Ensemble predictions of recovery rates

João A. Bastos*

CEMAPRE, ISEG, Technical University of Lisbon, 1200-781 Lisboa, Portugal

Forthcoming: *Journal of Financial Services Research*

Abstract

In many domains, the combined opinion of a committee of experts provides better decisions than the judgment of a single expert. This paper shows how to implement a successful ensemble strategy for predicting recovery rates on defaulted debts. Using data from Moody's Ultimate Recovery Database, it is shown that committees of models derived from the same regression method present better forecasts of recovery rates than a single model. More accurate predictions are observed whether we forecast bond or loan recoveries, and across the entire range of actual recovery values.

JEL classification: G17; G21

Keywords: Recovery rate; Loss given default; Forecasting; Ensemble learning; Credit risk

1 Introduction

Prudent people make predictions taking into consideration the opinion of several experts rather than trusting solely their own judgment or that of a single expert. Any predictive

*Tel.: +351 213 922 715; fax: +351 213 922 782; E-mail address: jbastos@iseg.utl.pt

model that produces better forecasts than the average outcome, by incorporating some of the information provided by explanatory variables, can be regarded as an "expert". Therefore, it follows that we should expect a committee of such experts to provide better predictions than a single expert. In this article, I show how to implement a successful ensemble strategy for predicting recovery rates on defaulted securities. This strategy is based on a simple predictor that combines the opinion of an ensemble of structurally similar models estimated on perturbed versions of the original data. This predictor produces more powerful forecasts of creditor recoveries than single models, and presents lower prediction errors across the entire range of actual recoveries. The gains obtained in predictive power are fundamental for lenders and investors wishing to estimate potential credit losses, helping them price and manage credit risk more efficiently.

This article contributes to the expanding strand of literature studying the "Loss Given Default": the loss incurred by creditors when a borrower defaults expressed as a proportion of the exposure at default. This ratio is one of the key elements for calculating regulatory capital requirements in the Basel II framework (Basel Committee on Banking Supervision, 2006). Modeling and forecasting the Loss Given Default or its complement, the recovery rate, is an interesting challenge. First, because the amount recovered after default is expressed as a fraction of the exposure at default, discounted recoveries are usually observed over the interval $[0, 1]$. Second, the distributions of discounted recoveries are frequently bimodal, containing many observations with very low recoveries and many with complete or near complete recoveries. A straightforward approach for modeling recoveries is to estimate a linear least squares regression (see, e.g. Acharya et al., 2007; Caselli et al., 2008; Grunert and Weber, 2009). A limitation of the linear model is that its support is the real line and, therefore, it does not guarantee that predicted values lie over the unit interval. This inconvenience may be overcome by fitting a (nonlinear) fractional regression estimated using quasi-maximum likelihood methods (Papke and Wooldridge, 1996). For instance, in the context of credit losses this approach is adopted by Dermine and Neto de Carvalho (2006). Alternatively, one may estimate a linear model on appropriately transformed recoveries. The most eminent example of this approach is Moody's LossCalc v2 (Gupton and Stein, 2005), in which recoveries are normalized via a beta distribution and a linear least squares regression is carried out on the transformed data. More recently, Bastos (2010) and Loterman et al. (2012) suggest that nonparametric predictive models, such as regression trees and neural networks, outperform parametric models in forecasting recoveries.

While the exploration of alternative models is fundamental for lenders wishing to better predict future credit losses, theoretical work in the area of machine learning suggests that a diverse ensemble of predictors may have greater predictive accuracy than a single predictor. Given a set of observations, how can we create an ensemble of predictors? One approach consists of using these observations to estimate a handful of models derived by different regression techniques, and combine the predictions in line with a convenient scheme. An alternative and often more powerful approach consists of creating a diversity of data sets by introducing randomness into the original data set, estimating models on the new data sets using the *same* regression technique, and then combine their predictions. The most basic procedure for perturbing data and combining predictors was proposed by Breiman (1996), giving it the acronym "bagging" (for bootstrap aggregating). This approach generates new data sets by sampling with replacement (i.e. bootstrapping) observations from the data. The new samples will not be identical to the primordial sample: some observations may be repeated while others may be left out. Then, for each bootstrap sample a predictive model is estimated using a single regression technique. The final prediction is a simple average of the individual predictions.

Since the same regression technique is applied to different data sets, the success of this strategy depends crucially on how the technique responds to small perturbations in the data. The resulting predictive models cannot be *too* similar, since numerous copies of the same forecast contain exactly the same amount of information as a single forecast. Therefore, this strategy must ensure diversity among the models it combines, and this diversity is only feasible if the regression method exhibits some instability with respect to changes in the data. Furthermore, a decision on the number of experts to be included in the ensemble must be made. The size of the committee is usually determined by minimizing a loss function such as the mean squared error. In general, substantial gains in predictive power are achieved by committees with few members, and the marginal impact of each additional expert quickly vanishes.

In the area of credit risk, several empirical studies have used ensemble strategies. However, these studies focused on bankruptcy prediction and credit scoring. The reader is referred to Verikas et al. (2010) for a survey of the literature. In the following sections, I explore the effectiveness of an ensemble strategy for forecasting recovery rates using data from Moody's Ultimate Recovery Database. This database covers US non-financial corporations and provides information on 4630 bonds and loans that defaulted in the period from 1987 to 2010. After providing an overview of the data, I discuss the choice of the regression method and the optimal size of the ensemble. Then, the out-of-sample predictive accuracy of the ensembles is evaluated using several accuracy measures. Results for the entire data set, the subsample of bonds, and the subsample of loans are reported. This is followed by an examination of which features of the defaulted debts determine the ensemble predictions of their recovery rates. Afterward, it is shown that when the regression method is stable with respect to perturbations in the data, this ensemble strategy is ineffective for forecasting recoveries. Finally, it is shown that ensembles of models estimated with a single regression method outperform a committee of two models estimated with different regression techniques.

2 Sample characteristics and variables

The data sample is Moody's Ultimate Recovery Database (URD), which covers US non-financial corporations holding over $50 million in debt at the time of default. The sample describes 4630 defaulted bonds and loans from 957 different issuers, covering the period from 1987 to 2010. Moody's provides three alternative valuations of nominal recoveries at the time of resolution. These are based on the settlement value taken at or close to default, the trading prices of the defaulted instruments at or post-emergence, and the value of the settlement instruments taken at the time of a liquidity event (Emery et al., 2007). The database indicates which recovery value Moody's considers to be the most representative of the actual recovery. The analysis is conducted using the recovery value recommended by Moody's. Discounted recoveries are obtained by discounting back to the last time interest was paid using the instrument's pre-petition coupon rate.

Table 1 reports the number of instruments and the mean discounted recovery rate broken down by year of default, Moody's industry, instrument type and collateral. Panel A shows that there is considerable variation in default rates across time. The number of defaulted instruments in the database increases in the recessions of the early 1990s, early 2000s and late-2000s. There is also substantial variation in mean recovery rates across time, and decreased recoveries are observed during the economic downturns.[1] Panel B reports the sample distribution and mean recovery rate broken down by Moody's industry classification. Again, there are big differences in recovery rates across industries. The highest mean recovery is observed in the Natural products sector, while the lowest is found in the Environment sector.

Panel C provides the sample breakdown by instrument type. Bonds make up approximately 60% of total instruments with loans comprising the remaining 40%. Among loans,

[1] For a discussion on the importance of the negative correlation between default probabilities and recovery rates see, e.g. Das (2007).

	Obs.	RR		Obs.	RR
Panel A. By Year					
1987	23	76%	1999	184	57%
1988	64	55%	2000	271	51%
1989	98	46%	2001	572	52%
1990	150	48%	2002	783	49%
1991	226	58%	2003	399	70%
1992	190	62%	2004	206	75%
1993	138	62%	2005	203	76%
1994	66	69%	2006	75	74%
1995	96	66%	2007	47	75%
1996	83	64%	2008	197	66%
1997	64	65%	2009	370	63%
1998	68	47%	2010	57	63%
Panel B. By Moody's industry					
Automotive	204	62%	Manufacturing	427	64%
Chemicals	74	64%	Media	358	64%
Construction	68	50%	Metals & mining	141	57%
Consumer products	385	66%	Natural products	93	82%
Distribution	519	52%	Other	67	57%
Energy	493	74%	Services	337	59%
Environment	51	30%	Technology	146	61%
Health care	157	56%	Telecommunications	469	42%
Industrials	69	67%	Transportation	314	50%
Leisure & entertainment	258	62%			
Panel C. By instrument type					
Junior Subordinated Bonds	69	18%	Term Loan	883	76%
Senior Subordinated Bonds	493	29%	Revolver	963	85%
Subordinated Bonds	372	29%			
Senior Unsecured Bonds	1263	49%			
Senior Secured Bonds	587	64%			
Panel D. By collateral type					
All or most assets	1348	82%	PP&E	342	59%
Capital Stock	183	69%	Second and third lien	204	55%
Inventory, accounts receivable & cash	218	96%	Unsecured	2273	41%
Other	62	84%			

Table 1: Summary statistics of Moody's Ultimate Recovery Database. This table shows the number of instruments (Obs.) and mean discounted recovery rate (RR), by year of default (Panel A), Moody's industry (Panel B), instrument type (Panel C) and collateral type (Panel D).

the data is roughly split between revolvers and term loans. On average, bank loans recover better than bonds, reflecting the typically higher position of loans in terms of claim priority. Furthermore, only 14% of the loans in the sample are not secured by any type of collateral. Revolvers exhibit higher mean recovery than term loans. This may be attributed to the greater frequency with which revolving lines of credit are monitored by lenders. The sample breakdown by collateral type is shown in Panel D. On average, debts secured by inventory, accounts receivable and cash exhibit higher recoveries, since these assets are easier to liquidate. Not surprisingly, unsecured instruments present the lowest mean recovery rate.

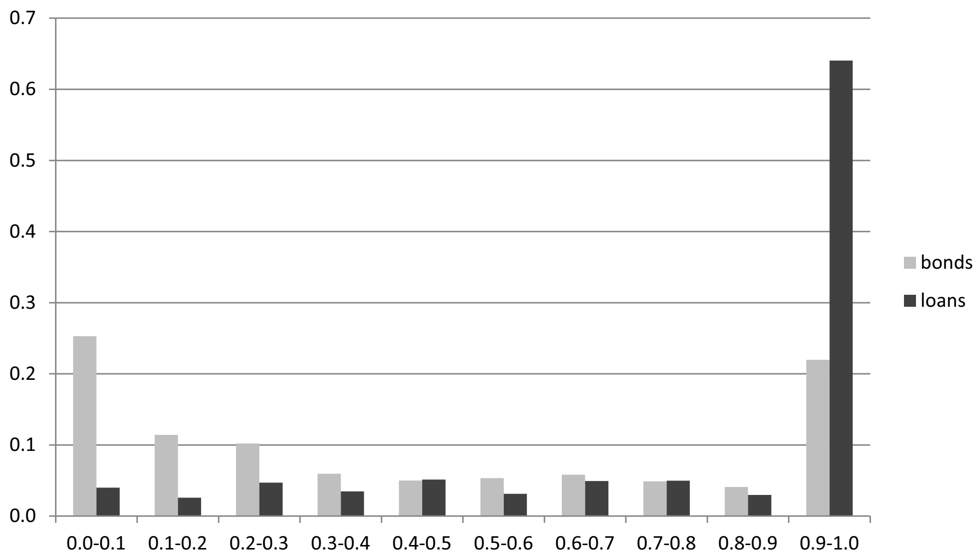

Figure 1: Distribution of discounted recovery rates for bonds and loans in Moody's Ultimate Recovery Database (1987–2010).

Figure 1 shows the distributions of discounted recovery rates for loans and bonds. The distribution for loans has a strong negative skew. More than 60% of these instruments resulted in complete or near complete recovery. On the other hand, the distribution of bond recovery rates is bimodal. In particular, the probability that defaulted bonds recover very little is substantial.

Table 2 shows which variables determine ultimate recoveries in a multivariate setting.

	coef.	p-value
Moody's Industry		
Automotive	0.214	<0.001
Chemicals	0.194	<0.001
Construction	0.134	0.001
Consumer products	0.224	<0.001
Distribution	0.117	<0.001
Energy	0.362	<0.001
Environment	-0.110	0.004
Health care	0.160	<0.001
Industrials	0.219	<0.001
Leisure & Entertainment	0.206	<0.001
Manufacturing	0.235	<0.001
Media	0.254	<0.001
Metals & Mining	0.144	<0.001
Natural products	0.399	<0.001
Other	0.189	<0.001
Services	0.220	<0.001
Technology	0.169	<0.001
Transportation	0.129	<0.001
Instrument type		
Revolver	0.219	<0.001
Term loan	0.188	<0.001
Senior secured bonds	0.141	0.002
Senior unsecured bonds	0.179	<0.001
Senior subordinated bonds	-0.001	0.989
Subordinated bonds	0.053	0.178
Collateral type		
All or most assets	0.086	0.001
Capital stock	0.051	0.106
Inventory & accounts receivable	0.161	<0.001
Other	0.105	0.011
PP&E	0.043	0.157
Second & Third Lean	0.011	0.727
Seniority		
Percentage above	-0.146	<0.001
Percentage below	0.388	<0.001
Instrument ranking	-0.024	0.004
Intercept	0.183	<0.001

Table 2: Determinants of ultimate recovery rates. This table shows which variables determine Moody's ultimate recovery rates in a multivariate setting. The coefficients and corresponding p-values were estimated by ordinary least squares with robust standard errors. The data sample consists of 4630 defaulted bonds and loans from 957 different issuers, covering the period from 1987 to 2010.

The coefficients and corresponding p-values were estimated by ordinary least squares with robust standard errors. The regressors include dummy variables for industry, collateral and instrument type. Furthermore, because there is strong empirical evidence that seniority has a strong impact on recoveries (see, e.g. Varma and Cantor, 2005), the following variables, readily available in the database, are also included in the regression: percentage above (the percentage of obligor's debt senior to the instrument); percentage below (or "debt cushion", the percentage of obligor's debt junior to the instrument); and the instrument rank within the obligor's liability structure (an ordinal variable in which lower values correspond to higher priority).

The reference group for Moody's industry dummies is the Telecommunications sector since it represents a large proportion of the debts and features the second lowest mean recovery rate. Most industries have statistically significant larger recoveries than the reference group. On the other hand, the Environment sector has statistically significant lower recoveries than the Telecommunications sector. The reference group for the instrument type dummies is the junior subordinated bonds. Table 2 suggests that other subordinated bonds do not have significantly different recoveries with respect to that group. On the other hand, the remaining instrument types have significantly higher recoveries at conventional levels.

With respect to collateral, debts protected by all or most assets, and inventory & accounts receivable have statistically significant higher recoveries than unsecured securities (the reference group). Also, collateral types gathered in the "other" category (which includes those with very few observations, such as real estate, oil and gas properties, and intellectual property) have statistically significant higher recoveries with respect to unsecured debts. Finally, the variables related to the instrument's priority in the liability structure have statistically significant effects on recoveries. As expected, percentage above and the instrument rank have a negative impact on recoveries, while percentage below has a positive effect.

3 Ensemble models of recovery rates

3.1 Bagging predictors

As mentioned in the introduction, a committee of experts may be obtained by creating a diverse ensemble of data sets from the primordial data. The simplest way to achieve this is to bootstrap the available observations to create new samples, and calculate the average of the predictions of models estimated on these samples using a single regression technique, which is labeled “base model”.

Let $\mathcal{L}$ denote a set of n observations $\{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ independently drawn from a probability distribution, where y_i is the response variable and $\mathbf{x}_i$ is a set of predictor variables. Furthermore, let $\phi(\mathbf{x}, \mathcal{L})$ denote a predictor for y estimated using $\mathcal{L}$. Now, assume that we have several datasets $\{\mathcal{L}_k\}$ each consisting of n independent observations drawn from the same underlying probability distribution. We can calculate the average of the predictions given by the models estimated with the individual datasets $\{\mathcal{L}_k\}$ to obtain an aggregated predictor for y,

$$\phi_A(\mathbf{x}) = E_{\mathcal{L}}\left[\phi(\mathbf{x}, \mathcal{L})\right]. \tag{1}$$

The mean-squared error, averaged over $\mathcal{L}$, of the single predictor is

$$\begin{aligned}
&E_{\mathcal{L}}\left[(y - \phi(\mathbf{x}, \mathcal{L}))^2\right] \\
&\quad = y^2 - 2yE_{\mathcal{L}}\left[\phi(\mathbf{x}, \mathcal{L})\right] + E_{\mathcal{L}}\left[\phi(\mathbf{x}, \mathcal{L})^2\right] \\
&\quad = y^2 - 2yE_{\mathcal{L}}\left[\phi(\mathbf{x}, \mathcal{L})\right] + E_{\mathcal{L}}\left[\phi(\mathbf{x}, \mathcal{L})\right]^2 + \mathrm{Var}_{\mathcal{L}}\left[\phi(\mathbf{x}, \mathcal{L})\right] \\
&\quad = (y - \phi_A(\mathbf{x}))^2 + \mathrm{Var}_{\mathcal{L}}\left[\phi(\mathbf{x}, \mathcal{L})\right].
\end{aligned} \tag{2}$$

Since in standard situations the variance of the predictor $\phi(\mathbf{x}, \mathcal{L})$ is positive, the squared error of the ensemble of predictors $\phi_A(\mathbf{x})$ is lower than the mean-squared error averaged over

$\mathcal{L}$ of a single predictor $\phi(\mathbf{x}, \mathcal{L})$. Moreover, the greater the instability of $\phi(\mathbf{x}, \mathcal{L})$ with respect to $\mathcal{L}$, the greater the difference between these errors will be.

Of course, in practice we have only one data set, and collecting more data is often prohibitive or simply impossible. The approach suggested in Breiman (1996) is to generate several data sets $\{\mathcal{L}^{(B)}\}$, each consisting of n observations sampled with replacement from the available data set. If N sets are generated, the aggregated predictor becomes

$$\phi_A(\mathbf{x}) \to \phi_B(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}, \mathcal{L}_i^{(B)}). \tag{3}$$

This predictor will perform better than a single predictor if: 1) the performance of the models estimated with bootstrap samples is not much worse than that of the model estimated with the original data set; 2) the models estimated with different bootstrap samples produce significantly different forecasts.

Below I show the pseudocode for the bagging procedure:

Input:

data set: $\mathcal{L} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n)\}$

base model: $\phi(\mathbf{x}, \mathcal{L})$

number of ensemble members: N

Process:

for $i = 1, 2, \ldots, N$

  sample with replacement n observations from $\mathcal{L}$ to obtain $\mathcal{L}_i^{(B)}$

  estimate predictor $\phi(\mathbf{x}, \mathcal{L}_i^{(B)})$

end for

Output:

$\phi_B(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}, \mathcal{L}_i^{(B)})$

3.2 Choosing the base model

Equation 2 shows that a crucial factor in creating a powerful ensemble is the stability of the base model. If slight perturbations in the data produce small changes in the estimated models, ensembles fail to provide significantly better forecasts than a single model. Therefore, ensembles are most effective when the base model has some instability with respect to perturbations in the data. For reasons that will become evident below, a good choice for the base model is a *regression tree* (Breiman et al., 1984). This is a fortunate circumstance, since this class of predictive models has shown good accuracy in forecasting credit recoveries (Bastos, 2010).

A regression tree is a model in which the data are recursively partitioned into smaller mutually exclusive subsets, and the partitions are represented by a sequence of logical *if-then-else* tests on the attributes of the observations. The algorithm begins with a "root" node containing all observations. It then searches all the possible binary splits of the data in order to find the explanatory variable and corresponding cut-off value that minimize the intrasubset variance of the response variable in the newly created daughter nodes. That is, the response variable will be more homogeneous in daughter nodes than in their parents. This procedure is then repeated for new daughter nodes until the reduction of variance is very small (e.g. less than 1% of the variance of the complete data set) and/or very few observations remain in the node. Starting from the upper-most node, observations are routed down the tree according to the values of the explanatory variables tested in successive nodes and, inevitably, end their path in a terminal node. The average value of the response variable in a terminal node will be the predicted value for new observations that reach that node.

Figure 2 illustrates the recursive segmentation of Moody's URD data using the tree induction algorithm. For illustrative purposes only, the splitting process was stopped earlier imposing a large, and not necessarily optimal, lower limit on the number of observations in the nodes. This tree splits the data into five distinct regions determined by three attributes

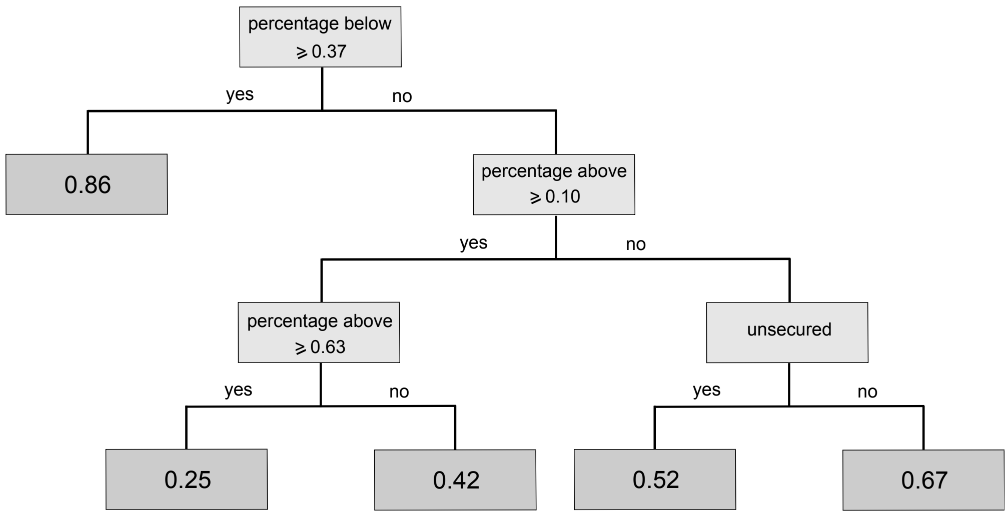

Figure 2: Regression tree to predict recovery rates from Moody's Ultimate Recovery Database. This tree was estimated with the 4630 observations in the database.

of the debts.[2] At the upper-most node it is asked if the instrument's percentage below is greater than 0.37. If the answer is positive, the predicted recovery is 0.86. If the opposite occurs, it is further inquired if the instrument's percentage above is greater than 0.10. If this is true, the tree examines the percentage above again. If the percentage above is smaller than 0.63 (and greater than 0.10) the predicted recovery is 0.42, otherwise it is 0.25. If the percentage above is smaller than 0.10, the question is asked if the debt is unsecured. If the answer is positive, the predicted recovery is 0.52; if the instrument is secured by collateral the predicted recovery is 0.67.

This simple model allows several conclusions to be drawn. First, higher debt cushions are associated with large recoveries, since the branch "percentage below ≥ 0.37" leads to a terminal node in which the predicted recovery is greater than those in the remaining nodes. Second, percentage above is negatively related to recoveries since the branch "percentage above < 0.10" leads to predicted recoveries that are larger than those in the opposite branch.

[2]The optimal lower limit on the number of observations in the nodes may be very low, even when the model precision is estimated out-of-sample to prevent over-fitting the data. This generates highly complex trees when the sample size is large. For Moody's URD, I found an optimal lower limit of just 2 observations per node, which segments the data into 170 distinct regions.

Furthermore, the branch “percentage above < 0.63” leads to a terminal node in which the predicted recovery is greater than that in the opposite node. The same reasoning leads us to the conclusion that unsecured debts are associated with lower recoveries.

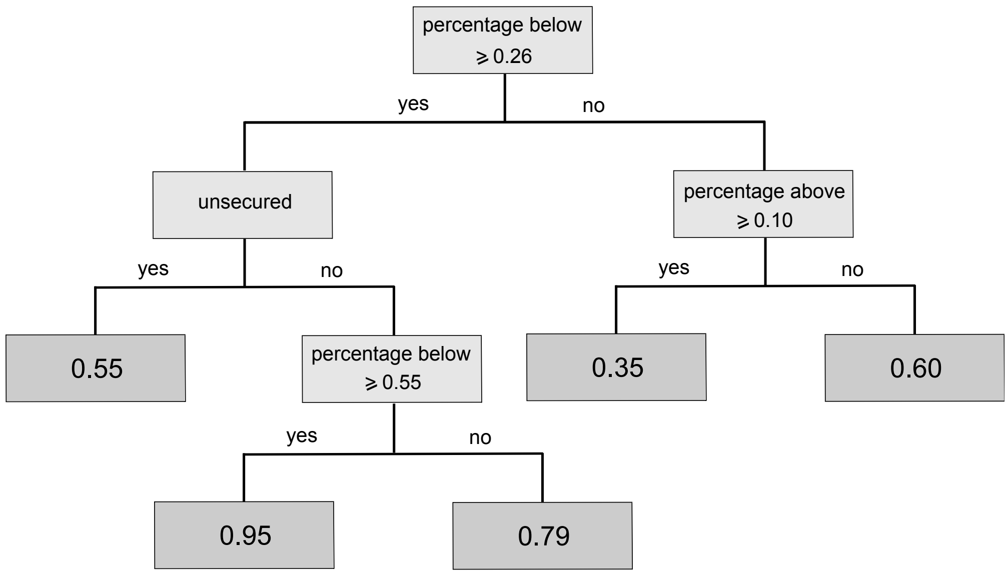

Figure 3: Regression tree estimated with a bootstrap sample consisting of 4630 observations sampled with replacement from the complete data set. The comparison of this tree with that in Figure 2 illustrates a crucial feature of regression tree models: small changes in the data generate different tree structures, and different forecasts for the response variable (the values in the terminal nodes are different from those in Figure 2). An ensemble is a set of regression tree models estimated with different bootstrap samples. The random nature of the sampling procedure guarantees that these trees have different structures and provide different forecasts.

The recursive segmentation of the data induced by this algorithm implies that the estimated model will be rather unstable with respect to perturbations in the data. In fact, small changes in the data may easily result in a different variable or cut-off value being chosen at a particular node, and a significantly different structure for the subtree beneath that node (Witten and Frank, 2005). Figure 3 illustrates this point by showing a regression tree estimated with a bootstrap sample, and imposing the same lower limit on the number of observations in the nodes as for the tree estimated with the full data. Comparing this tree with the one in Figure 2, it can be seen that the explanatory variable at the upper-most node is still the instrument's percentage below. However, the optimal cut-off value now is 0.26. This slightly different cut-off value leads to substantially different subtrees beneath this node. The branch on the left is not a terminal node but a question of whether the debt is unsecured. The branch on the right leads to a test on the debt's percentage above with the same cut-off value as in the tree estimated with the full data. However, for the bootstrap sample this test ends the splitting process.

As a concluding remark, an additional reason for electing the regression tree for the base model is that its predictions are averages of the response variable. Because recovery rates are bounded to the interval [0,1], regression trees give predictions that also lie in this interval. Furthermore, as the ensemble predictions are averages of the individual base model predictions, they will be bounded to [0,1] as well.

3.3 Choosing the number of ensemble members

After electing a base model, one must decide how many experts should be included in the committee. This decision is based on the minimization of a loss function, such as the mean squared error. Let y and $\hat{y}$ denote the actual and predicted recovery rates, respectively, and n denote the number of debts in the sample. The mean squared error (MSE) is defined as

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{4}$$

Models with lower MSE tend to give smaller differences between the actual and predicted recoveries and, on average, predict actual recoveries more accurately.

Figure 4 shows the out-of-sample MSE of ensembles of regression trees as a function of the size of the committee.[3] While this plot refers to ensembles estimated using the entire data

[3]Out-of-sample estimates of errors on unseen data are obtained through 10-fold cross-validation. In this procedure, the data are divided into 10 groups of approximately the same size. Nine groups are used for estimation and one group is used for evaluating an out-of-sample error. Each of the 10 groups is in turn set aside to serve temporarily as an independent test sample. Then, the out-of-sample errors of the 10 samples

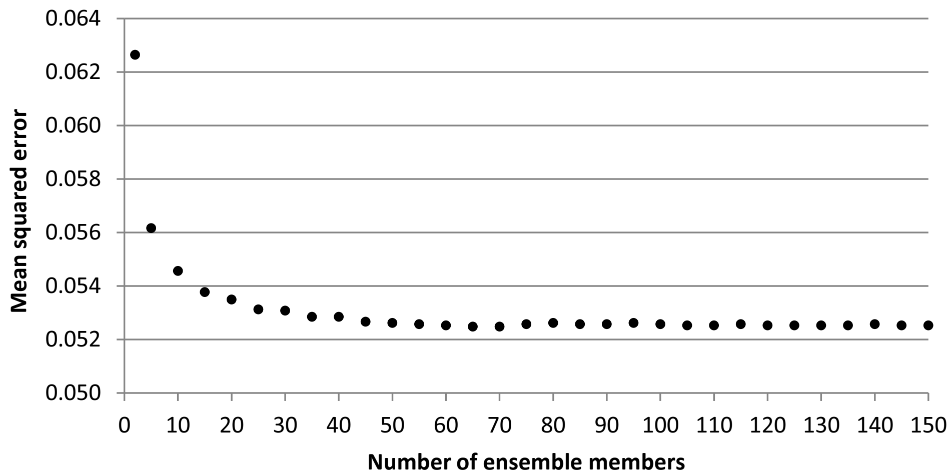

Figure 4: Out-of-sample mean squared error of recovery rate forecasts given by an ensemble of regression trees as a function of the number of trees in the committee.

from Moody's URD, the results are similar when bonds and loans are considered separately. The regressors are the dummy variables for industry, collateral and instrument type, and the variables describing the seniority of the debt within the obligor's liability structure that were introduced in Section 2.

Figure 4 shows that the prediction error of the combined "opinions" decreases as the number of experts in the committee increases. These errors should be compared with the out-of-sample MSE of a single tree estimated with the full data which is 0.068. Clearly, even an ensemble with very few members has better accuracy than a single tree. However, substantial gains in predictive power are obtained by the first few members, and the marginal impact of each additional opinion decreases rapidly. If computing time is an issue, practical results are obtained with just 15 or 20 members. After the inclusion of around 60 members, the out-of-sample error stabilizes and further improvements in predictive accuracy are unattainable. The size of the committee should be greater than this critical number.

are combined to obtain an estimate of the error on unseen data using all available observations.

3.4 Evaluating predictive accuracy

In addition to the mean squared error introduced in Equation 4, I consider the following measures of predictive accuracy. The mean absolute error (MAE) is defined as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \,. \tag{5}$$

Models with lower MAE also predict actual values more accurately, on average. However, the MSE puts more weight on large errors than on small ones, whereas the MAE does not. Therefore, the MSE should be preferred over MAE when large prediction errors are more damaging. Relative errors measure the predictive accuracy with respect to a simple "model" that always predicts the average outcome of the response variable, thereby neglecting the information provided by the explanatory variables. The relative squared error (RSE) and the relative absolute error (RAE) are defined as:

$$\mathrm{RSE} = 100\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\ (\%), \quad \mathrm{RAE} = 100\frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i - \bar{y}|}\ (\%). \tag{6}$$

Models with RSE and RAE smaller than 100% provide, on average, better forecasts than the simple predictor in terms of squared and absolute error, respectively. The last measure of predictive accuracy is the statistical correlation between y and $\hat{y}$:

$$\rho_{y,\hat{y}} = \frac{\mathrm{Cov}(y,\hat{y})}{\sqrt{\mathrm{Var}(y)\mathrm{Var}(\hat{y})}}. \tag{7}$$

Of course, the model with the highest correlation outperforms the others.

Table 3 shows the out-of-sample accuracy measures for recoveries predicted by a single tree and an ensemble of trees. The ensemble contains 100 members, but any committee size within the range of the plateau in Figure 4 could be chosen. For the sake of completeness, the results for a linear least squares regression are also shown. In addition, Table 3 shows

Panel A: All data					
	Linear model	Single tree	Ensemble	LM-ST	ST-E
Mean squared error	0.086	0.068	0.053	-21%	-22%
Mean absolute error	0.238	0.179	0.163	-25%	-9%
Relative squared error (%)	56.83	44.81	34.79	-21%	-22%
Relative absolute error (%)	66.73	50.14	45.70	-25%	-9%
Correlation coefficient	0.657	0.746	0.809	14%	8%
Panel B: Bonds					
	Linear model	Single tree	Ensemble	LM-ST	ST-E
Mean squared error	0.099	0.078	0.060	-21%	-23%
Mean absolute error	0.260	0.198	0.179	-24%	-10%
Relative squared error (%)	70.46	55.54	42.85	-21%	-23%
Relative absolute error (%)	77.20	58.61	52.94	-24%	-10%
Correlation coefficient	0.543	0.673	0.758	24%	13%
Panel C: Loans					
	Linear model	Single tree	Ensemble	LM-ST	ST-E
Mean squared error	0.060	0.056	0.043	-7%	-23%
Mean absolute error	0.189	0.152	0.141	-19%	-7%
Relative squared error (%)	65.89	61.75	47.02	-7%	-23%
Relative absolute error (%)	75.03	60.51	56.11	-19%	-7%
Correlation coefficient	0.583	0.625	0.730	7%	17%

Table 3: Comparison of predictive accuracy. Out-of-sample predictive accuracy measures of recovery rate forecasts given by a linear least squares regression, a single regression tree and an ensemble of 100 regression trees. The last two columns show the percent variation of the accuracy measures between the linear model and a single tree (LM-ST), and between the single tree and the ensemble (ST-E). Panel A refers to the entire data set, panel B refers to a subsample including bonds only, and panel C refers to the subsample including loans only.

the accuracy measures when we consider the entire data set (panel A), the subsample of bonds (panel B) and the subsample of loans (panel C).

A single tree gives better forecasts of recoveries than the linear model across all measures. For instance, a single tree decreases the out-of-sample MSE of the complete data set by about 21% compared to the linear model. The corresponding numbers for the subsamples of bonds and loans are 21% and 7%, respectively. This is not unexpected, since the linear model forces a linear dependence between recovery rates and the explanatory variables, while the regression tree does not assume any underlying relationship, and is capable of detecting nonlinear dependencies.

When we compare ensembles of trees with single trees we find even bigger improvements in predictive accuracy. Across all measures and across the three data sets the ensembles outperform the single tree models. In fact, further reductions in MSE of about 22%, 23% and 23% are observed in the full sample, the subsample of bonds, and the subsample of loans, respectively. The corresponding reductions in terms of MAE are 9%, 10% and 7%. Accordingly, the ensembles present lower relative errors across the three data sets. Finally, the recoveries predicted by the ensembles also show higher correlation with actual recoveries.

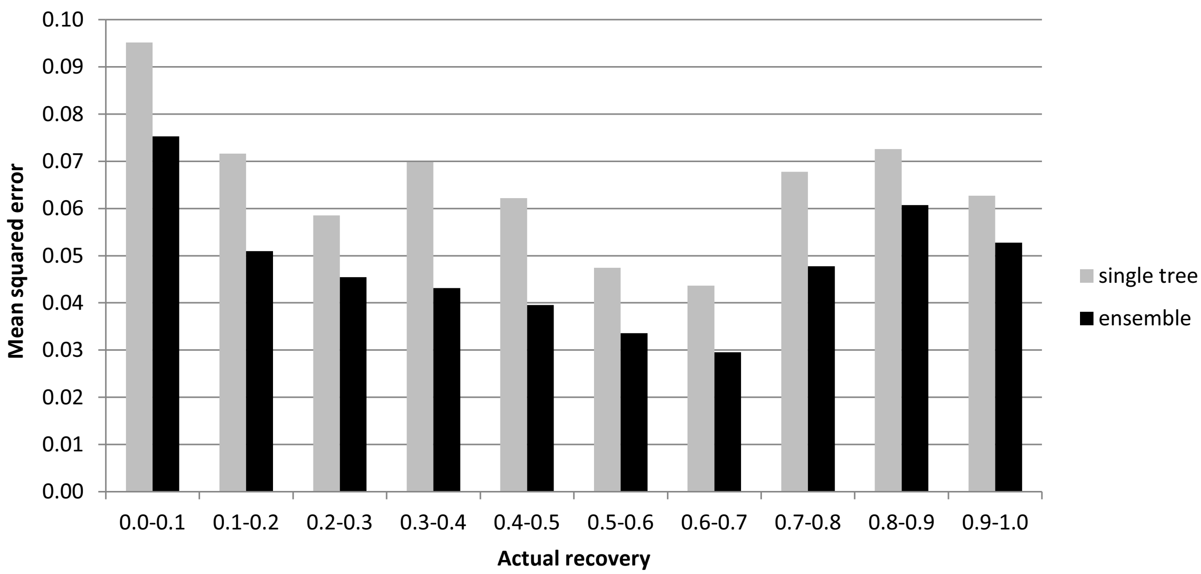

Figure 5: Out-of-sample mean squared errors versus actual recoveries given by a single tree and an ensemble of trees.

Figure 5 shows out-of-sample mean squared errors versus actual recoveries given by a single tree and the ensemble of trees. These results refer to the full data set. The ensemble presents lower out-of-sample prediction errors across all recovery values. In particular, the ensemble is more powerful at predicting the low recoveries which are more damaging to the lender.

Also of note is that the errors are systematically biased for the loans sample. This is due to the large number of defaulted loans with full recovery (see Figure 1), and due to predicted values being bounded to the interval [0,1]. Defining the error as $\varepsilon = y - \hat{y}$, about 70% of the loan recovery rate forecasts have positive errors (the ensemble underestimates the recovery rate). On the other hand, the distribution of recoveries for the bonds sample is more symmetrical. For this sample, about 46% of the bond recovery rate forecasts have positive errors.

3.5 Which variables determine ensemble forecasts?

One potential disadvantage of creating ensembles of experts is that the final prediction rule is substantially more complex than that of a single expert. Therefore, it is more difficult to understand which attributes of the observations are contributing to the improved forecasts.

When regression trees are used as the base model, a straightforward approach to understand the relative importance of the explanatory variables is to calculate the frequency with which those variables appear in the decision nodes of the trees in the committee. This information can be complemented with the inspection of individual trees or the coefficients of a parametric model in order to discern the direction of the partial effects. Figure 6 reports this frequency for the ensembles estimated with data from Moody's URD. The bars in black, dark gray, and light gray, correspond to the full sample, and the subsamples of bonds and loans, respectively.

Overall, the relative participation of the industry, collateral and instrument type dummies in the decision nodes is rather small. On the other hand, the variables that characterize the priority of the debts in the liability structure frequently participate in the decision nodes. These variables also reveal an interesting asymmetry between bonds and loans. For the subsample of bonds, percentage above is the most important variable. This is not unexpected, since bonds tend to reside at the bottom of the obligor's liability structure in terms of claim priority. Therefore, the percentage of debt senior to these debts must have a significant (and negative) impact on how they recover. Conversely, percentage below is the

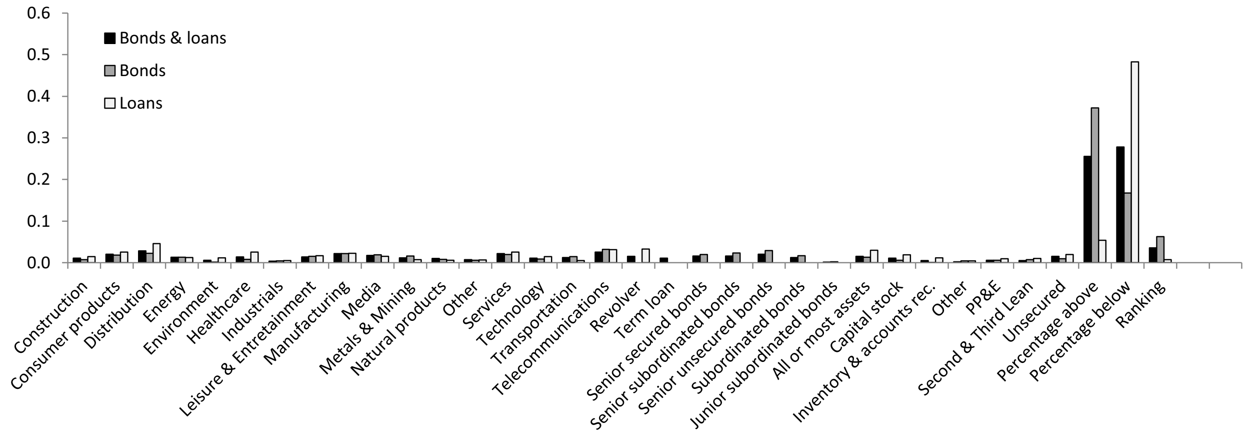

Figure 6: Frequency with which explanatory variables are employed in the decision nodes of tree ensembles. The bars in black, dark gray, and light gray, correspond to the full sample, and the subsamples of bonds and loans, respectively.

most important variable for loans. Given that loans tend to reside at the top of the liability structure, the debt cushion below these instruments must have a significant (and positive) impact on how they recover.

3.6 Ensembles of linear models

The instability of regression trees in response to small perturbations in the data was the reason they were adopted as the base model. Nevertheless, how effective is an ensemble strategy if the base model is fairly robust to small changes in the data? Figure 7 sheds light on this question. It shows how the out-of-sample mean squared error of recovery rate forecasts given by an ensemble of linear least squares regressions evolves as a function of the number of members in the committee.

First, the MSE as a function of the size of the ensemble is over a very limited range. Because the OLS regression is fairily robust with respect to changes in the data, linear models estimated with different bootstrap samples are very similar to each other and produce very

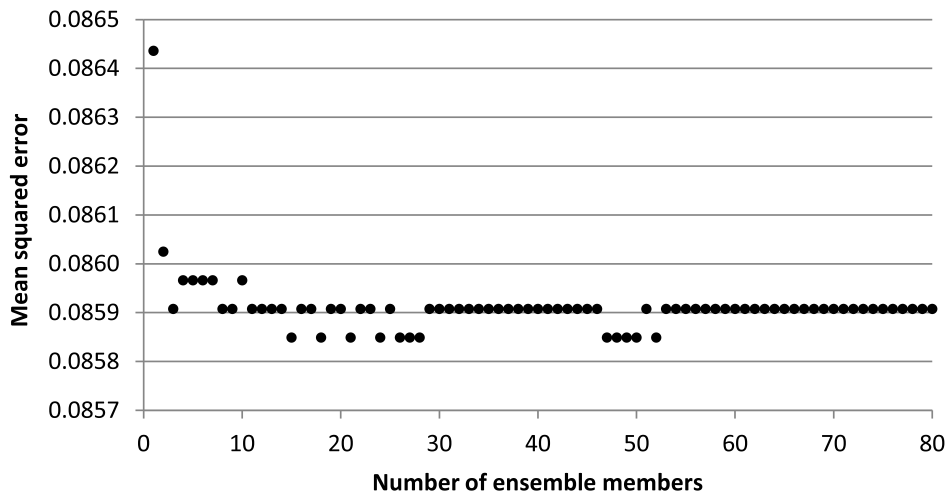

Figure 7: Out-of-sample mean squared error of recovery rate forecasts given by an ensemble of linear least squares regressions as a function of the number of members in the ensemble.

similar forecasts. Therefore, ensemble forecasts are also very similar regardless of the size of the ensemble. An apparent and rather small gain in predictive power is observed when the size of the committee increases. However, Figure 7 is somewhat misleading. First, the out-of-sample MSE of an ensemble with few members is actually larger than the out-of-sample MSE of a single linear model estimated with the full data (0.08591). In fact, when the committee contains very few linear models, we are just averaging predictions from very similar models that were estimated with just a fraction of the data.[4] As the size of the committee increases, and eventually all observations in the primordial data set are used in the estimation, the accuracy of these ensembles tends toward that of the single linear model estimated with the full data.

[4]The bootstrap samples leave many observations out. For reasonably large data sets such as Moody's URD, on average 36.8% of the observations in the bootstrap samples are duplicates.

3.7 Committee of different regression techniques

If the regression technique is unstable with respect to different bootstrap samples, the members of the ensemble are significantly different from each other, and can be regarded as distinct experts. Therefore, we may regard an ensemble as a large committee of experts with similar expertise, but sufficiently different to complement one another. On the other hand, the combination of different regression techniques estimated with the same data may be regarded as a small committee of experts with different kinds of expertise. Table 4 reports the out-of-sample predictive accuracy measures given by a model in which the predicted recoveries are the average of the predictions of two different regression techniques: a linear least squares regression and a regression tree. Comparing these measures with those in Table 3, we find that the combined forecasts of this committee are better than the individual forecasts of the linear model. However, they are barely comparable to the predictions of the most skilled expert: the regression tree. Consequently, the forecasts of this committee are outperformed by those of the ensemble of regression trees.

	All data	Bonds	Loans
Mean squared error	0.067	0.077	0.053
Mean absolute error	0.201	0.221	0.167
Relative squared error (%)	44.28	54.86	57.91
Relative absolute error (%)	56.53	65.64	66.40
Correlation coefficient	0.748	0.675	0.648

Table 4: Predictive accuracy of a committee of two regression techniques. Out-of-sample predictive accuracy measures of recovery rate forecasts given by a model in which the predicted values are the average of the predictions of a linear least squares regression and a regression tree.

4 Conclusions

This article shows that the combined opinion of a committee of models provides better forecasts of recovery rates than single models. The analysis is based on a simple ensemble strategy called “bagging”. This procedure generates new data sets by bootstrapping observations from the original data, and calculating the average of the predictions of models estimated on these samples using a single regression method. Because the effectiveness of this strategy depends on the instability of the regression technique in response to small perturbations in the data, a decision tree induction algorithm is chosen for the base model.

Using data from Moody’s Ultimate Recovery Database, it is shown that the recovery rate forecasts given by an ensemble of regression trees outperform the forecasts given by a single regression tree, whether we consider the entire data set or samples containing only bonds or loans. Also, the ensembles outperform a single tree across all actual recovery values. In particular, the ensembles are more powerful at predicting low recoveries, which are more ruinous to lenders and investors. It is also shown that an ensemble of linear models fails to increase predictive performance compared to a single linear model. Moreover, tree ensembles have greater predictive power than a committee formed by a linear least squares regression and a regression tree.

One potential drawback of ensembles is that it may not be easy to understand which variables are contributing to the improved forecasts. While it is often the case that simplicity has to be sacrificed in order to achieve a higher degree of precision, if we calculate the frequency with which explanatory variables are employed in the decision nodes of the ensemble we can understand in intuitive terms the relative importance of the variables in the model. This approach showed that the variables describing the instrument’s position in the liability structure in terms of claim priority contribute more often to the recovery rate forecasts provided by the ensemble.

Acknowledgments

I would like to thank João Andrade e Silva, Telmo Pina e Moura, Joaquim Ramalho and an anonymous referee for their valuable comments and suggestions. All the remaining errors are mine. Financial support from FCT - Fundação para a Ciência e Tecnologia (grants PTDC/EGE-ECO/119148/2010 and UTA_CMU/MAT/0006/2009) is also gratefully acknowledged.

References

Acharya, V.A., Bharath, S.T., Srinivasan, A., 2007. Does industry-wide distress affect defaulted firms? Evidence from creditor recoveries. *Journal of Financial Economics* 85, 787-821.

Basel Committee on Banking Supervision, 2006. International convergence of capital measurement and capital standards. Bank for International Settlements.

Bastos, J.A., 2010. Forecasting bank loans loss-given-default. *Journal of Banking & Finance* 34, 2510-2517.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. Wadworth International Group, Belmont, California.

Breiman, L., 1996. Bagging predictors. *Machine Learning* 24, 123-140.

Caselli, S., Gatti, S., Querci, F., 2008. The sensitivity of the loss given default rate to systematic risk: New empirical evidence on bank loans. *Journal of Financial Services Research* 34, 1-34.

Das, S.R., 2007. Basel II: Correlation Related Issues. *Journal of Financial Services Research* 32, 17-38.

Dermine, J., Neto de Carvalho, C., 2006. Bank loan losses-given-default: A case study. *Journal of Banking & Finance* 30, 1291-1243.

Emery, K., Cantor, R., Keisman, D., Ou, S., 2007. Moody's Ultimate Recovery Database. Moody's Investors Service.

Grunert, J., Weber, M., 2009. Recovery rates of commercial lending: Empirical evidence for German companies. *Journal of Banking & Finance* 33, 505-513.

Gupton, G.M., Stein, R.M., 2005. LossCalc V2: Dynamic prediction of LGD. Moody's Investors Service.

Loterman, G., Brown, I., Martens, D., Mues, C., Baesens, B., 2012. Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting* 28, 161-170.

Papke, L.E., Wooldridge, J.M., 1996. Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics* 11, 619-632.

Varma, P., Cantor, R., 2005. Determinants of recovery rates on defaulted bonds and loans for North American corporate issuers: 1983-2003. *Journal of Fixed Income* 14, 29-44.

Verikas, A., Kalsyte, Z., Bacauskiene, M., Gelzinis, A., 2010. Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey *Soft Computing* 14, 995-1010.

Witten, I.H., Frank, E., 2005. Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers.