Can corporate credit ratings be explained by simple decision heuristics?

Jürgen Bohrmann*
Gunter Löffler

University of Ulm

May 2011
preliminary

Abstract

We compare the empirical performance of models that predict corporate credit ratings with key financial indicators. In addition to commonly used models, including ordered probit regression, multivariate linear regression, and machine learning algorithms, we also use two simple heuristics. One of these heuristics, Categorization by Elimination (CBE), outperforms the competitors in terms of predictive ability. CBE is a fast and frugal heuristic suggested in the psychological literature as a realistic description of efficient human decision making. Our results are therefore consistent with the presence of subjective components in rating decisions. They also show that such components can be represented through a structured algorithm.

JEL classification: C35, C45, C52, D81, G24

Keywords: credit rating, heuristics, categorization by elimination, ordered probit, bounded rationality

Both authors are from the Institute of Finance, University of Ulm, Helmholtzstrasse 18, 89069 Ulm, Germany. This research was supported by Deutsche Forschungsgemeinschaft (DFG) through the SFB 649 “Economic Risk”.

* Corresponding author, juergen.bohrmann(at)uni-ulm.de

1. Introduction

Credit rating agencies such as Fitch, Moody's and Standard & Poor's have been issuing ratings for corporate debt since the first half of the 20th century. Today, the majority of large companies in developed markets possess a credit rating.[1] Over time, the rating process has not undergone major changes. Rating agencies combine quantitative information such as accounting ratios with qualitative assessments of management quality and other factors. The final rating decision is made by a rating committee. According to rating agencies, it does not rest on a fixed weighting algorithm (Standard and Poor's (2008)).

Recent events have spurred criticism of rating quality. Firms such as Lehman Brother's and AIG collapsed even though they were highly rated until the problems became evident. As part of their response, regulators have called for more transparency about rating decisions.[2] The presumption is that greater transparency will increase consistency in rating assignments and help market participants to better assess the information content of a rating. Larger transparency would be straightforward to achieve if rating agencies used fixed rules for weighting well-defined input variables. As soon as qualitative information and judgmental weightings come into play, however, increasing transparency may turn out to be difficult.

In this paper, we show that a better understanding of the rating process does indeed require a modelling of the subjectivity of the rating process. We propose that decision heuristics studied in the field of psychology (cf. Gigerenzer, Todd, and ABC Research Group (1999)) are a plausible model for the information aggregation performed at the end of the rating process. Specifically, we employ the Categorization by Elimination characteristic suggested by Berretty et al. (1997). The heuristic is hierarchical and noncompensatory. Available pieces of information, or *cues*, are used in a certain order;

[1] Over 6,000 corporate issuers held a long-term bond, corporate family or loan rating from Moody's in 2008, over 5,500 corporate issuers held a Standard & Poor's rating in 2010 and Fitch had over 1,700 rated corporate issuers globally in 2010.

[2] http://www.sec.gov/news/press/2011/2011-113.htm; http://ec.europa.eu/internal_market/consultations/docs/2010/cra/cpaper_en.pdf.

once a categorization decision is made, it cannot be changed by cues form a lower hierarchical level. In a regression by contrast, each variable affects the final outcome, even though its coefficient may be very small; variables that are relatively unimportant individually can push the prediction into a direction that is not consistent with the most important variable. The categorization heuristic does neither require a computer nor a pocket calculator. It is therefore not just an "as if" model of behavior but a plausible description of actual human decision-making.

For a large data set of bond ratings for US corporates, we use the heuristic to derive rating decisions from a set of predictors that include leverage, interest coverage, profitability and other commonly used variables. Compared to statistical approaches such as linear regression, ordered probit or neural neutworks, the heuristic leads to better out-of-sample predictions of actual rating decisions.

The results are consistent with the presence of subjective components in rating decisions. They also show that such components can be represented through a structured algorithm, which can help to clarify the rating process. Of course, the fact that the heuristic still leaves a large part of unexplained variation in rating decisions justifies some caution. Empirically, it could turn out to be difficult to attain much higher levels of transparency because the extent to which subjective components can be made transparent may be limited. On the other hand, the paper focuses on the information aggregation stage and derives the predictor variables from public sources. With proprietary information from rating agencies, a further improvement in the representation of rating decisions should be possible.

With the decision-making literature, our paper is the first to document for a broad dataset that decision heuristics examined in the psychological literature can be very valuable for explaining the behavior of important financial market players. Extant psychological research tests the empirical validity using small samples and decision tasks which are common in everyday life but of limited economic importance.

The academic literature on the determinants of credit ratings goes back to the 1960s. Horrigan (1966), Pogue et al. (1969) and West (1970) employed multivariate linear regression analysis. Pinches et al.

(1973) and other researchers used discriminant analysis, before Kaplan et al. (1979) proposed ordinal models instead of linear models in order to match the nonlinear ordinal character of credit ratings. In recent years, many authors favor ordered probit or logit regressions (Blume, Lim and MacKinlay (1998), Amato et al. (2004), Jorion et al. (2009) and Caporale et al. (2009)), but linear regression continues to be used (Kisgen (2008)). In addition, neural networks and support vector machines have been investigated (Huang et al. (2004) and Lee (2007)).

The empirical validity of simplified decision rules has already been advocated by Dawes (1979). The Categorization by Elimination of Beretty et al. (1997) was inspired by Gigerenzer and Goldstein (1996). Recent practical applications are presented in Goldstein and Gigerenzer (2009).

2. The rating process and heuristics

2.1 A description of the rating process

"Bear in mind, however, that a rating is, in the end, an opinion. The rating assignment is as much an art as it is a science."

Standard & Poor's, Corporate Ratings Criteria (2008, p. 3)

According to rating agencies, a corporate credit rating is a current opinion of an issuer's relative creditworthiness, i.e. an opinion on the issuer's willingness and capacity to repay its debt.[3] Briefly, the rating process can be described as follows. If a rating is to be assigned for the first time or an existing rating is to be reviewed, the rating agency designates a rating analyst to prepare the rating decision. Analysts usually focus on one or few different industries in order to build up specific expertise. The lead analyst is backed by other analysts and support team members. Together, they collect quantitative and

[3] The information give in this section largely builds on Standard and Poor's (2008).

qualitative information relevant for assessing the issuer's creditworthiness. Important sources of information are financial statements, the firm's own projections, meetings with the rated firm's managers, and an analysis of the firm's competitors and the market it operates in. The lead analyst then presents the results of the analysis to the decision-making body – the rating committee. The committee members discuss the results and decide on the rating. Before it is published, the issuer is allowed to respond. If the firm appeals and brings in new information, the committee reconvenes.

For several reasons, reproduction of a rating decision faces several changes. Some key indicators such as management quality or competitive position are based on a qualitative analysis, making them difficult to reproduce. Through their contacts with the firm's management, agencies may have gained information that is not publicly available.[4] Finally, rating agencies do not reveal in detail how the information is aggregated. Standard and Poor's (2008), for example, describe eleven key financial ratios used in their analysis of financial risk but do not describe how the ratios are weighted relative to each other. According to Standard and Poor's (2008), the reason for not providing information on aggregation schemes is not that it is proprietary. Rather they state that the agency does not employ such fixed schemes: "Note that we do not have any predetermined weights for these categories. The significance of specific factors varies from situation to situation" (Standard and Poor's (2008), p. 22).

Such a description is not necessarily inconsistent with the observation that rating agencies have developed statistical models that aggregate information in a well-defined way. Moody's, for example, has published both statistical default prediction models (Falkenstein et al. (2001)) as well as a model that maps financial ratios into ratings (Metz and Cantor (2006)). Developing such a model can make sense even if the rating agency does not incorporate it in its own rating process. For example, the model could be marketed to risk managers who need to assess the risk of unrated companies. Metz and Cantor

[4] Until 2010, rating agencies were exempt from US regulation FD, meaning that managers could communicate material non-public information to rating agencies. Because of the possibility of confidentiality arrangements, it is not obvious whether the regulatory change will substantially change information flows between issuers and rating agencies.

(2006) also hint at the possibility that the presented model can serve as "as an initial input in the rating process" but point out "that we can never construct a perfect map, since we simply cannot include all the factors which determine ratings" (Metz and Cantor (2006), p. 1).

Of course, one could argue that rating agencies may employ algorithms that they do not fully disclose. With full disclosure, competing agencies, new entrants but also investors and issuers could apply the algorithms, eroding the agency's business model. The analysis of the present paper helps to assess whether the agencies' description of their rating process is consistent with reality. If it is true that the rating process is not fully automated but rather contains subjective elements, one should be able to improve rating prediction through a realistic modeling of the rating analysts' decision process. To develop such a model, we turn to psychological research on heuristics.

2.2 Heuristics

The type of heuristics that we apply to credit rating agencies goes back to the research of cognitive psychologists Gigerenzer and Goldstein (1996). For binary choices of the type "In terms of population, is city A larger than city B?" they propose what they call the "Take the Best" rule. To answer the question, decision makers employing this rule would first consider the information – henceforth called cue – which they expect to have the largest validity, where validity is measured as the percentage of correct answers among the answers generated by a cue. Example cues for the city size problem could be: "Did the city host the olympic summer games at some point in the past?", "Is the city a national capital?", or "Does the city have an international airport?". Assuming that these cues were listed in the order of their validity and taking the problem to be "Is Madrid larger than Munich" one would examine each cue in the order of validity, and stop once a cue generates an answer. In the example given here, the first cue would not discriminate as both cities have an international airport. The decision maker would proceed to the second cue. Since Madrid is a national capital while Munich is not, the decision maker would then fix the answer, and discard any remaining cues.

The Take the Best rule is hierarchical (cues are examined in a pre-specified order) and noncompensatory (some pieces of information are neglected and cannot influence the decision). Heuristics such as Take the Best are often dubbed fast and frugal, as the less than complete use of available information increases the speed of decision making.

Take the Best and related heuristics have been inspired by Simon's (1956) theory of bounded rationality, which acknowledges that in many decision-making situations, it is either impossible or costly to acquire perfect information. Also, the cognitive capacity of decision makers may be limited. In such situations, following simplified decision rules can be advantageous because the perfect solution may be infeasible or too costly to obtain. Findings by Gigerenzer and Goldstein (1996, 1999) suggest that the Take the Best heuristic is ecologically valid. In many different situations, from ranking cities according to size to ranking professors according to their salary, they obtain that it performs at least as well as more elaborate methods such as linear regression.

Heuristics also feature prominently in another strand of psychological research, the heuristics and biases literature initiated by Tversky and Kahneman (1974). In this strand as well as in much of the behavioral finance literature that is built upon it, simplified decision rules are usually associated with biased or inaccurate decisions that suffer from violations of the laws of logic and probability. The latter suggests that predictive accuracy could easily be increased by avoiding the fallacies associated with the heuristic. Proponents of fast and frugal heuristics, by contrast, would argue that they enable efficient decision making that cannot improved upon easily (cf. Gigerenzer (2009)).

To model credit rating decisions, we cannot use the Take the Best heuristic as it is only applicable to binary decisions. A generalization to mult-category decision problems is the Categorization by Elimination (CBE) heuristic, which was introduced by Berretty et al. (1997). It bears resemblance to Tversky's (1972) Elimination by Aspects. CBE can be applied to both unordered and ordered categories. Berretty et al. (1997) give the illustrative example of categorizing birds (eagle, sparrow or stork?), while we apply it to ratings. Assume that there are seven rating categories and that there are a

number of cues, e.g. financial ratios such as leverage, return on assets, and so forth. In a first step, we need to select the cues that have the potential to enter the process, and sort them according to their empirical validity in predicting credit ratings. There are different ways for measuring validity. In the subsequent analysis, we suggest to use the Pseudo-R^2 of an ordered probit analysis of ratings in which the respective cue is the only explanatory variable. As a final preparatory step, we need to map cue values into rating categories. We associate each category with a range for the cue value. If the cue value is inside the range of category k, it is associated with rating category k. Ranges for different categories can overlap, meaning that one cue value can be associated with several ratings. Two standard procedures for define the ranges are (i) define the range through the maximum or minimum in the data used for training the algorithm (ii) use empirical confidence intervals. In the paper, we employ the first approach.

The category is best described with a simple example illustrated in Figure 1. Assume that we have four cues: leverage (highest validity), return on assets, operating margin, interest coverage (lowest validity). The assumed binning structure is shown in Figure 1. The highest cue, leverage, is associated with ratings AA and A. As this set is not unique and therefore does not lead to a rating decision, we proceed to the next cue, which is return on assets. Here, it is associated with ratings BBB, BB and B. According to CBE, the new set is the intersection of the previous set with the set of categories associated with the current cue. If the intersection is empty as is the case here, CBE calls to stay with the previous set and proceed to the next cue. This discarding of information is consistent with the hierarchical structure of the algorithm. In the example, we move on to operating margin, which is associated with ratings BB, BBB and A. The intersection with the previous set is {A}. The considered set is now unique, and the CBE algorithm stops. Hence, we would not consider information contained in the fourth cue, which exemplifies the noncompensatory character of the algorithm.

In practical applications, we may not arrive at a unique set of categories even after all available cues have been considered. In such a case, Berretty et al. (1997) suggest to take the average of the categories that remain at the end. Alternatively, one could randomly select a category from the remaining ones.

To complement the exemplary description of the algorithm, Figure 2 shows a flow diagram. Let us briefly summarize the algorithm in a more formal language. At the start, a set of cues with their binning structure have to be given. Then as a first step, the cue with the highest validity is taken and we compare the value of the first cue to our binning structure for the first cue. We eliminate all possible rating categories that are inconsistent with the value of the first cue, yielding a set S of possible rating categories S. If only one rating category remains in this set S, the algorithm already stops here. If not, we proceed to the cue C* with the next highest validity and once again we check which rating categories are consistent with the value of cue C*. We denote the possible rating categories for cue C* by S*. Now we intersect the set S with the S* to see whether further rating categories can be eliminated. We denote the intersection of S and S* by S again. If S contains only one rating category, we are finished and choose this category. If S contains more than one category and if there are still further cues that can be checked according to their order of validity, we proceed as before. If the intersection S is empty, we reset S to the previous value, and proceed to the next cue. In the case that no additional cues are available and more than one category remains in the set S, we choose the average rating of the categories in S.

Let us briefly ponder whether the CBE algorithm could be a plausible description of actual rating behavior. Binning financial indicators into categories is similar to the representation given in Standard and Poor's (2008, Table 2), There, financial ratios are binned into five categories of financial risk. The financial risk categorization is then brought together with the business risk assessment, resulting in a rating. Technically, binning ranges can be determined based on empirical ranges. The necessary information is available to rating agencies. Standard and Poor's (2007), for example, contains median values per rating category for several financial ratios. However, CBE could be a realistic description of

the decision process even if the decision makers did not use well defined binning structure. Based on their experience, member of the rating committee could follow a mental mapping. Finally, note that the application is not computationally expensive. It could be applied by individual analysts during the rating discussion in the committee.

3. Data and methodology

3.1 Cue selection and variable definition

We study long-term issuer ratings from Standard & Poor's for US corporates. Ratings as well as annual financial information are obtained from COMPUSTAT for the years 1985 to 2007. Information about stock returns is from CRSP. When we link ratings and stock market information to accounting data, we use a conservative lag of six months to ensure data availability. For a firm with fiscal-year end December, for example, annual financial data for 2005 are coupled with the rating from June 2006.

In accordance with the literature (e.g. Blume, Lim and MacKinlay (1998)) all accounting variables that we use as cues for rating prediction are averaged over the last three years in order to approximate the through-the-cycle approach of rating agencies. Looking through the cycle means that rating agencies abstract from fluctuations they regard as temporary (Standard and Poor's (2008), p. 22); since they do not provide information on how they separate temporary and long-term components, a simple, backward-looking average is used here.

If there is only a two-year history for an observation, we use the two-year average instead of the three-year average so that we do not lose too many observations. If there is only information about one year, the observation is discarded. All variables are winsorized at the 1% and the 99% levels to alleviate potential outlier problems, which is again common practice in the literature.

We use two different sets of cues for our rating prediction models. The first set takes the variables used in the influential study of Blume, Lim and MacKinlay (1998) – hereafter: BLM. The predictor variables

of BLM have been adopted by other studies (Amato et al. (2004) and Jorion et al. (2009). They include: pretax interest coverage; operating income to sales; long-term debt to assets; total debt to assets; market capitalization; equity return beta; equity return standard error. Precise variable definitions are given in Table 1. Note that market capitalization as well as betas and standard errors are not averaged. As in Amato et al. (2004) and Jorion et al. (2009), betas and standard errors are cross-sectionally demeaned. The variables capture business risk (market capitalization, beta, standard error), financial leverage (long-term debt to assets, total debt to assets), profitability (operating income to sales), and interest coverage.

The Table also reports the Pseudo-R^2 of ordered probit regressions, in which the S&P rating is explained only with the variable in question. These Pseudo-R^2 values measure the strength of the bivariate relationship between ratings and a given variable; they will be used to measure the empirical validity of a cue in the Categorization by Elimination algorithm.

For the second set of cues, shown in Table 2, we have considered a larger set of 31 candidates. In addition to the BLM variables, we consider ratios mentioned in publications of rating agencies (Standard and Poor's (2008) and Metz and Cantor (2006)). Conceptually, the differences between variable definitions are often small. In their definition of interest coverage, for example, Metz and Cantor (2006) propose to include rental payments and preferred dividends, which are not included by BLM (1998). Due to the large similarity among subsets of these 31 candidates, we restrict the further analysis eight variables. Within each of four groups (business risk, leverage, profitability, coverage), we choose two variables that have the strongest bivariate relationship with ratings, again measured through separate ordered probit regressions.

Our merged COMPUSTAT-CRSP initial data set contains 29,005 firm-years with S&P long-term issuer ratings ranging from 1985 to 2007. Deleting all observations with missing data on any of our seven BLM cues listed in Table 1, we end up with a sample size of 21,235 firm-years for the data set with the BLM variables. Deleting all observations with missing data on any of the eight cues, we arrive at a

sample size of 10,301 firm-years. The large reduction in firm-years that comes about when moving from the BLM variables to the expanded set of predictors is one reason why we conduct the analysis separately for the two sets of variables. Descriptive statistics of the data sets are shown in Table 3.

3.2 Models for rating prediction

Categorization by Elimination (CBE)

This paper is the first to apply this heuristic to the prediction of corporate credit ratings. The structure of the algorithm was already explained in section 2.2. Here, we only briefly summarize the main assumptions made for implementation. The set of cues considered is given either by the BLM variables from Table 1 or the expanded set of predictors from Table 2. For each variable j and each letter rating category k, a range is defined using the minimum and maximum values of variable j that are observed for rating category k in the estimation sample. The validity of cues, which determines the order in which they are checked, is measured with the unconditional Pseudo-R^2 from Tables 1 and 2.

Tallying

Weighting each cue equally is referred to as tallying. Such unit-weight linear models sometimes perform as well as multivariate linear regression in terms of predictive accuracy. Dawes (1979) and Dawes et al. (1974) first concluded that tallying is a serious competitor for linear regression; Czerlinski et al. (1999) reaffirm their findings on 20 different and independent data sets. Averaged over all data sets, multivariate regression achieved a higher accuracy for the in-sample fit, but tallying slightly outperformed linear regression in out-of-sample predictions. Einhorn et al. (1975), however, points out that tallying can achieve prediction accuracy comparable to that of multivariate linear regression only under special circumstances.

Compared to CBE, tallying is not a frugal heuristic because it makes use of all available cues. The decision-making process cannot be stopped after the first discriminating cue, but all information at hand is used. In our empirical study, we make a modification to the original tallying rule. We do not add up the individual decision cues. Instead we add up the individual cue-implied ratings with equal weights and then take their average. Implied ratings are determined through ordered probit regressions.

Multivariate linear regression

In contrast to tallying, multivariate linear regression differentiates individual decision cues according to their importance by estimating a separate coefficient for each of them. In line with the literature (Horrigan (1966), Pogue et al. (1969), West (1970), and Kisgen (2008), we code the rating information in a linear fashion (AAA = 1, … , CCC/CC/C = 7).

Ordered probit regression

The ordered probit regression has one main advantage over linear regression, which makes it so far the state-of-the-art in the empirical literature on rating prediction (cf. Kamstra et al. (2001)). Ordered probit does not rely on the assumption of a linear rating scale. Rather, it endogenously determines cutoff points for each rating category. These cutoff points are monotonic in the score, but they need not to be equidistant.

More formally, the ordered probit approach can be described as follows. Let R_{it} be the rating of issuer *i* at time *t* and X_{it} a vector of our decision cue variables available at time *t* that impact on issuer *i*'s rating. R_{it} is defined as above with AAA corresponding to 1, AA to 2, …, and all CCC/CC/C ratings corresponding to 7. Now let us introduce a latent variable Z_{it} that maps values of X_{it} in R_{it}. The linear equation $Z_{it} = \beta X_{it} + \varepsilon_{it}$ with β as a vector of coefficients and ε_{it} as an error term links our

observable decision cues X_{it} to the latent variable Z_{it}. Furthermore, Z_{it} is linked to our numerical rating variable R_{it} according to:

$$R_{it} = \begin{cases} 1 & if\ Z_{it} \in (-\infty, \mu_1) \\ r & if\ Z_{it} \in [\mu_{r-1}, \mu_r), \quad r = 2, 3, \ldots, 6 \\ 7 & if\ Z_{it} \in [\mu_6, \infty) \end{cases}$$

where the μ_i represent our cutoff points for the rating categories. We use the same link and the estimated cutoff points to associate a prediction $\hat{\beta} X$ with a rating.

Linear discriminant analysis

Discriminant analysis maximizes the variance between classification groups and minimizes the variance within classification groups. Pinches et al. (1973) and Kumar et al. (2006) used discriminant analysis in order to forecast ratings.

C4.5 – a decision tree algorithm

As one of two machine learning algorithms, we also include a decision tree algorithm in the list of competitors. C4.5 is a decision tree algorithm developed by Quinlan (1993). We have chosen C4.5 because of its easily accessible free-source implementation in the WEKA data mining environment[5] and because of the ostensible similarities of the decision tree and the CBE approach. Furthermore, C4.5 ranks among the most powerful statistical classifiers that can deal with continuous cue values (cf. Martinelli et al. (1999)). Based on a learning sample of preclassified instances grouped together with their cue values, the decision tree algorithm recursively builds up the decision tree. At each node of the tree, the algorithm chooses one cue that – in terms of information entropy – most effectively splits the

[5] The name of WEKA's Java implementation of C4.5 is J48.

set of issuers to be classified into subsets. After splitting up the subsets according to a node-specific threshold value for the most effective cue, the C4.5 algorithm then recursively gets deeper into these subsets until subsets only contain issuers with the same rating category.

C4.5 shows some similarities to the CBE approach. The decision tree structure is common to both approaches, also a stop of classification after the first decision node is possible for C4.5. Despite these similarities, one should not forget that C4.5 is a model of unbounded rationality. The decision tree structure can become overwhelmingly complex for C4.5, whereas the maximum number of decision nodes used in CBE is given by the number of cues. C4.5 might create several decision nodes for the same cue, each with differing threshold values. Furthermore, the creation of the C4.5 decision tree is a recursive procedure, where the algorithm makes a full recursive optimization at each node. Thus, C4.5 does not suggest itself as a realistic model of human decision making.

Neural network

Our second applied machine learning algorithm is the WEKA implementation of a three-layer neural network with backpropagation learning. We have experimented with different learning rates, momentum and number of hidden layers in order to optimize the out-of-sample prediction performance. This should allow the neural network to come as close as possible to the ideal of an unboundedly rational model. Finally, we have settled on a learning rate of 0.3, a momentum of 0.2 and a training time of 10,000 seconds.

Huang et al. (2004) give a comprehensive overview on applying backpropagation neural networks to the prediction of corporate credit ratings. The studies performed in the literature so far vary greatly in the number of rating categories (from 2 to 16), the number of cues used (from 4 to 87), the composition of cues, the proportion of training set size to validation set size, and the overall size of the rating data set (from 47 to 3,886 observations).

4. Empirical results on predictive accuracy

We compare the rating prediction accuracy of the Categorization by Elimination algorithm with the competitor models tallying, multivariate linear regression, ordered probit, linear discriminant analysis, C4.5 and the neural network. We compare the models' performances for different sets of decision cues, different sample sizes, and both for in-sample data fitting as well as for out-of-sample prediction.

The primary accuracy measure that we use is the hit rate. The hit rate is defined as the percentage of correctly predicted observations. Again, ratings are coded from AAA=1 to C=7; predictions are rounded to the nearest integer. If the actual rating of the firm-year observation was BBB or 4, and our rating prediction model made a prediction of 4.4, then we would count it as a hit. If the prediction was 4.6, it would not be counted as a hit. The number of hits divided by the number of predictions made is the hit rate.

4.1 Analysis with seven cues from Blume, Lim and MacKinlay (1998)

In this section, we use the variables suggested by BLM (1998) as rating predictors. For CBE and tallying we use the seven BLM cues as defined in Table 1. This also holds for the other models, except for pretax interest coverage, which is transformed into a piecewise linear function via four transforms C1, C2, C3 and C4 to account for skewness of interest coverage.[6] Using the pretax interest coverage transforms for the heuristics CBE and tallying does not seem to be intuitive. Viewed in isolation, the variables used to implement the transformation do not lend themselves easily to a psychologically

[6] The transformation is proposed by BLM (1998), and is implemented as follows: first set observations with negative numerator to zero, then those with negative denominator to 100, and cap all other observations at 100. Call this modified coverage *c*. Then define four variables: C1 = min(*c*, 5); C2 = max(0, min(*c*,10) - 5); C3 = max(0, min(*c*,20) - 10);C4 = max(0, *c* - 20).

plausible processing On the other hand, unreported analysis shows that using only the original pretax interest coverage cue without transformations for the other prediction models would decrease the prediction accuracy of linear discriminant analysis, linear regression, ordered probit, C4.5 and the neural network.

In Table 4 we validate the prediction models in-sample. Here, training and validation data sets are identical. To examine stability over time, we present results for expanding estimation windows. The first rating prediction is performed on all observations from 1985 to 1995, i.e. the prediction model is trained and validated on the years 1985 to 1995. The second training and validation is performed on the period from 1985 to 1996, the third from 1985 to 1997 and so forth. In Table 5 we perform out-of-sample rating predictions using the same expanding estimation windows as in the in-sample analysis. For the out-of-sample validation, we use the years 1985 to 1995 as our first training data set and then predict the ratings for the subsequent year 1996. The last training set in our out-of-sample exercise is 1985 to 2006, on which a prediction for 2007 is based.

Multivariate linear regression and ordered probit regression estimate their coefficients and cutoff points with the training data set. With these estimated coefficients and cutoff points at hand, the cue values of the out-of-sample period are transformed in a rating. C4.5 builds the full decision tree with all nodes and threshold values based on the training set and then makes predictions with this tree structure for the validation set. The neural network learns its weights for all neurons in all layers and then uses this neuronal structure to make predictions on the validation set. Both CBE and tallying build their binning structure with the held of the training data and then use this binning structure for making classifications out-of-sample.

In the in-sample analysis of Table 4, C4.5 achieves the highest average in-sample prediction accuracy. It is followed by CBE, the neural network, linear regression, ordered probit, linear discriminant analysis and finally tallying. When looking at the out-of-sample performance presented in Table 5, CBE leads the board followed by the neural network, C4.5, ordered probit, linear regression, linear discriminant

analysis and once again tallying at the end. Differences appear significant. The average hit rate of CBE is 59.9%, compared to 51.8% for the neural network. The dominance of CBE is also stable over time. With the exception of 2001, CBE achieves the highest hit rate among the competing prediction models.

The relative ranking of prediction models other than CBE is consistent with the academic literature. Kim (1993), Maher et al. (1997) and Kumar et al. (2006) have demonstrated on their data that neural networks achieve higher prediction accuracy out-of-sample than linear models such as linear regression or linear discriminant analysis. Kaplan et al. (1979) have stressed that ordinal models such as ordered probit or logit are more appropriate than linear models for modeling corporate credit ratings. Chaveesuk et al. (1999) and Huang et al. (2004) have shown that for some specifications neural networks can outperform ordered logit regression in terms of prediction accuracy. Outside the rating prediction literature, Berretty et al. (1997) have reported that CBE can come close to the performance of neural networks in categorizing mushrooms, iris flowers or wine. Furthermore, Brighton (2006) has shown that simple heuristics can beat both C4.5 and neural networks on the majority of small exemplary data sets in his study.

Following the heuristics literature (cf. Berretty et al. (1997), Czerlinski et al. (1999) or Brighton (2006)), we used the hit rate – the proportion of correct inferences made – as our primary prediction accuracy measure. As alternative accuracy measures, we now consider in Table 6 the distribution of rating prediction errors and the mean absolute prediction errors.[7]

The mean absolute error is highest for tallying, both in- and out-of-sample, while C4.5 leads to the lowest in-sample error. For out-of-sample predictions, CBE beats almost all other competitor models with respect to hit rates and is ranked second with respect to mean absolute prediction errors. Only the neural network shows a lower mean absolute prediction error than CBE.

[7] Metz and Cantor (2006, p. 3) advise against the least-squares criterion as a valid accuracy measure for rating prediction models: "A least-squares criteria would prefer a model which had 18 issuers with one notch errors and one issuer with a nine notch error (total squared errors being 99) to a model which had 18 issuers with a zero notch error and one issuer with a ten notch error (total squared errors being 100). But users of the model would almost certainly prefer the latter, since, for all intents and purposes, a nine notch error is every bit as bad as a ten notch error, but a zero notch error is much better than a one notch error."

Looking at the cumulative shares of rating predictions that deviate from the actual rating by a certain number of rating categories, we can explain the small discrepancy between the hit rate results and the mean absolute error results. CBE achieves a higher hit rate than the other models at the cost of a higher dispersion of rating prediction errors around the actual rating. The distribution of rating errors is wider for CBE than for the ordered probit and also the two machine learning models.

4.2 Sensitivity analysis: Different cues, different sample sizes

In this section, we use a simulation study to examine whether the results of the previous section are robust to the choice of predictor variables, and whether sample size matters for relative accuracy. We use the data set with 10,301 firm-year observations that remain if the eight predictor variables of Table 2 are considered. We train our models and predict ratings with either four cues – standard error of the market model, long-term debt to assets, Moody's ROA and Moody's definition of interest coverage – or with eight cues – the aforementioned four cues and in addition market capitalization, Moody's debt to assets, S&P's return on capital and S&P's FFO interest coverage. We have chosen the variables such that each of the categories business risk, leverage, profitability and interest coverage is represented by one variable (if four cues are considered) or two variables (if eight cues are considered). Within each category, the cue with the highest bivariate Pseudo-R^2 was chosen. As before, all cues are winsorized at the 1% and 99% levels; accounting variables are averaged over three years.

In-sample and out-of-sample analysis are conducted as in the previous section, using expanding estimation windows from 1985-1995, 1985-1996 and so forth. Table 7 shows the average hit rates, averaged over all estimation windows. Relative accuracy is very similar to the previous section. In-sample, the CBE algorithm ranks second behind C4.5; out-of-sample, CBE again outperforms the other competitors, including C4.5. The results are therefore robust to changes in the definition and number of cues used for prediction.

It is interesting to see that for almost all prediction models prediction accuracy decreases comparing in-sample to out-of-sample results. This holds true for analyses with both four and eight cues. Only the heuristics models CBE and tallying are able to increase their prediction accuracy when switching from in-sample to out-of-sample validation. One likely explanation is that outliers in the cue variables can have a large effect on the CBE algorithm because they can widen the binning ranges. If a subset of the data is used for estimating the ratings, the impact of outliers will be smaller or equal compared to the use of the complete data.

Next, we use different sample sizes in order to find out whether there is a relationship between sample size and prediction accuracy of these models. We use sample sizes of 100, 1,000 and 10,000 firm-year observations. The subsamples of 100, 1,000 and 10,000 observations are randomly drawn without replacement across issuers and years from the full set of 10,301 observations. In order to get stable results for prediction accuracy in the smaller sample sizes, we draw 1,000 independent random sets of 100 observations, 1,000 independent random sets of 1,000 observations and 1,000 independent random sets of 10,000 observations. After running the prediction models separately on the generated random sets, we average our prediction accuracy measure. This method is often referred to as repeated random sub-sampling validation.

For the out-of-sample analysis, the randomly drawn samples are split into two halves. The first half is used as training set, while the second half serves as the validation data set for prediction. We refer to this procedure as cross-validation.

The results of our simulation study are shown in Table 8. For in-sample validation, once again C4.5 shows the highest prediction accuracy, whereas in the out-of-sample validation CBE outperforms all other rating prediction models for all different sample sizes. When looking at Table 8 in more detail, some peculiarities are striking. In-sample, the prediction accuracy of most models decreases with increasing sample size. The opposite appears for out-of-sample predictions. How can these patterns be explained? In the in-sample analysis, the models are trained and validated on the same observations.

The smaller the sample size, the larger is the number of parameters relative to the number of observations, making it easier to fit the data. This remark also holds for CBE, which requires the estimation of binning ranges. In the out-of-sample exercise, learning and validation data sets are not identical. When the model learns on a larger sample, it becomes more likely that the model finds general relationships instead of peculiarities that are not representative for the data.

4.3 Determinants of CBE's performance

To learn more about the determinants of CBE's success, we analyze the number of cues used in the course of the decision process, and split the CBE hit rate into two components – the hit rate of unique decisions and the hit rate of averaging decisions. We again use the simulation framework of the previous section to differentiate according to the sample size and the number of cues used. Results are presented in Table 9.

The average number of cues considered is very high. For large samples, it is close to 100%. The percentage of cases in which the algorithm stops with a unique rating – what we call unique decisions – is relatively small, at least if the sample size is 1,000 or larger. The hit rate for these unique decisions is considerably smaller than the hit rate of the averaging decisions. The two observations clarify the empirical character of CBE when applied to rating decisions. Though the algorithm is hierarchical and possibly discards information, most cues are considered in the application of this paper. Also, the decisions rarely lead to a clear-cut rating decision.

5. Conclusion

We have performed a comparative study of rating prediction models including ordered probit regression, multivariate linear regression, linear discriminant analysis, sophisticated machine learning algorithms as well as two simple decision-making heuristics. One of these heuristics, Categorization by

Elimination (CBE), has achieved a relatively high prediction accuracy. In out-of-sample predictions, it outperforms the complete set of competitors in terms of the hit rate. The result is robust to the number and the definition of decision cues.

CBE is a fast and frugal heuristic suggested in the psychological literature as a realistic description of human decision making. Our results are therefore consistent with the presence of subjective components in rating decisions. They also show that such components can be represented through a structured algorithm. While this could help to increase transparency about the rating process, the fact that the heuristic still leaves a large part of unexplained variation in rating decisions justifies some caution. Empirically, it could turn out to be difficult to attain much higher levels of transparency because the extent to which subjective components can be made transparent may be limited.

Though the paper cannot establish that rating analysts actually follow the heuristic, it is interesting to note that it would lend itself easily to an application in a rating committee. The algorithm is hierarchical, meaning that the order of discussion is pre-specified as long as committee members agree on the ordering of cues; in addition, it does not require a re-assessment of prior evaluations. It is noncompensatory, meaning that some cues are either not considered at all or dismissed in between. It does not require computations. In consequence, CBE would be an efficient way of structuring the decision process.

References

Amato, J. and C. Furfine (2004). Are credit ratings procyclical? *Journal of Banking and Finance*, 28, pp. 2641-2677.

Bank of International Settlement (2010). BIS Quarterly Review – Statistical Annex. *http://www.bis.org/publ/qtrpdf/r_qa1003.pdf#page=114, accessed on 20th May 2010.*

Berretty, P. M., Todd, P. M. and P. W. Blythe (1997). Categorization by elimination: A fast and frugal approach to categorization. *In M. G. Shafto and P. Langley (Eds.), Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, Mahwah, pp. 43-48.

Blume, M. E., Lim, F. and C. MacKinlay (1998). The declining credit quality of U.S. corporate debt: myth or reality? *Journal of Finance*, 53, pp. 1389-1413.

Brighton, H. (2006). Robust inference with simple cognitive models. *In Lebiere, C. and R. Wray (Eds.), Between a rock and a hard place: Cognitive science principles meet AI-hard problems. Papers from the AAAI Spring Symposium*, *AAAI Press*, Menlo Park, pp. 17-23.

Caporale, G., Matousek, R. and C. Stewart (2009). Rating assignments: Lessons from international banks. *CESIFO working paper No. 2618.*

Casey, C. J. (1980). Variation in accounting information load: The effect on loan officers' predictions of bankruptcy. *Accounting Review*, 55, pp. 36-49.

Chaveesuk, R., Srivaree-Ratana, C. and A. E. Smith (1999). Alternative neural network approaches to corporate bond rating. *Journal of Engineering Valuation and Cost Analysis*, 2, pp. 117-131.

Chewning, E. G. and A. M. Harrell (1990). The effect of information load on decision makers' cue utilization levels and decision quality in a financial distress decision task. *Accounting, Organizations and Society*, 15, pp. 572-542.

Czerlinski, J., Gigerenzer, G. and D. G. Goldstein (1999). How good are simple heuristics? *In Gigerenzer, G., Todd, P. M. and the ABC Research Group (Eds.), Simple heuristics that make us smart*, *Oxford University Press*, 1st edition, New York, pp. 97-118.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34, pp. 571-582.

Dawes, R. M. and Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81, pp. 95-106.

Dimson, E. (1979). Risk measurement when shares are subject to infrequent trading. *Journal of Financial Economics*, 7, pp. 197-226.

Einhorn, H. J. and R. M. Hogarth (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, 13, pp. 171-192.

Falkenstein, E. G., Ibarra, E., Kocagil, A. E. and J. Sobehart (2001). RiskCalc Public – Europe: Rating methodology. *Moody's Investors Service*.

Gigerenzer, G. and H. Brighton (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1, pp. 107-143.

Gigerenzer, G. and D. G. Goldstein (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, pp. 650-669.

Gigerenzer, G., Todd, P. M. and the ABC Research Group (1999.). *Simple heuristics that make us smart*, Oxford University Press, 1st edition, New York.

Goldstein, D. G. and G. Gigerenzer (1999). The recognition heuristic: How ignorance makes us smart. *In Gigerenzer, G., Todd, P. M. and the ABC Research Group (Eds.), Simple heuristics that make us smart*, *Oxford University Press*, 1st edition, New York, pp. 37-72.

Horrigan, J. O. (1966). The determination of long term credit standing with financial ratios. *Journal of Accounting Research, Supplement*, pp. 44-62.

Huang, Z., Chen H., Hsu C., Chen W. and S. Wu (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, 37, pp. 543-558.

Jorion, P., Shi, C. and S. Zhang (2009). Tightening credit standards: The role of accounting quality. *Review of Accounting Studies*, 14, pp. 123-160.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47 (2), pp. 263-291.

Kamstra, M., Kennedy, P. and T. Suan (2001). Combining bond rating forecasts using logit. *Financial Review*, 37, pp. 75-96.

Kaplan, R. and G. Urwitz (1979). Statistical models of bond ratings: A methodological inquiry. *Journal of Business*, 52, pp. 231-261.

Kim, J. W. (1993). Expert systems for bond rating: A comparative analysis of statistical, rule-based and neural network systems. *Expert Systems*, 10, pp. 167-171.

Kisgen, D. (2006). Credit ratings and capital structure. *Journal of Finance*, 61, pp. 1035-1072.

Kumar, K. and S. Bhattacharya (2006). Artificial neural network vs. linear discriminant analysis in credit ratings forecast: A comparative study of prediction performances. *Review of Accounting and Finance*, 5, pp. 216-227.

Lee, Y.-C. (2007). Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications*, 33, pp. 67-74.

Maher, J. J. and T. K. Sen (1997). Predicting bond ratings using neural networks: A comparison with logistic regression. *Intelligent Systems in Accounting, Finance and Management*, 6, pp. 59-72.

Martinelli, E., de Carvalho, A., Rezende, S. and A. Matias (1999). Rules extractions from banks' bankruptcy data using connectionist and symbolic learning algorithms. *In: Proceedings of Computational Finance Conference, January 1999*, New York.

Metz, A. and R. Cantor (2006). Moody's credit rating prediction model. *Moody's Investors Service*, Special Comment.

Pinches, G. E. and K. A. Mingo (1973). A multivariate analysis of industrial bond ratings. *Journal of Finance*, 3, pp. 1-18.

Pogue, T. F. and R. M. Soldofsky (1969). What's in a bond rating? *Journal of Financial and Quantitative Analysis*, 4, pp. 201-228.

Quinlan, J. R. (1993). Programs for machine learning. *Morgan Kaufmann Publishers*, 1st edition, San Mateo, pp. 5-13.

Schaeffer, J., Burch, N., Björnsson, Y., Kishimoto, A., Müller, M., Lake, R., Lu, P. and S. Sutphen (2007). Checkers is solved. *Science*, 317, pp. 1518-1522.

Schroder, H. M., Driver, M. J. and S. Streufert (1967). Human information processing. Holt, Rinehart and Winston, 1st edition, New York.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63, pp. 129-139.

Standard & Poor's (2002). Corporate ratings criteria. *www.corporatecriteria.standardandpoors.com*.

Standard & Poor's (2007). 2007 adjusted key US industrial and utility financial ratios. *www2. standardandpoors.com/spf/pdf/fixedincome/CreditStats_2007_Adjusted_Key_Financial_Ratios.pdf.*

Standard & Poor's (2008). Corporate ratings criteria. *www.corporatecriteria.standardandpoors.com*.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, 79, pp. 281-299.

Tversky, A. and D. Kahneman (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185, pp. 1124-1131.

West, R. R. (1970). An alternative approach to predicting corporate bond ratings. *Journal of Accounting Research*, 8, pp. 118-125.

Table 1: Cues for BLM (Blume, Lim, MacKinlay) rating prediction study

The table presents the seven cues that Blume, Lim and MacKinlay (1998) applied in their ordered probit regression to model ratings. The first four financial accounting ratios – pretax interest coverage, operating income to sales, long-term debt to assets and total debt to assets – are averaged over three years in order to account for business cycle effects. Market capitalization, beta and standard error have not been averaged. All cues in the table are winsorized at the 1%- and 99%-level. The data is from the COMPUSTAT database and COMPUSTAT was merged with CRSP for betas and standard errors. In the rightmost column, the Pseudo-R^2 of single regressions of the S&P long-term issuer rating on the respective cue is reported.

Cues (explanatory variables)	Definition of variable	Pseudo-R^2 in bivariate ordered probit regression with dependent variable 'rating'
Pretax interest coverage	(Operating income after depreciation + interest expense) / interest expense	0.6%
Operating income to sales	EBITDA / sales	0.7%
Long-term debt to assets	Long-term debt / total assets	8.6%
Total debt to assets	(Long-term debt + debt in current liabilities) / total assets	6.3%
Market capitalization	Natural logarithm of real market capitalization	12.6%
Beta	Beta of market model estimated with 200 daily returns using the Dimson procedure with one lead and lag of the value-weighted market return	0.6%
Standard error	Standard error of market model estimated with 200 daily returns using the Dimson procedure with one lead and lag of the value-weighted market return	20.4%

Table 2: Selected cues for rating prediction simulation study

The table presents the eight cues that we have selected from Blume, Lim and MacKinlay (1998), Standard and Poor's (2006), Standard and Poor's (2008) and Metz and Cantor (2006).
The list of potential cues selected from the rating literature comprises a total of 31 variables. We order the cues according to the Pseudo-R^2 in ordered probit regressions of the S&P long-term issuer rating on the respective cue. For each of the categories – business risk, leverage, profitability, interest coverage – the two cues with the highest Pseudo-R^2 are reported here and are used in our simulation studies.

Select cues	Definition of variable; source	Pseudo-R^2 in bivariate ordered probit regression with dependent variable 'rating'
Standard error	Standard error of market model estimated with 200 daily returns using the Dimson procedure with one lead and lag of the value-weighted market return; *BLM (1998) – proxy for business risk*	20.4%
Market capitalization	Natural logarithm of real market capitalization; *BLM (1998) – proxy for business risk*	12.6%
Long-term debt to assets	Long-term debt / total assets; *BLM (1998) – leverage*	8.6%
Moody's debt to assets	(Long-term debt + debt in current liabilities) / total assets; *Metz and Cantor (2006) – leverage*	6.3%
Moody's ROA	Income before extraordinary items / two-year average of total assets; *Metz and Cantor (2006) – profitability*	6.0%
S&P's return on capital	Operating income after depreciation / two-year average of (long-term debt + debt in current liabilities + deferred taxes + stockholder's equity + minority interest); *S&P (2006) – profitability*	4.2%
Moody's interest coverage	(EBIT – interest capitalized + (1/3)*rental expense) / (interest expense + (1/3)*rental expense + preferred dividends / 0.65); *Metz and Cantor (2006) – interest coverage*	3.6%
S&P's FFO interest coverage	Total funds from operations / interest expense; *S&P (2008) – interest coverage*	0.5%

Table 3: Descriptive statistics

The table presents descriptive statistics for the initial data set with 29,005 firm-year observations between 1985 and 2007, a data set with 21,235 firm-year observations for which the Blume, Lim, MacKinlay (BLM, 1998) variables are available, and a data set with 10,301 firm-year observations for which a set of eight variables is available. Except for the number of observations, we report the means of the respective variables per rating category.

Variables	Rating 1 AAA	Rating 2 AA	Rating 3 A	Rating 4 BBB	Rating 5 BB	Rating 6 B	Rating 7 CCC-C	Firm-year observations in data set
Number of observations	412 1.4%	2,101 7.2%	6,841 23.6%	8,107 28.0%	6,251 21.5%	4,713 16.3%	580 2.0%	*Initial data set* 29,005 100.0%
Long-term debt to assets	12.4%	14.2%	18.8%	26.2%	35.7%	45.0%	51.2%	28.9%
Interest coverage (Moody's)	13.1	8.9	7.0	4.4	3.4	0.9	-0.8	4.3
Operating income to sales	30.0%	23.5%	23.5%	19.5%	12.0%	-17.0%	-347%	6.0%
Total assets ($ billion)	79.3	60.4	28.1	9.0	3.1	1.8	1.7	15.7
Number of observations	309 1.5%	1,661 7.8%	5,428 25.6%	5,864 27.6%	4,486 21.1%	3,113 14.7%	374 1.7%	*BLM variables* 21,235 100.0%
Long-term debt to assets	12.2%	14.6%	18.8%	25.3%	35.2%	43.3%	49.2%	28.4%
Interest coverage (Moody's)	14.2	8.3	6.4	4.3	3.0	1.0	-0.6	4.2
Operating income to sales	30.2%	23.3%	22.9%	18.5%	9.6%	1.0%	-103%	14.9%
Total assets ($ billion)	88.4	58.3	27.1	8.9	3.1	1.9	1.7	16.2
Number of observations	125 1.2%	681 6.6%	2,205 21.4%	2,679 26.0%	2,514 24.4%	1,858 18.1%	239 2.3%	*Eight variables* 10,301 100.0%
Long-term debt to assets	9.7%	16.7%	20.9%	27.3%	36.4%	46.3%	53.4%	31.3%
Interest coverage (Moody's)	14.5	8.4	6.2	3.9	2.9	1.1	-0.4	4.0
Operating income to sales	21.3%	22.8%	21.4%	18.8%	3.2%	9.3%	-358%	5,5%
Total assets ($ billion)	41.9	28.2	15.8	7.7	2.8	1.5	1.4	8.7

Table 4: In-sample hit rates of rating prediction models using seven variables from Blume, Lim and MacKinlay (1999)

The table contains the hit rates – i.e. the percentage of correctly predicted observations – for the rating prediction models linear discriminant analysis (LDA), linear regression, ordered probit regression, tallying, Categorization by Elimination (CBE), the decision-tree algorithm C4.5 and a multilayer perceptron neural network.
The cues used for rating prediction are those used by BLM (1998). The cues are winsorized at the 1%- and 99%-quantiles. For CBE, they are ordered according to their unconditional Pseudo-R^2 in an ordered probit regression of ratings. Training and validation data coincide.

Training & validation data sets		**Hit rates of the competitor models**						
Training & validation	# obs. validation	LDA	Linear regression	Ordered probit	Tallying	CBE	C4.5	Neural network
1985-1995	8,170	46.7%	45.9%	49.1%	25.1%	62.9%	85.1%	56.3%
1985-1996	9,178	46.4%	45.8%	48.7%	25.6%	63.2%	85.4%	56.5%
1985-1997	10,262	46.1%	45.9%	48.9%	25.9%	62.7%	84.6%	56.7%
1985-1998	11,428	46.2%	46.1%	49.2%	26.2%	63.6%	83.9%	56.9%
1985-1999	12,600	46.6%	46.5%	49.3%	26.6%	63.8%	82.9%	56.7%
1985-2000	13,745	46.6%	46.9%	49.3%	26.9%	49.8%	83.4%	55.5%
1985-2001	14,877	46.5%	47.0%	49.8%	27.2%	50.4%	84.2%	56.2%
1985-2002	15,991	46.5%	47.3%	49.9%	27.4%	50.3%	84.8%	54.6%
1985-2003	17,102	46.3%	47.6%	49.8%	27.7%	50.8%	84.7%	55.1%
1985-2004	18,159	46.0%	47.6%	49.8%	28.0%	51.2%	84.0%	54.2%
1985-2005	19,159	45.2%	47.5%	49.5%	28.2%	51.1%	83.4%	53.4%
1985-2006	20,050	46.9%	47.3%	49.4%	24.4%	51.4%	85.3%	56.8%
1985-2007	21,235	46.5%	47.0%	49.0%	28.4%	50.7%	83.0%	51.6%
Average over estimation windows		**46.3%**	**46.8%**	**49.3%**	**26.7%**	**55.5%**	**84.2%**	**55.4%**

Table 5: Out-of-sample hit rates of rating prediction models using seven variables from Blume, Lim and MacKinlay (1999)

The table contains the hit rates – i.e. the percentage of correctly predicted observations – for the rating prediction models linear discriminant analysis (LDA), linear regression, ordered probit regression, tallying, Categorization by Elimination (CBE), the decision-tree algorithm C4.5 and a multilayer perceptron neural network.
The cues used for rating prediction are those used by BLM (1998). The cues are winsorized at the 1%- and 99%-quantiles. For CBE, they are ordered according to their unconditional Pseudo-R^2 in an ordered probit regression of ratings. Estimation windows are expanding. They start in 1985 and end in the year prior to the one for which an out-of-sample prediction is made.

Training & validation data sets			**Hit rates of the competitor models**						
Training	Validation	# obs. validation	LDA	Linear regression	Ordered probit	Tallying	CBE	C4.5	Neural network
1985-1995	1996	1,008	38.9%	47.7%	50.6%	30.6%	65.8%	53.2%	51.0%
1985-1996	1997	1,084	40.9%	48.4%	50.9%	30.0%	66.0%	52.6%	55.1%
1985-1997	1998	1,166	40.6%	50.0%	52.8%	29.5%	66.0%	53.4%	53.9%
1985-1998	1999	1,172	42.7%	51.1%	53.6%	30.4%	65.5%	54.9%	56.5%
1985-1999	2000	1,145	47.0%	50.5%	52.6%	30.5%	61.3%	55.1%	58.9%
1985-2000	2001	1,132	46.6%	47.3%	50.1%	30.8%	53.0%	52.2%	55.7%
1985-2001	2002	1,114	43.9%	49.8%	52.7%	30.5%	54.9%	51.9%	52.2%
1985-2002	2003	1,111	41.9%	46.8%	49.0%	31.2%	58.0%	53.1%	51.3%
1985-2003	2004	1,057	37.7%	48.7%	49.7%	32.8%	60.7%	53.3%	48.6%
1985-2004	2005	1,000	37.7%	46.3%	49.2%	33.2%	59.2%	49.4%	53.5%
1985-2005	2006	891	33.9%	43.5%	44.5%	33.8%	58.5%	47.3%	46.3%
1985-2006	2007	1,185	32.2%	37.3%	39.1%	31.0%	50.3%	43.3%	38.6%
Average over estimation windows			**40.3%**	**47.3%**	**49.6%**	**31.2%**	**59.9%**	**51.6%**	**51.8%**

Table 6: Alternative measures for prediction accuracy

The table presents the cumulative distribution of deviations between predicted and actual ratings. The category of +/- 0.5 rating grades corresponds to the hit rate. The results shown in this table are for predictions based on the seven variables from Blume, Lim and MacKinlay (1999).

Vali-dation	Model analysed	Cumulative share of ratings (in %) with deviation between predicted and actual rating smaller or equal than …							Mean absolute error
		+/- 0.5	+/- 1.0	+/- 2.0	+/- 3.0	+/- 4.0	+/- 5.0	+/- 6.0	
In-sample	LDA	46.3	90.7	98.1	99.5	100.0			0.65
	Linear regression	46.8	79.6	98.4	99.9	100.0			0.75
	Ordered probit	49.3	95.2	99.8	100.0				0.56
	Tallying	26.7	69.2	96.3	100.0				1.08
	CBE	55.5	83.0	98.0	99.9	100.0			0.63
	C4.5	84.2	97.8	100.0					0.18
	Neural network	55.4	96.6	99.2	100.0				0.49
Out-of-sample	LDA	40.3	83.6	95.7	99.1	99.9	100.0		0.81
	Linear regression	47.3	79.4	98.0	99.9	100.0			0.75
	Ordered probit	49.6	94.8	99.6	100.0				0.56
	Tallying	31.2	73.9	97.0	100.0				0.98
	CBE	59.9	86.0	98.5	99.9	100.0			0.55
	C4.5	51.6	91.5	99.0	100.0				0.58
	Neural network	51.8	95.0	99.2	100.0				0.54

Table 7: Average hit rates of rating prediction models using expanding estimation windows and sets of four or eight variables

The table contains the average hit rates – i.e. the average percentage of correctly predicted observations – for the rating prediction models linear discriminant analysis (LDA), linear regression, ordered probit regression, tallying, Categorization by Elimination (CBE), the decision tree algorithm C4.5 and a multilayer perceptron neural network.

Cues are selected from Blume, Lim and MacKinlay (1998), Standard and Poor's (2006, 2008) and Metz and Cantor (2006), comprising a total of 31 variables. We order the cues according to the Pseudo-R^2 in ordered probit regressions of the S&P rating on the respective cue. For each of the categories – business risk, leverage, profitability, interest coverage – the two cues with the highest Pseudo-R^2 are used in our simulation studies. For the analyses with 4 cues we have chosen the highest-scoring cues for the categories business risk, leverage, profitability and interest coverage – i.e. the standard error, long-term debt to assets, Moody's ROA and Mooody's interest coverage. For the analyses with 8 cues, market capitalization, Moody' debt to assets, S&P's return on capital and S&P's FFO interest coverage join the cue selection. The cues are winsorized at the 1%- and 99%-quantiles and are ordered according to their validity.

For the in-sample analysis, the training and validation sets are identical. Estimations are performed on expanding windows (1985-1995, 1985-1996,…, 1985-2007). For the out-of-sample analysis, the training sets are also expanding (1985-1995, 1985-1996,…, 1985-2006). The validation is always performed on the observations of the subsequent year. The hit rates shown in the table are hit rates averaged over the estimation windows.

		Average hit rates of the competitor models (over all estimation windows)						
Validation	# of cues used	LDA	Linear regression	Ordered probit	Tallying	CBE	C4.5	Neural network
In-sample	4	46.2%	47.4%	51.2%	29.2%	57.2%	70.6%	52.1%
In-sample	8	46.5%	48.8%	52.3%	27.4%	58.9%	84.5%	57.0%
Out-of-sample	4	39.8%	47.1%	50.5%	33.3%	64.7%	49.6%	50.6%
Out-of-sample	8	37.2%	46.0%	49.0%	30.3%	68.1%	51.6%	52.4%

Table 8: Average hit rates of rating prediction models using time-independent cross-validation and sets of four or eight variables

The table contains the average hit rates – i.e. the average percentage of correctly predicted observations – for the rating prediction models linear discriminant analysis (LDA), linear regression, ordered probit regression, tallying, Categorization by Elimination (CBE), the decision tree algorithm C4.5 and a multilayer perceptron neural network.

Cues are selected from Blume, Lim and MacKinlay (1998), Standard and Poor's (2006, 2008) and Metz and Cantor (2006), comprising a total of 31 variables. We order the cues according to the Pseudo-R^2 in ordered probit regressions of the S&P rating on the respective cue. For each of the categories – business risk, leverage, profitability, interest coverage – the two cues with the highest Pseudo-R^2 are used in our simulation studies. For the analyses with 4 cues we have chosen the highest-scoring cues for the categories business risk, leverage, profitability and interest coverage – i.e. the standard error, long-term debt to assets, Moody's ROA and Mooody's interest coverage. For the analyses with 8 cues, market capitalization, Moody' debt to assets, S&P's return on capital and S&P's FFO interest coverage join the cue selection. The cues are winsorized at the 1%- and 99%-quantiles and are ordered according to their validity.

The rating prediction models are tested on samples of 100, 1,000 and 10,000 observations that are randomly drawn from the 1985-2007 data set. For each sample size, 1,000 simulation draws are performed. The results of these 1,000 simulations are then aggregated to receive the average hit rate per rating prediction model. For the out-of-sample analysis, each of the randomly drawn sample is split into two halves.

			Average hit rates of the competitor models						
Validation	# of cues used	Sample size (training/ validation)	LDA	Linear regression	Ordered probit	Tallying	CBE	C4.5	Neural network
In-sample	4	100	47.0%	43.0%	47.5%	33.0%	63.3%	80.6%	55.4%
	4	1,000	42.0%	41.6%	46.0%	28.5%	57.8%	75.9%	50.7%
	4	10,000	42.1%	41.5%	45.8%	26.4%	51.1%	68.1%	49.6%
In-sample	8	100	54.0%	47.9%	51.1%	35.5%	65.9%	89.0%	60.8%
	8	1,000	44.2%	46.1%	48.5%	30.2%	58.2%	83.5%	53.9%
	8	10,000	44.1%	47.3%	48.6%	27.5%	57.5%	82.9%	53.6%
Out-of-sample	4	50/50	33.5%	37.3%	38.2%	33.0%	47.5%	33.6%	35.2%
	4	500/500	34.9%	40.8%	42.4%	32.1%	58.4%	43.2%	49.8%
	4	5,000/5,000	35.3%	41.6%	43.1%	28.7%	61.9%	48.7%	50.1%
Out-of-sample	8	50/50	32.6%	35.1%	36.5%	32.2%	42.1%	28.8%	37.1%
	8	500/500	33.7%	38.6%	41.3%	34.1%	56.0%	45.9%	51.3%
	8	5,000/5,000	32.4%	39.5%	41.4%	29.0%	60.5%	53.4%	54.1%

Table 9: Performance decomposition of CBE

The table contains detailed results on the CBE rating prediction algorithm. Unique decisions are those decisions where CBE stops before the last cue that could be considered. All other decisions are averaging decisions.

Panel 1: In-sample analysis

Sample size	# of cues used	CBE – overall hit rate	CBE performance split			
			Unique decisions	Unique decision hit rate	Averaging decisions	Averaging decision hit rate
Data from Table 7						
100	3.68 / 4	63.3%	24.0%	58.4%	76.0%	64.8%
1,000	3.98 / 4	57.8%	3.0%	30.2%	97.0%	58.6%
10,000	3.99 / 4	51.1%	0.2%	20.0%	99.8%	51.2%
Data from Table 7						
100	6.50 / 8	65.9%	31.7%	63.0%	68.3%	67.2%
1,000	7.89 / 8	58.2%	6.1%	35.8%	93.9%	59.7%
10,000	7.99 / 8	57.5%	0.5%	25.9%	99.5%	57.6%
Data from Table 4						
21,235	6.92 / 7	55.9%	0.8%	25.5%	99.2%	56.1%

Panel 2: Out-of-sample analysis

Sample size (training/ validation)	# of cues used	CBE – overall hit rate	CBE performance split			
			Unique decisions	Unique decision hit rate	Averaging decisions	Averaging decision hit rate
Data from Table 7						
50/50	3.01 / 4	47.5%	49.5%	32.3%	50.5%	62.4%
500/500	3.90 / 4	58.4%	9.9%	24.2%	90.1%	62.2%
5,000/5,000	3.99 / 4	61.9%	0.4%	18.0%	99.6%	62.1%
Data from Table 7						
50/50	4.32 / 8	42.1%	68.2%	35.2%	31.8%	56.9%
500/500	7.43 / 8	56.0%	21.4%	28.1%	78.6%	63.6%
5,000/5,000	7.95 / 8	60.5%	2.0%	25.6%	98.0%	61.2%
Data from Table 5						
21,235	6.85 / 7	59.9%	0.1%	21.3%	99.9%	60.0%

Figure 1 – Example for the application of the Categorization by Elimination (CBE) to rating decisions.

In the example, the following cues are considered: leverage (highest validity), return on assets, operating margin, interest coverage (lowest validity). A given cue value is associated with a rating if it falls in the range defined for the rating. CBE starts with the most valid cue, proceeding to the one with the next highest validity.

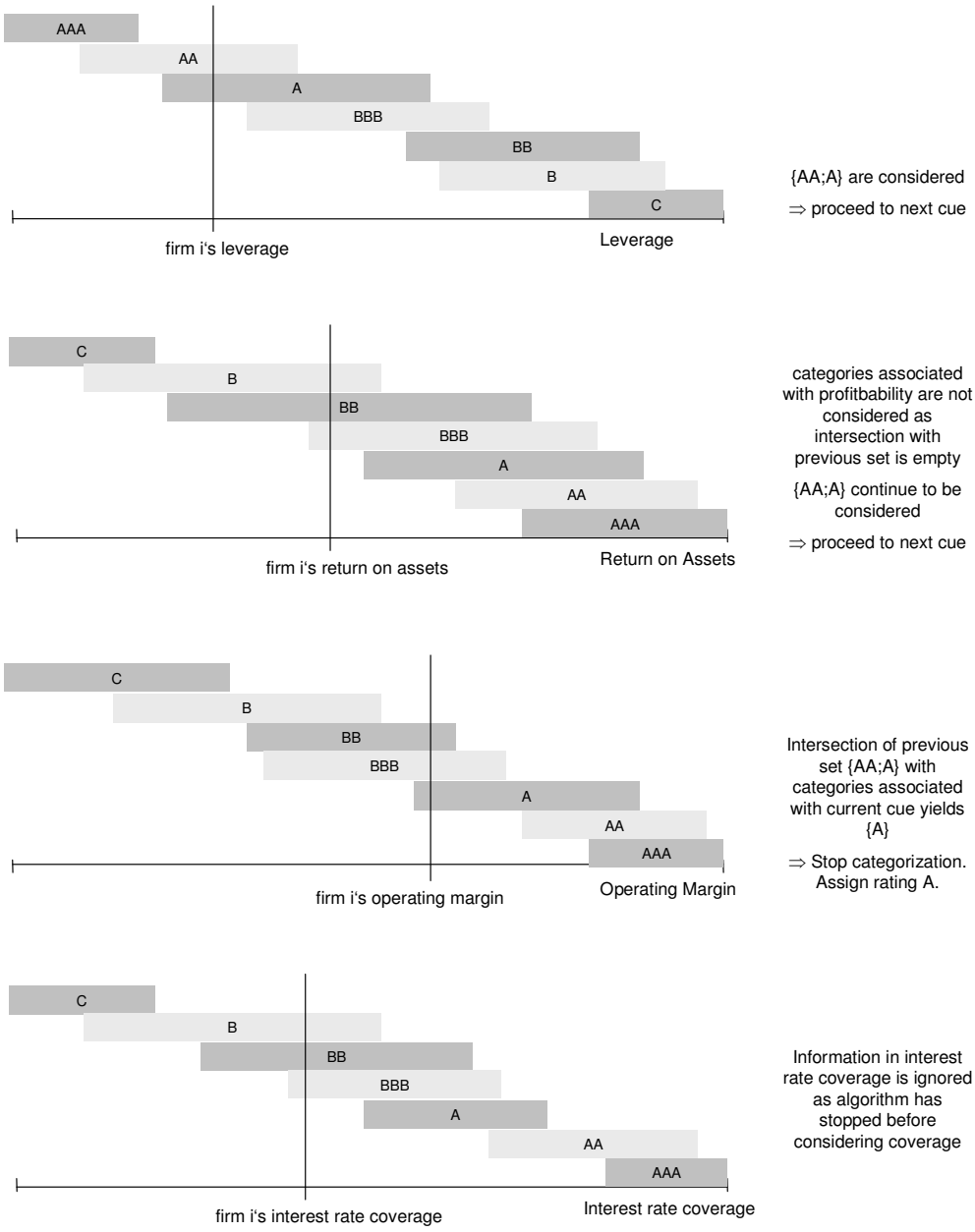

Figure 2 – Flow diagram of the Categorization by Elimination algorithm (Source: Berretty et al. (1997))