at least 4.55 months despite a good previous payment history.

An analysis of the relative importance of the explanatory variables was done for each tree model. Such analysis determines which variables are the most important ones for the learning process that takes place in the training sample.[15] Figure 5 illustrates it for each case.

[15] This is done via surrogate splits. In other words, what would be the accumulated effect on the impurity function if the split were done with another explanatory factor? This procedure is repeated for every node.

[Figure 5 About Here]

It is worth mentioning that the most important explanatory factors are the Previous Payment History and the percentage of Total Debt Stock to Gross National Product for both tree models. Table 10 contains the performance matrices for both tree models in the training and test sample.

Table 10. Performance Matrices for the Tree-based models.

Training Sample	Predicted 0	Predicted 1
Actual 0	177	24
Actual 1	17	131

Test Sample	Predicted 0	Predicted 1
Actual 0	56	14
Actual 1	9	40

a) Performance matrices for the tree-based model (Gini)

Training Sample	Predicted 0	Predicted 1
Actual 0	175	26
Actual 1	16	132

Test Sample	Predicted 0	Predicted 1
Actual 0	48	22
Actual 1	8	41

b) Performance matrices for the tree-based model (Information)

The Tree-based models have a very similar performance in the training data. The only difference comes from the ability to predict the non-rescheduling cases in the test data, where the tree model grown with the Gini impurity function works better.

3.4 Neural Networks models

In this paper a Neural Network with eight inputs, eight nodes and one hidden layer (a 8-8-1 Neural Network) ended up being the one that delivered the lowest mean squared error (MSE) in the test sample. Different Neural Network's architectures were also experimented with and did not deliver any significant improvement in terms of the test sample MSE. The logistic function was used in the hidden layer as well as in the output node. Such function only produces values between zero and one. A range for output between zero and one is needed in order to have probabilistic values.[16]

In the optimization process an early stop criterion was employed because of the bias-variance tradeoff shown in the Figure 6.

[Figure 6 About Here]

It can be seen from Figure 6 that the MSE decreases in the training data as the number of epochs (iterations) increases. Exactly the opposite occurs in the test sample as soon as the minimum is found. The MSE increases because of the bias-variance tradeoff or process over-fitting. In order to select the best neural network architecture, the performances on the test sample were compared when the minimum MSE in the test sample was found.

The relative importance measures for each explanatory factor in the estimated Neural Network model are shown in Figure 7.

[16] The neural networks toolbox provided by MATLAB™ was used to run the 8-8-1 Neural Network. A variable learning rate optimization algorithm was employed instead of the slower and more common Levenberg-Marquardt (LM).

[Figure 7 About Here]

As can be seen in Figure 7 the most important variable is the Previous Payment History, the rest of the variables play roughly the same insignificant role in understanding the problem at hand. Table 11 shows the performances matrices for both training and test sample with a cutoff value of 40%.

Table 11. Performance Matrices for the Neural Network model.

	Predicted 0	Predicted 1
Actual 0	179	22
Actual 1	16	132
Training Sample		

	Predicted 0	Predicted 1
Actual 0	56	14
Actual 1	9	40
Test Sample		

4. Comparing the areas under the ROC curves derived from the methodologies used above

The classification error provides a way of comparing the performance of methodologies for a single probability cutoff point. Table 12 contains such numbers for each methodology explained above.

Table 12. Classification Error Measures.

	Cutoff value = 0.40			
	Training Sample		Test Sample	
Method	Non-ReSCE	ResCE	Non-ReSCE	ResCE
Logit	0.119	0.149	0.214	0.163
MARS (IL=1)	0.109	0.162	0.200	0.184

MARS (IL=2)	0.109	0.115	0.200	0.184
MARS (IL=3)	0.104	0.101	0.243	0.327
Tree (Function = Gini)	0.119	0.115	0.200	0.184
Tree (Function = Information)	0.129	0.108	0.314	0.163
Neural Network	0.109	0.108	0.200	0.184

It can be seen from Table 12 that if the losses from misclassifying a debt rescheduling are greater than those from misclassifying a non-rescheduling, then the tree with Information impurity function and the Logit model will be the models to consider as a consequence of having the lowest (16.30%) classification error for reschedulings.[17]

However, the classification error does not tell us anything in regard to the discrimination power of the methodologies for any other cutoff point. A ROC curve, as explained before, sweeps the cutoff point from zero to one. The area under it is indeed a measure of the overall discrimination power of a methodology. For visualization purposes, Figure 8 shows the ROC curves for the Logit and Gini index Tree-based models.

[Figure 8 About Here]

The drawback of analyzing ROC curves by just a glance is the fact that two different ROC curves may have the same area under it. Table 13 presents the area under the ROC curve along with its 95% confidence interval and a common accuracy measure for each methodology used above.

[17] The goal for any function approximation method is to perform better in the test sample (to be able to generalize) by not over-fitting the training data.

Table 13. Areas under ROC curves and accuracy measures.

		Training Sample		Test Sample	
Method	Specification	MSE	ROC Area	MSE	ROC Area
Logit	Full model	0.106	0.881 ± 0.0397	0.143	0.857 ± 0.0728
MARS (IL=1)	IL =1	0.108	0.892 ± 0.0374	0.149	0.847 ± 0.0743
MARS (IL=2)	IL = 2	0.097	0.912 ± 0.0335	0.161	0.814 ± 0.0804
MARS (IL=3)	IL = 3	0.083	0.924 ± 0.0321	0.200	0.788 ± 0.0833
Neural Network	architecture (8-8-1)	0.097	0.914 ± 0.0335	0.149	0.847 ± 0.0720
Tree-based model	Impurity Function = Gini	0.102	0.898 ± 0.0357	0.157	0.836 ± 0.0753
Tree-based model	Impurity Function = Information	0.102	0.905 ± 0.0337	0.190	0.817 ± 0.0795

A model that it is able to generalize is looked for. In other words, in the presence of an independent database this particular model will be able to make accurate and precise predictions. To determine which model is better or preferable, the selection must be made in terms of the discrimination and accuracy power on the test data.

The discrimination power and accuracy of the estimated models diminish in the test sample, being the Logit model the most consistent in terms of discrimination power and accuracy while the MARS (IL=3) the most inconsistent of all.

The critical ratio z test is used to determine whether two areas are statistically different from one another or not.[18] A critical ratio z equal to 1.29—i.e., one out of roughly 10 samples produces different areas due to random sampling—is considered enough to establish a statistical difference between the discrimination powers of two methodologies

[18] See Hanley and McNeil (1983) to find out how to calculate such critical ratio.

being compared against each other. Table 14 shows the resulting preferences between models derived from the ratio numbers for the training sample. A matrix of preferences between models is read in the following way: the model in a row is more preferred (MP) or equally preferred (EP) or less preferred (LP) than the model in a column.

Table 14. Matrix of preferences between models for the training sample.

Method	Logit	MARS (IL=1)	MARS (IL=2)	MARS (IL=3)	Tree(Function = Gini)	Tree(Function = Information)	Neural Network
Logit	NA	EP	LP	LP	EP	LP	LP
MARS (IL=1)		NA	LP	LP	LP	LP	EP
MARS (IL=2)			NA	LP	EP	EP	EP
MARS (IL=3)				NA	MP	MP	MP
Tree (Function = Gini)					NA	EP	EP
Tree (Function = Information)						NA	EP
Neural Network							NA

In the training sample, the MARS model with three-way interactions is more preferred than the other methodologies employed in this investigation. However, it is necessary to see if this result is due to over-training. Consequently, the same analysis on the test sample was performed. The results are shown in Table 15.

Table 15. Matrix of preferences between models for the test sample.

Method	Logit	MARS (IL=1)	MARS (IL=2)	MARS (IL=3)	Tree(Function = Gini)	Tree(Function = Information)	Neural Network
Logit	NA	EP	EP	MP	EP	EP	EP
MARS (IL=1)		NA	EP	MP	EP	EP	EP
MARS (IL=2)			NA	MP	EP	EP	LP
MARS (IL=3)				NA	LP	LP	LP
Tree (Function = Gini)					NA	EP	EP
Tree (Function = Information)						NA	EP
Neural Network							NA

The MARS (IL=3) model clearly underperforms when compared to the rest of the methodologies in the test sample. Consequently, from the third row in Table 15, it can be said that the MARS model with two-variable interactions is equally preferred than the tree grown under the Information impurity function. This finding is important because it contradicts what the MSE criterion says. Moreover, according to the last column of Table 15, the Neural Network is equally preferred to the Logit, MARS additive model and trees models. For someone that had interpretation needs, the Tree-based models and the MARS additive model would be preferred to the Neural Network.

Barney and Alse (2001) find that their models are equally reliable. According to Table 15, only seven out of twenty-one comparisons between models indicate something else other than equal reliability of models. In fact, six of them occur because of the over-parameterized MARS (IL=3) model. The other one is due to the Neural Network model outperforming the MARS (IL=2) model.

5. Conclusions

To the authors' knowledge, only one study in the area of sovereign debt had used a variety of statistical methodologies to test the reliability of their predictions and to compare their performance against one another. However, those comparisons across models had been made only in terms of different probability cutoff points and mean squared errors. Moreover, the issue of interpretability of models for debt rescheduling had not been addressed in terms of interactions among explanatory variables with their correspondent debt rescheduling threshold level. This paper tackled these two issues and provided basis functions and easy to follow decision rules to distinguish debt rescheduling from non-rescheduling outcomes.

First, this paper evaluated and compared the discrimination power of various traditional and modern statistical methodologies for both the training and the test debt-related data samples. This was done by a pair-wise statistical comparison of the areas under the Receiver Operating Characteristic (ROC) curves. Logit, Neural Networks, MARS (Multivariate Adaptive Regression Splines), and Tree-based models were the function approximation algorithms used here.

In relation to the predictive power of the models, fourteen out of twenty-one (66.66%) cases indicated equal reliability of models when compared with each other. This finding is in line with what was found in Barney and Alse (2001) who rely on comparing mean squared errors across models. However, the MSE criterion sometimes could be misleading as to what model to use. A measure like the area under the ROC provides an indicator to assess and compare the discrimination power of models, something more desirable to have than a simple accuracy measure for understanding and predicting sovereign debt rescheduling. Moreover, the finding that the Neural Network was not more reliable than the MARS additive and Tree-based models makes a good case to use the latter two methodologies in order to come to grips with sovereign debt rescheduling.

Second, this paper did an analysis of deviance for the Logit model in order to find out which variables were more important than others regarding sovereign debt rescheduling. Along the same line of inquiry, an analysis of the relative importance of the explanatory variables was done for the Neural Network, MARS, and Tree-based models. All of the models ranked the previous payment history as the most important explanatory variable for the macroeconomic phenomena at hand. Future research should look into the implications of defining sovereign debt rescheduling in a different way – i.e., there is sovereign debt rescheduling this period if and only if there is rescheduling this period and there was no rescheduling during the previous period. Also, future research should look into ways of building up a good sovereign credit history. Furthermore, a model should be analyzed within the framework of partial areas under ROC curves to determine if it is more preferred than any other.

The findings of this paper will allow researchers, policy makers and financial analysts to decide, when confronted with equally preferred models for the test sample, which model to rely on based on the their needs of interpretation.

Acknowledgments

Arnulfo Rodriguez thanks David Kendrick, Juan Pablo Graf and Jose Luis Negrin for useful comments made on earlier versions of this paper and Dean Corbae and Scott Freeman for their academic support. He also thanks Peter Ward and the Mexican Center of LLILAS at The University of Texas at Austin for financial support (E.D. Farmer fellowship). Pedro N. Rodriguez thanks CONACYT (Mexico) for financial support (Fellowship: 170328).

Table 1. Explanatory variables' notation.

EDT/GNP(%): the percentage of Total Debt Stock to Gross National Product
TDS/XGS(%): the percentage of Total Debt Service to Exports of Goods and Services
PCGNPG(%):the growth rate of the per capita Gross National Product
RES/MGS(months): the ratio of International Reserves to Imports of Goods and Services
EGR(%): the growth rate of Exports of Goods and Services
IR(%): the CPI growth rate
MGS/GNP(%): the percentage of Imports of Goods and Services to Gross National Product
PPH: the previous payment history or the lag of the dependent variable

Table 2. Results of the Logit regression.

	Coefficient	Std. Error	Z-Value
(Intercept)	-1.68	0.234	-7.19
EDT/GNP(%)	1.72	0.470	3.66
TDS/XGS(%)	0.30	0.175	1.71
PCGNPG(%)	0.19	0.187	1.01
RES/MGS(months)	0.10	0.183	0.56
EGR(%)	0.01	0.178	0.06
IR(%)	0.13	0.119	1.09
MGS/GNP(%)	0.16	0.236	0.70
PPH	3.37	0.335	10.07
Log-likelihood	-125.54		
Likelihood Ratio	224.66		
A.I.C.	269.08		

Table 3. Analysis of Deviance**.**

Terms	AIC	Residual Dev.	LRT	P-Value
PPH	275.31	271.31		
EDT/GNP(%)	264.10	258.10	13.21	2.79E-04
TDS/XGS(%)	262.99	254.99	3.11	7.80E-02
PCGNPG(%)	262.94	252.94	2.05	1.53E-01
IR(%)	263.88	251.88	1.06	3.04E-01
MGS/GNP(%)	265.39	251.39	0.49	4.84E-01
RES/MGS(months)	267.09	251.09	0.30	5.80E-01
EGR(%)	269.08	251.08	0.01	9.51E-01

Table 4. Performance Matrices for the Logit model.

	Predicted 0	Predicted 1
Actual 0	177	24
Actual 1	22	126
Training Sample		

	Predicted 0	Predicted 1
Actual 0	55	15
Actual 1	8	41
Test Sample		

Table 5. Results of MARS additive model**.**

MARS Debt Reschedulings Estimates
Linear GCV=.1116
Cubic GCV=.1125
Cost per degree of freedom =2

Explanatory variables	Coefficient	Variable
Constant	0.938	
Basis function 1 (BF1)	-0.653	1 if PPH=0, 0 otherwise
Basis function 4 (BF4)	-0.303	Max(0,.266-EDT/GNP(%))

Table 6. Results of MARS (IL=2) model.

MARS Debt Reschedulings Estimates

Linear GCV=.1052

Cubic GCV=.1074

Cost per degree of freedom =3

Explanatory variables	Coefficient	Variable
Constant	0.598	
Basis function 1 (BF1)		1 if PPH=0, 0 otherwise
Basis function 4 (BF4)	-0.396	Max(0,.45-EDT/GNP(%))*BF1
Basis function 8 (BF8)	0.084	Max(0,2.712-RES/MGS(months))
Basis function 9 (BF9)	-0.163	BF1*BF8
Basis function 14 (BF14)		Max(0, PCGNPG(%)-.738)
Basis function 16 (BF16)	0.106	Max(0, TDS/XGS(%)+1.789)*BF14

Table 7. Results of MARS (IL=3) model.

MARS Debt Reschedulings Estimates
Linear GCV=.1020
Cubic GCV=.1071
Cost per degree of freedom =3

Explanatory variables	**Coefficient**	**Variable**
Constant	0.966	
Basis function 1 (BF1)		1 if PPH=0, 0 otherwise
Basis function 2 (BF2)		1 if PPH=1, 0 otherwise
Basis function 3 (BF3)	0.941	Max(0,.45-EDT/GNP(%))*BF1
Basis function 4 (BF4)	7.943	Max(0,EDT/GNP(%)-.45)*BF1
Basis function 5 (BF5)		Max(0,RES/MGS(months)-1.644)*BF1
Basis function 6 (BF6)		Max(0,1.644-RES/MGS(months))*BF1
Basis function 7 (BF7)	-0.388	Max(0,EDT/GNP(%)+.246)*BF6
Basis function 8 (BF8)	0.549	Max(0,-.243-EDT/GNP(%))*BF6
Basis function 9 (BF9)	-0.36	Max(0,RES/MGS(months)+1.229)*BF2
Basis function 10 (BF10)	-0.332	Max(0, TDS/XGS(%)-.928)*BF9
Basis function 13 (BF13)	0.984	Max(0,-.005-IR(%))*BF9
Basis function 14 (BF14)	0.326	Max(0, TDS/XGS(%)+1.789)*BF5
Basis function 17 (BF17)	-0.897	Max(0, MGS/GNP(%)+10.171)*BF4
Basis function 18 (BF18)	0.163	Max(0, TDS/XGS(%)+.883)*BF9

Table 8. Performance Matrices for MARS additive and interaction models.

	Predicted 0	Predicted 1
Actual 0	179	22
Actual 1	24	124
Training Sample		

	Predicted 0	Predicted 1
Actual 0	56	14
Actual 1	9	40
Test Sample		

a) Performance Matrices for MARS (IL=1)

	Predicted 0	Predicted 1
Actual 0	179	22
Actual 1	17	131
Training Sample		

	Predicted 0	Predicted 1
Actual 0	56	14
Actual 1	9	40
Test Sample		

b) Performance Matrices for MARS (IL=2)

	Predicted 0	Predicted 1
Actual 0	180	21
Actual 1	15	133
Training Sample		

	Predicted 0	Predicted 1
Actual 0	53	17
Actual 1	16	33
Test Sample		

c) Performance Matrices for MARS (IL=3)

Table 9. Tree-based model Total Error Rate in the test data as a function of the number of splits.

Complexity Parameter	Number of Splits	Total Error Rate
0.72839	0	58.82%
0.01867	1	19.33%
0.01000	3	19.33%

a) Gini Impurity Function

Complexity Parameter	Number of Splits	Total Error Rate
0.728385	0	58.82%
0.017077	1	19.33%
0.010000	3	25.21%

a) Information Impurity Function

Table 10. Performance Matrices for the Tree-based models.

	Predicted 0	Predicted 1
Actual 0	177	24
Actual 1	17	131
Training Sample		

	Predicted 0	Predicted 1
Actual 0	56	14
Actual 1	9	40
Test Sample		

a) Performance Matrices for the tree model (Gini)

	Predicted 0	Predicted 1
Actual 0	175	26
Actual 1	16	132
Training Sample		

	Predicted 0	Predicted 1
Actual 0	48	22
Actual 1	8	41
Test Sample		

b) Performance Matrices for the tree model (Information)

Table 11. Performance Matrices for the Neural Network model.

	Predicted 0	Predicted 1
Actual 0	179	22
Actual 1	16	132
Training Sample		

	Predicted 0	Predicted 1
Actual 0	56	14
Actual 1	9	40
Test Sample		

Table 12. Classification Error Measures.

	Cutoff Value = 40%			
	Training Sample		Test Sample	
Method	Non-ResCE	ResCE	Non-ResCE	ResCE
Logit	0.119	0.149	0.214	0.163
MARS(IL=1)	0.109	0.162	0.200	0.184
MARS(IL=2)	0.109	0.115	0.200	0.184
MARS(IL=3)	0.104	0.101	0.243	0.327
Tree(Function=Gini)	0.119	0.115	0.200	0.184
Tree(Function=Information)	0.129	0.108	0.314	0.163
Neural Network	0.109	0.108	0.200	0.184

Table 13. Areas under ROC curves and accuracy measures.

		Training Sample		Test Sample	
Method	Specification	MSE	ROC Area	MSE	ROC Area
Logit	Full Model	0.106	0.881 ±. 0395	0.143	0.857 ± .0728
MARS	IL=1	0.108	0.892 ± .0374	0.149	0.847 ± .0743
MARS	IL=2	0.097	0.912 ± .0335	0.161	0.814 ± .0804
MARS	IL=3	0.083	0.924 ± .0321	0.200	0.788 ± .0833
Neural Network	architecture(8-8-1)	0.097	0.914 ± .0335	0.149	0.847 ± .0720
Tree model	Impurity function=Gini	0.102	0.898 ± .0357	0.157	0.836 ± .0753
Tree model	Impurity function=Information	0.102	0.905 ± .0337	0.190	0.817 ± .0795

Table 14. Matrix of preferences between models for the training sample.

Method	Logit	MARS(IL=1)	MARS(IL=2)	MARS(IL=3)	Tree(Function=Gini)	Tree(Function=Information)	Neural Network
Logit	NA	EP	LP	LP	EP	LP	LP
MARS(IL=1)		NA	LP	LP	LP	LP	EP
MARS(IL=2)			NA	LP	EP	EP	EP
MARS(IL=3)				NA	MP	MP	MP
Tree(Function=Gini)					NA	EP	EP
Tree(Function=Information)						NA	EP
Neural Network							NA

Table 15. Matrix of preferences between models for the test sample.

Method	Logit	MARS(IL=1)	MARS(IL=2)	MARS(IL=3)	Tree(Function=Gini)	Tree(Function=Information)	Neural Network
Logit	NA	EP	EP	MP	EP	EP	EP
MARS(IL=1)		NA	EP	MP	EP	EP	EP
MARS(IL=2)			NA	MP	EP	EP	LP
MARS(IL=3)				NA	LP	LP	LP
Tree(Function=Gini)					NA	EP	EP
Tree(Function=Information)						NA	EP
Neural Network							NA

Figure 1. The Receiver Operating Characteristic curve.

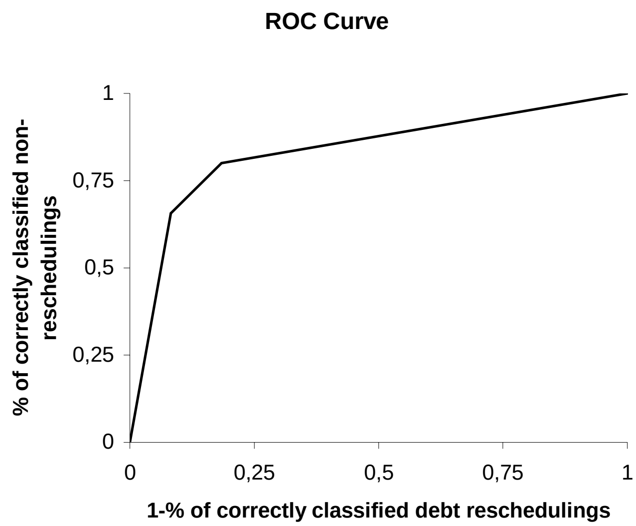
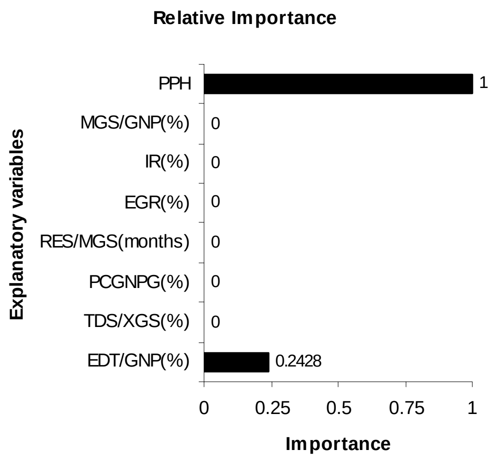
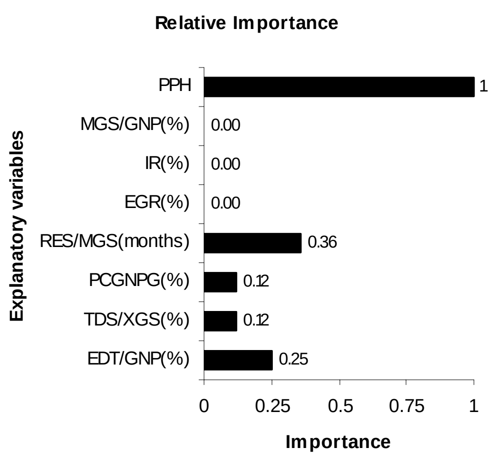


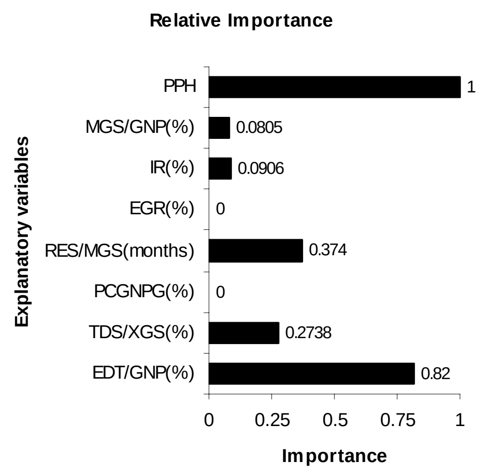
Figure 2. Relative Importance Measures for MARS models.



a) Relative Importance Mesure for MARS (IL=1)



b) Relative Importance Mesure for MARS (IL=2)



c) Relative Importance Mesure for MARS (IL=3)

Figure 3. Tree model grown under the Gini Impurity Function.

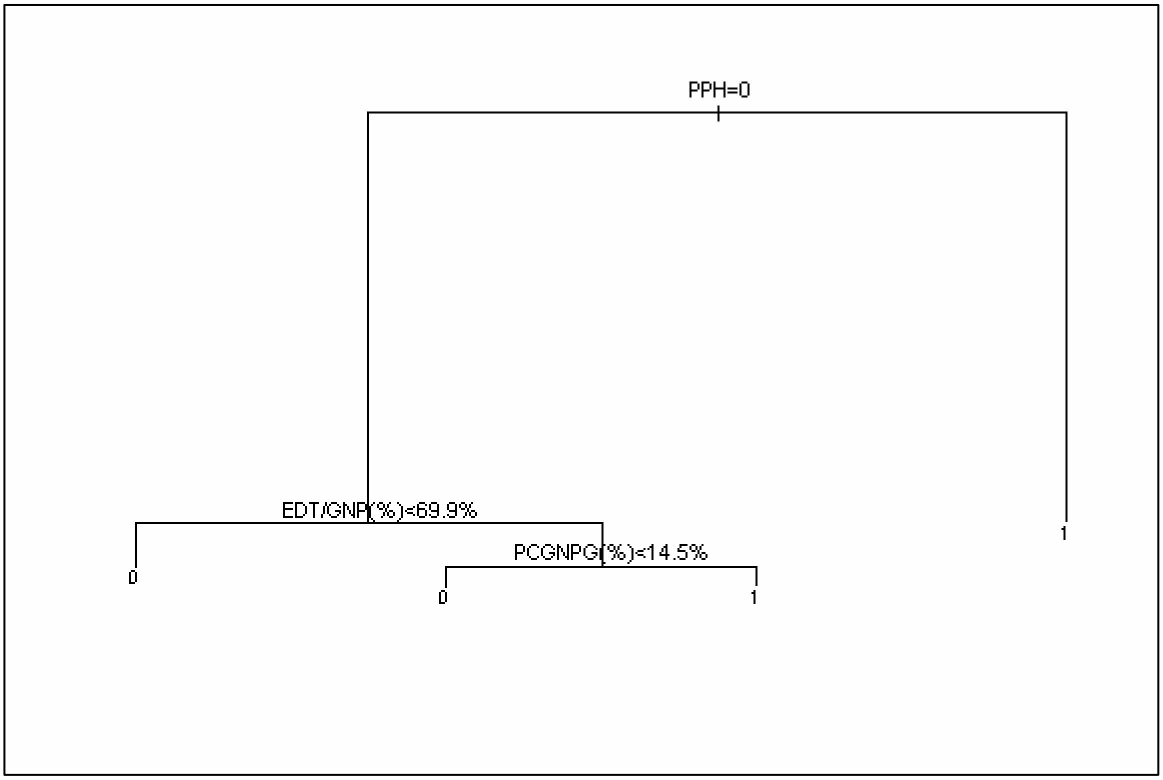


Figure 4. Tree model grown under the Information Impurity Function.

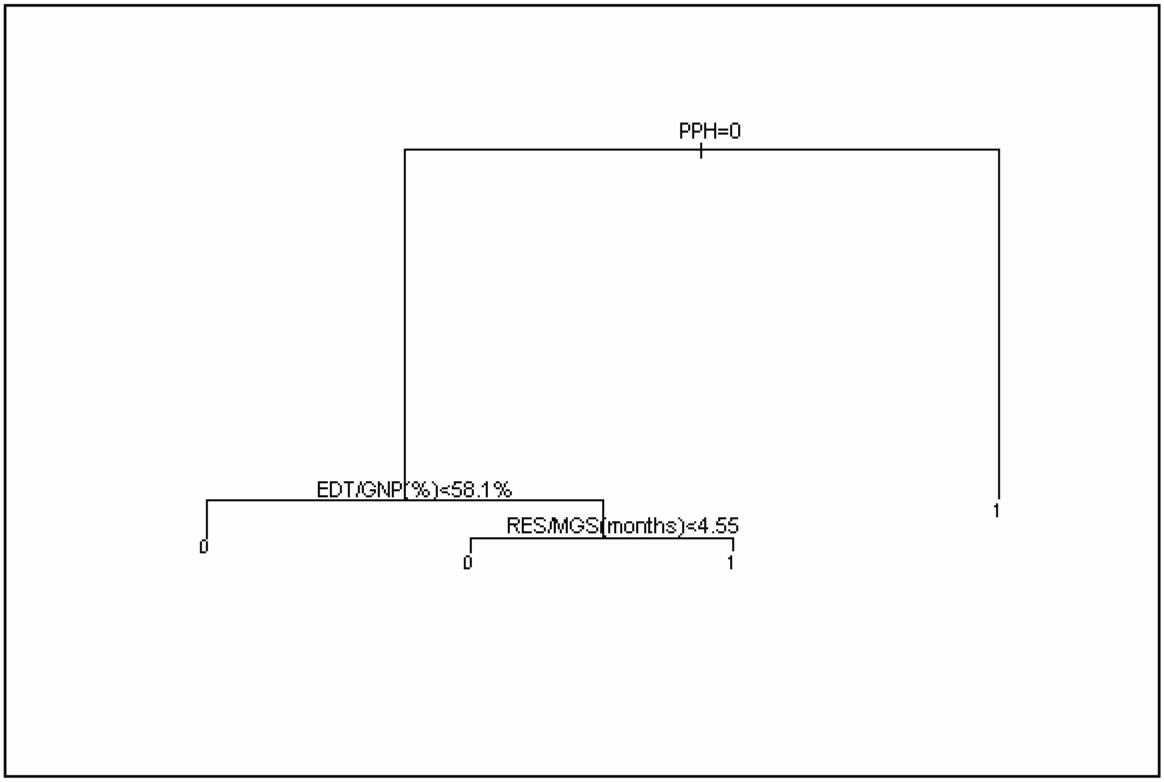
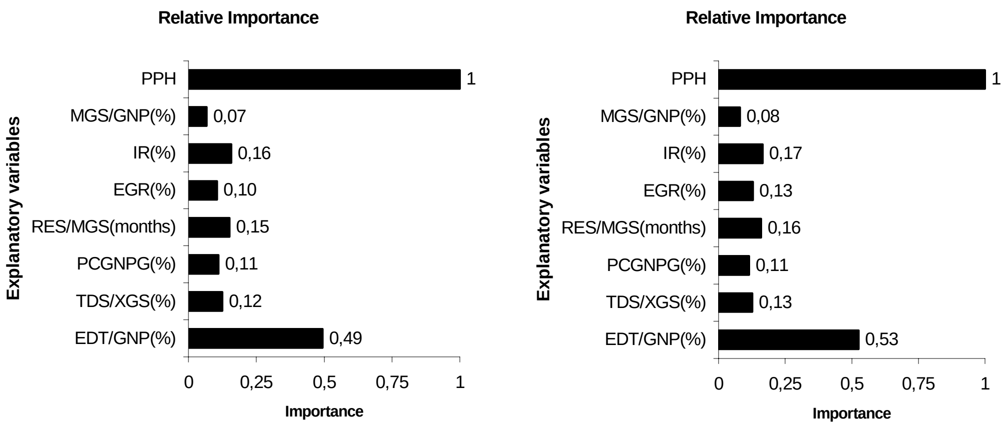


Figure 5. Relative Importance Measure for the Tree-based models.



a) Relative Importance Measure for the tree model with Gini impurity function

b) Relative Importance Measure for the tree model with Information impurity function

Figure 6. The bias-variance tradeoff.

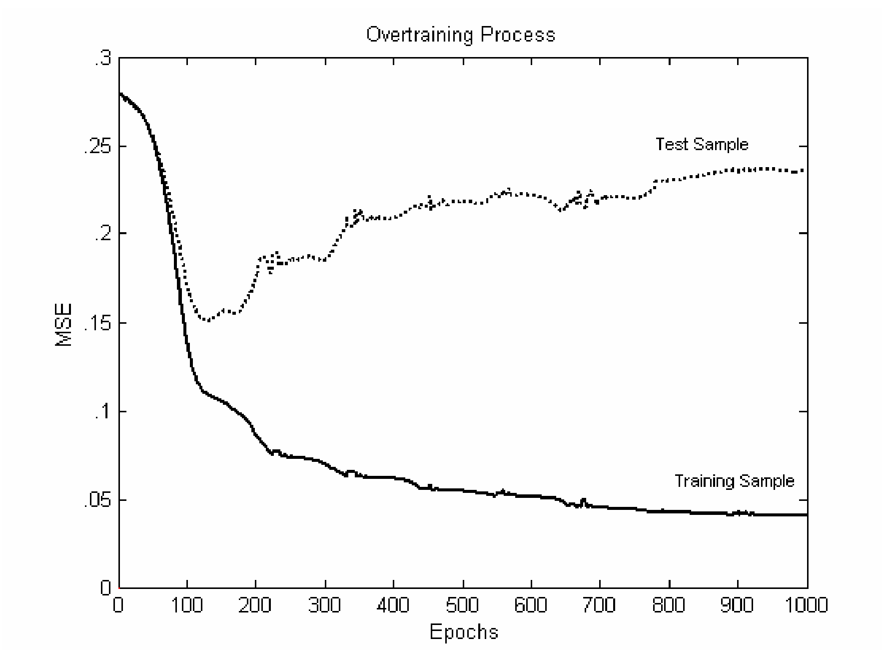


Figure 7. Relative Importance Measure for the Neural Network model.

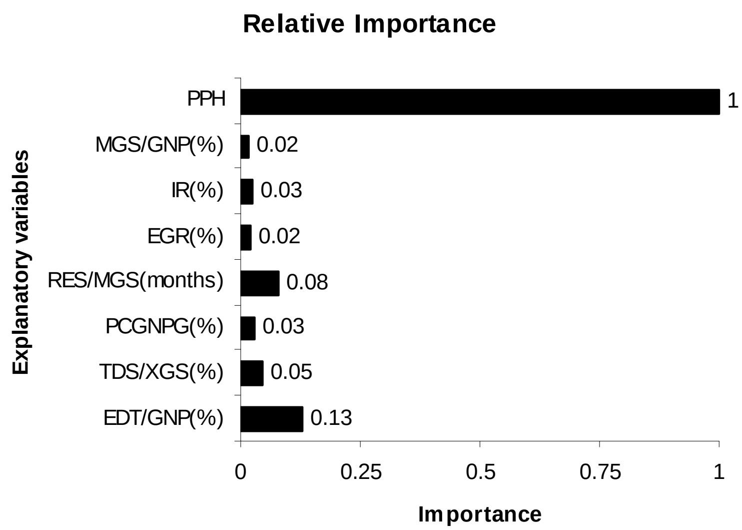
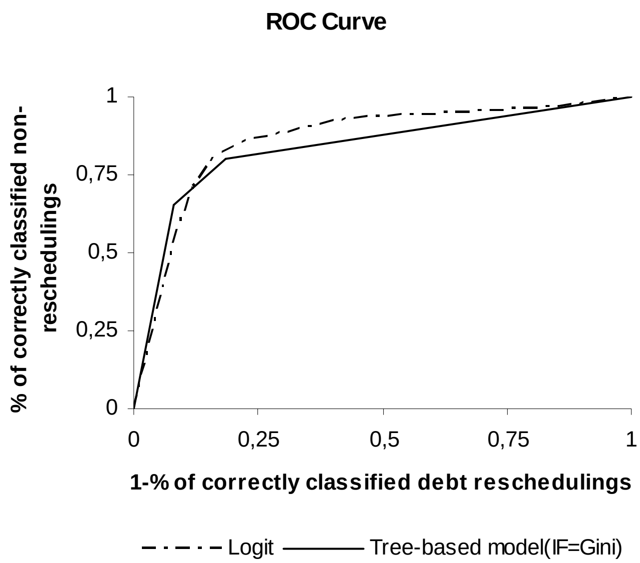


Figure 8. ROC curves for the Logit and the Tree-based model grown under the Gini impurity function in the Test Sample.



References

Barney DK, Alse JA. 2001. Predicting LDC Debt Rescheduling: Performance Evaluation of OLS, Logit, and Neural Network Models. *Journal of Forecasting* **20**: 603-615.

Bradley A. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* **30** (7): 1145-1159.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and Regression Trees*. Wadsworth Inc., Pacific Glove.

Craven P, Wahba G. 1979. Smoothing Noisy Data with Spline Functions – Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numerische Mathematik* **31** (4): 317-403.

Esposito F, Malerba D, Semeraro G. 1997. A Comparative Analysis of Methods for Pruning Decision Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19** (5): 476-491.

Friedman JH. 1991. Multivariate Adaptive Regression Splines. *The Annals of Statistics* **19**: 1-67.

Friedman JH. 1994. An Overview of Predictive Learning and Function Approximation. In *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, Cherkassky, V, Friedman, JH, Wechsler, H. Springer-Verlag.

Galindo J, Tamayo P. 2000. Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics* **15**: 107-143.

Hajian-Tilaki KO, Hanley JA, Joseph L, Collet J. 1997. A Comparison of Parametric and Nonparametric Approaches to ROC Analysis of Quantitative Diagnostics Tests. *Medical Decision Making* **17**: 94-102.

Hamber D. 1975. The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Graph. *Math Psych* **12**: 387-415.

Hand DJ. 1997. *Construction and Assessment of Classification Rules*. John Wiley & Sons, Chichester.

Hanley JA, McNeil BJ. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **143** (1): 29-36.

Hanley, JA, McNeil, BJ. 1983. A Method of Comparing the Areas under Receiver Operating Characteristic (ROC) Curves Derived from the Same Cases. *Radiology* **148** (3): 839-843.

Hastie T, Tibshirani R, Friedman J. 2002. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer-Verlag: New York.

Hospital Ramón y Cajal Homepage, *Material docente de la Unidad de Bioestadística Clínica*. http://www.hrc.es/bioest/M_docente.html#tema5 [1 October 2003].

Lehmann EL. 1998. *Nonparametrics*: *Statistical Methods based on Ranks.* Prentice Hall: New Jersey.

Metz CE, Herman BA, Shen J. 1998. Maximum Likelihood Estimation of Receiver Operating Characteristic (ROC) Curves from Continuously-Distributed Data. *Statistics in Medicine* **17**: 1033-1053.

Sephton P. 2001. Forecasting Recessions: Can We Do Better on MARS™? The Federal Reserve Bank of St. Louis, March/April 2001.

World Debt Tables. Various editions.

World Development Indicators 2002.