Selection vs. Averaging of Logistic Credit Risk Models

Evelyn Hayden
Raiffeisen Bank International

Alex Stomper*
Humboldt University Berlin (HU) and ECGI

Arne Westerkamp
Vienna University of Economics and Business

August 6, 2012

Abstract

We evaluate the relative performance of logistic credit risk models that were selected by means of standard stepwise model selection methods and "average" models obtained by Bayesian model averaging (BMA). Our bootstrap analysis shows that BMA should be considered as an alternative to the stepwise model selection procedures that are currently often used in practice.

Keywords: Credit risk models, logistic models, stepwise model selection, Bayesian model averaging
JEL Classification: C 51, C 52, G 10

1 Introduction

In credit risk modeling, standard practice tends to ignore model uncertainty. Analysts construct credit risk models based on model selection heuristics but use the models as if they knew that they were correct. This practice is standard despite the existence of methods for taking model uncertainty into account: Bayesian model averaging (BMA) is a coherent way to form a weighted average of a class of possible models, based on weights that depend on the relative likelihood of each model, given the data.

In this paper, we compare two strategies for logistic credit risk modeling: (i) stepwise model selection based on heuristics that are widely used in practice, and (ii) Bayesian model averaging. We focus on logistic credit risk models because this model class is the current industry standard.[1] Within the logistic model class, our analysis compares the predictive performance of the credit risk models that result from the two modeling strategies, based on the same real-world data set.

In the first step of the analysis, we base our comparison on the raw data. We consider three standard measures of model performance, i.e. the accuracy ratio (AR), the Brier score (BS), and the logarithmic score (LS).[2] A bootstrap analysis reveals that BMA yields credit risk models

*Corresponding author. Email: stompera@hu-berlin.de The views expressed in this paper are those of the authors and not necessarily those of the Raiffeisen Bank International.

[1] In logistic regression, the log-odds (or "logits") of the probability of default are modeled as a linear function of the independent predictors

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \text{a} + \sum_i b_i x_i, \tag{1}$$

where a is the intercept, b_i are the coefficients and x_i are the independent variables. It is available in every statistical software package.

[2] The model performance measures will be described in detail below.

with a better predictive performance than that of models selected by means of the stepwise model selection heuristics. This result is consistent with theoretical considerations and with empirical results of related prior analyses based on medical data.[3]

In the second part of our analysis, we investigate possible reasons for our finding that BMA beats stepwise model selection as a strategy for constructing logistic credit risk models. We obtain a surprising result: the stepwise model selection procedures perform as well as BMA if the data are first transformed in order to satisfy certain linearity assumptions of the logistic model class.[4] BMA thus appears not only as a remedy for model uncertainty within the class of logistic credit risk models, but also for violations of key assumptions behind this model class. This result is consistent with an intriguing argument of George (1999) that BMA within a model class can "approximate models outside the model class."

In summary, we find that BMA is highly commendable for credit risk modeling as an alternative to the stepwise model selection procedures that are currently commonly used. Moreover, BMA-based logistic credit risk modeling appears to be more robust than stepwise model selection with respect to violations of linearity assumptions behind the logistic model class. The latter finding is particularly relevant in situations in which a credit risk model should be based on raw data about debtors in order to enhance the model's transparency. In our experience, such situations will often occur in practice because loan officers tend to lack an intuitive grasp of the data transformations that are required in order to avoid violations of linearity assumptions behind logistic credit risk models. Our analysis suggests that BMA is commendable for constructing models based on easy-to-understand untransformed data about credit applicants.

This paper is structured as follows. In the next section, we describe the model selection and model averaging procedures that are the subject of our analysis. Section 3 presents our bootstrap analysis. Section 4 concludes.

2 Model Selection vs. Averaging

A large set of possible default predictor variables are commonly used in credit risk modeling. If this set contains k variables, there are 2^k possible prediction models. Model selection procedures are algorithms for selecting one of these 2^k models. Model averaging procedures form weighted averages of different models. This section presents the theories behind the model selection and -averaging procedures that we use in our analysis, along with critical appraisals of these procedures. Moreover, we will reference all software packages that we used in order to implement the procedures.

2.1 Stepwise Model Selection

Theory: Stepwise model selection methods are based on test statistics for testing the relative explanatory power of two models: a model i and a model $i+1$ that differs from model i in terms of one explanatory variable. The model selection can proceed either "forward" or "backward", i.e. from smaller to larger models ($i = 0, 1, ...$) or in the opposite direction ($i = k, k-1, ...$, where k denotes the number of candidate default prediction variables). In forward selection, each "step" is a test of the null hypothesis that model $i+1$ differs from model i by an additional

[3] We are aware of two studies in the medical arena that compare BMA to stepwise selection for logistic models. Viallefont, Raftery, and Richardson (2001) find that BMA gives high weights to models containing the "true" risk factors while stepwise selection methods (based on p-values) very often select irrelevant variables. Wang, Zhang, and Bakhai (2004) compare BMA and various stepwise model selection methods with respect to model performance in predicting coronary events out-of-sample. We know of only one other paper on Bayesian methods in credit risk modeling: Zhang and Härdle (2010) analyze the predictive performance of Bayesian additive classification trees. We consider BMA of standard logistic regression models, rather than of classification trees.

[4] A detailed description can be found below. In short, each default predictor variable is transformed into Hodrick-Prescott filtered empirical log-odds of default associated with the variable.

explanatory variable with a coefficient equal to zero. This null hypothesis is tested against the alternative that the coefficient is non-zero, i.e. that model $i+1$ contains an additional relevant variable. If the null hypothesis can be rejected for some possible extension of model i, then model i is replaced by a model $i+1$ that takes the place of model i in the next step of the algorithm.[5] In backward selection, the roles of models i and $i+1$ are reversed.

Stepwise model selection procedures have been criticized in the literature. In each step i, these procedures rely on test statistics that are conditional on previous steps, but are not properly adjusted. Moreover, the procedures tend to be unstable. For example, Derksen and Keselman (1992) found that the procedures are overly susceptible with respect to correlation of possible predictor variables.

Implementation: In our empirical analysis, the stepwise model selection processes are implemented using STATA's *stepwise* command. We consider both forward and backward selection.[6] In each step of the selection processes, the requisite hypotheses tests are specified as likelihood ratio tests. In forward selection, new variables are included if the likelihood ratio test rejects the above-stated null with a p value below 5%.[7] In backward selection, variables are excluded if their marginal contribution to the model's likelihood ratio is insignificant at the 10% level.[8] The selection processes terminate once no further variable eligible for inclusion (in forward selection) or exclusion (in backward selection) can be found.

2.2 Bayesian Model Averaging

Theory: Bayesian model averaging (BMA) is a way to take model uncertainty into account by forming a weighted average of all possible models, rather than selecting one model deemed "the best". In the model averaging, each candidate model receives a weight that depends on the model's likelihood, given the data and a prior.

To explain the intuition behind BMA in the most accessible way, we consider the case in which there are just two candidate models, i.e. models M_1 and M_2; the general case with m possible models is presented in Appendix 1. If there are just two candidate models, the right "mix" of these models is determined by the following ratio of posterior probabilities with which each model is the correct model, given the data:

$$\psi = \frac{\text{pr}(M_2|D)}{\text{pr}(M_1|D)} = \frac{\text{pr}(D|M_2)}{\text{pr}(D|M_1)} \frac{\text{pr}(M_2)}{\text{pr}(M_1)}, \tag{2}$$

where Bayes' rule has been used in order to obtain the product on the right-hand side. The first factor of this product is crucial for understanding BMA and is referred to as the Bayes factor. This factor is the ratio of the marginal probabilities of observing the data at hand under the models M_1 and M_2. A Bayes factor above one indicates that model M_2 is more likely to be the correct model than model M_1. The second factor of the product is the ratio of the prior probabilities of models M_1 and M_2 being correct. It is often assumed that $\text{pr}(M_2) = \text{pr}(M_1)$ such that the optimal "mix" of the models M_1 and M_2 is fully determined by the Bayes factor, and, thus, by the data.

Given the ratio ψ, the "mixing" of the models M_1 and M_2 occurs by taking a weighted average of the models' coefficient vectors. The coefficient vector of the "mixed" model is given

[5] There are, of course, usually many ways to extend a model i by adding an additional explanatory variable in order to obtain a model $i+1$. The algorithm chooses the model $i+1$ for which the null hypothesis can be rejected most resoundingly.

[6] We also used combinations of these two basic procedures, but the resulting models were almost always identical to those obtained by means of pure forward or pure backward selection.

[7] The corresponding STATA command is stepwise, pe(0.05) lr logit.

[8] The corresponding STATA command is stepwise, pr(0.10) lr logit.

by

$$\bar{\beta} = \frac{\psi}{1+\psi}\beta_1 + \frac{1}{1+\psi}\beta_2. \tag{3}$$

where β_1 and β_2 denote the coefficient vectors of the models M_1 and M_2. The higher the ratio ψ, the more will the coefficient vector of the mixed model, $\bar{\beta}$, resemble the coefficient vector of model M_1, rather than that of model M_2. This specification is intuitively appealing because a higher value of ψ indicates a higher probability with which model M_1 is indeed the correct model, given the data at hand. If one model contains a variable that is not contained in other model(s), the averaged coefficients of this variable is "shrunk" in that the average is taken across some models in which the variable effectively has a coefficient of zero (since it is not contained in the models); BMA is, therefore, a shrinkage estimator.

The performance of BMA has been analyzed theoretically. A succinct summary appears in Raftery and Zheng (2003): "[W]hen used for model selection, the Bayes factor minimizes the total error rate (sum of Type I and Type II error probabilities); BMA point estimators and predictions minimize mean squared error (MSE); BMA estimation and prediction intervals are calibrated; and BMA predictive distributions have optimal performance in terms of the log score performance measure." Our analysis can be seen as being inspired by these statements. For example, we will compute log scores in order to analyze the performance of models generated by means of BMA, and expect to find that these models outperform models obtained by means of stepwise forward or backward selection.

Implementation: We use a standard implementation of BMA that is part of the free statistical computing software "R", namely the routine bic.glm in the package BMA.[9] This routine does actually not average over the entire space of all 2^k conceivable models, but rather restricts the BMA to a subset of models that are sufficiently likely to be correct and, at the same time, sufficiently parsimonious. The first criterion requires that the set will only contain models that are sufficiently likely relative to the most likely model, given the data. We specify this criterion for inclusion of a model as requiring the model's likelihood to be larger than 5% of the highest likelihood observed.[10] The second criterion excludes models that are overly complex in that they contain smaller models (with fewer variables) that are more likely.[11]

3 A Bootstrap Analysis

We conduct a bootstrap analysis in order to compare the model selection and model averaging procedures that were described above. Our bootstrap analysis resembles those of Li, Zhang, Rosenzweig, Wang, and Chan (2002) and Engelmann, Hayden, and Tasche (2003b) in that we draw 1,000 bootstrap samples by sampling with replacement. Each sample contains 3,738 observations which are further split into two subsamples. The first subsample contains 60% of the observations and is used for constructing credit risk models. The second subsample contains 40% of the observations and is used for model validation. We assign the observations to the two subsamples such that the fraction of defaults is held constant.

Each bootstrap sample will serve as the data input for a "horse-race" between three logistic credit risk models: one model that we obtain by BMA, and two models obtained through stepwise model selection, one of which is the result of forward selection while the other results from backward selection. We compare the models' "out-of-sample" performance, i.e. their performance in predicting defaults in the validation sample. The comparison is based on the models' predicted default probabilities, which can be interpreted as "credit scores".

[9] For BMA of logistic models, the routine has to be called with the setting familiy=binomial.
[10] In terms of the bic.glm routine's code, the relevant setting is OR=20.
[11] In terms of the bic.glm routine's code, the relevant setting is strict=FALSE.

We use three common measures of the model performance: the accuracy ratio (AR), the Brier score (BS), and the logarithmic score (LS). All of them are used and analyzed extensively in the fields of Bayesian econometrics and (meteorological) forecasting. Gneiting and Raftery (2007), Johnstone, Jose, and Winkler (2011) and Winkler and Jose (2011) provide excellent overviews and further references. The fact that different standard utility functions necessarily imply different optimal "scoring functions", i.e., performance measures, is one of the reasons that the simultaneous reporting of different performance measures is recommended (see for example, Fildes and Goodwin, 2007; Gneiting, 2011; Winkler and Murphy, 1992). The three performance measures employed in this study are described and defined in Appendix 2. To interpret our results below, it suffices to know that (i) the AR is an ordinal model performance measure, while the BS and the LS are cardinal measures, and (ii) that the BS is an inverse measure of performance: a smaller value of BS indicates better performance.[12]

3.1 The Data

Our analysis is based on a data set that contains all debtors of the Austrian Tourism Bank. There are 3,738 firm-year observations concerning an unbalanced panel of 918 small-and-medium-sized enterprises (SMEs) during the years 2000-07. We use the data in order to construct the default prediction variables defined in Altman and Sabato (2007) for predicting defaults of SMEs. These variables are summarized in Table 1. To reduce the impact of outliers, we replace all values above the 1% (99%) quantile with the 1% (99%) quantile itself (i.e. we winsorize all variables at the 1%-level).

The dependent variable of our analysis is an indicator variable that marks firm-years in which a firm formally applied for permission to miss a payment on a loan by more than 90 days. We interpret such an application as a "default" and exclude any firm-years that follow the first default of a firm. Based on these conventions, our sample features a default rate of 1.6%.

3.2 Results

Our bootstrap analysis yields realizations of each performance measure (AR, BS, and LS) for 1,000 models constructed by means of BMA, stepwise forward selection (SFS), and stepwise backward selection (SBS), respectively. To summarize these results, we first present OLS regressions of the model performance measures on dummy variables indicating the methods by which the models were obtained. By leaving out the dummy indicating BMA, we obtain coefficients that measure differences in model performance between BMA and the two stepwise model selection procedures. In all regressions, we control for cross-sectional variation in the fraction of default-observations that are included in the bootstrap samples.

The OLS estimates appear in Table 2. They suggest that the choice between BMA and stepwise model selection methods matters only little if model performance is measured in terms of AR. The average AR of models obtained through BMA is only about 4% (5.8%) higher (smaller) than that of models selected by means of SFS (SBS).[13] However, BMA seems to substantially outperform the stepwise model selection methods in terms of the BS. The average BS of models obtained through BMA exceeds that of models selected by SFS (SBS) by 13% (32%).[13] A comparison in terms of the LS yields similar results.

So far, our estimates suggest that BMA is superior to stepwise model selection. As an ordinal measure of model performance, the AR only contains information about the discriminatory power of different credit risk models. Our results suggest that BMA yields models with rather

[12] As an ordinal model performance measure, the AR measures a model's performance only in terms of the model's ability to rank debtors relative to each other. A model may thus appear to be perfect even if it over- or under-estimates default rates. Cardinal model performance measures depend on predicted vs. actual default rates, rather than just on the ranking of debtors.

[13] These percentages are relative to the regression intercept.

Table 1: Candidate default predictors

Starred variables are variables that were not used by Altman and Sabato (2007) but were suggested to us by the bank that provided us with the data. All other variables are defined as in Altman and Sabato (2007).

Category	Candidate Default Predictors
Leverage	Short term debt / Equity (book value)
	Equity (book value) / Total liabilities
	Liabilities / Total assets
	Liabilities / Sales*
	Liabilities / Ordinary cash flow*
Liquidity	Cash / Total assets
	Working capital / Total assets
	Cash / Net sales
	Intangible / Total assets
	Ordinary cash flow / Short term liabilities*
	Current assets / Short term liabilities*
Profitability	EBIT / Sales
	EBITDA / Total assets
	Net income / Total assets
	Retained earnings / Total assets
	Net income / Sales
	Ordinary cash flow / Sales
Coverage	EBITDA / Interest expenses
	EBIT / Interest expenses
Activity	Sales / Total assets
	Account payable / Sales
	Account receivable / Liabilities
Industry-specific	Personnel expenses / Total expenses*
	Goods and material employed / Total expenses*
	Services / Personnel expenses*

Table 2: Regressions explaining out-of-sample model performance

Ordinary least squares regression of out-of-sample model performance measures on dummy variables that indicate different model selection procedures, and the fraction of default observations included in the underlying bootstrap samples. Results from Bayesian model averaging represent the left-out category. t-statistics in parenthesis, asterisks denote the significance level: *** p<0.01, ** p<0.05, * p<0.1

	Accuracy Ratio *higher* is better	Brier Score *smaller* is better	Log Score *higher* is better
stepwise forward	-0.0164***	0.000111***	-4.842***
	(-3.396)	(4.767)	(0.671)
stepwise backward	0.0237***	0.000281***	-4.408***
	(4.918)	(12.08)	(0.671)
bootstrap default rate	6.379***	0.933***	-4910***
	(6.664)	(201.6)	(133.4)
Constant	0.412***	0.000872***	-37.86***
	(26.49)	(11.60)	(2.166)
Observations	3000	3000	3000
R^2	0.037	0.932	0.321

similar discriminatory power as the models selected through stepwise model selection heuristics. Both types of models can thus be used for comparing debtors in terms of their credit scores. If credit scores are however used for computing default probabilities, BMA yields models that tend to outperform models selected through stepwise heuristics.

The results of our regression analysis can be checked in a number of ways. Similar results are obtained if Wilcoxon matched pairs signed rank tests are used to test for differences in mean model performance. In addition, one can *separately* compare the relative performance of the models that BMA and stepwise model selection yield, for each of the 1,000 bootstrap samples. Appendix 2 presents suitable test statistics for tests comparing models based on the AR and the BS.[14] The results show that BMA regularly outperforms stepwise model selection if model performance is measured in terms of the BS. In terms of AR, the comparison yields no clear conclusion.[15]

Besides using different statistical tests to check the robustness of our results, we can also check whether the results depend on other features of our analysis. Such checks show that the results are not driven by possible problems of multi-collinearity, insufficient winsorization, specifics of our stepwise model selection procedures, etc.

There is, however, one feature of our analysis that does turn out to have some bearing on the results of our comparison of BMA and stepwise model selection: whether we use the variables defined by Altman and Sabato (2007) as such, or instead transform all of these variables by mapping them into smoothened log-odds of default. This data transformation is motivated by a key assumption of logistic models: that there is a multivariate linear relation between the explanatory variables and the log-odds of the event modeled.[16] In credit risk modeling, this assumption will be violated if a logistic model is based on variables that are nonlinearly related to the log-odds of defaults. It is however possible to mitigate such violations of linearity

[14] No test statistics are available for testing model performance in terms of the LS.

[15] In terms of AR, BMA significantly outperforms SFS (SBS) in 485 (358) of the 1,000 bootstrap samples, and yields significantly worse models in 308 (481) cases. In terms of the BS, BMA significantly outperforms SFS (SBS) in 312 (404) cases and yields significantly worse models in only 116 (115) cases.

[16] See footnote 1

Table 3: Regressions explaining out-of-sample performance of models based on log-odds of default

Ordinary least squares regression of out-of-sample model performance measures on dummy variables that indicate different model selection procedures, and the fraction of default observations included in the underlying bootstrap samples. Results from Bayesian model averaging represent the left-out category. t-statistics in parenthesis, asterisks denote the significance level: *** p<0.01, ** p<0.05, * p<0.1

	Accuracy Ratio *higher* is better	Brier Score *smaller* is better	Log Score *higher* is better
stepwise forward	0.0180***	1.92e-05	0.287
	(4.768)	(0.810)	(0.323)
stepwise backward	0.0358***	6.51e-05***	0.471
	(9.480)	(2.740)	(0.323)
bootstrap default rate	2.343***	0.905***	-4993***
	(3.124)	(191.7)	(64.18)
Constant	0.555***	0.00109***	-30.29***
	(45.58)	(14.16)	(1.042)
Observations	3000	3000	3000
R^2	0.032	0.925	0.669

assumptions by specifying logistic models based on transformed variables: each variable is simply mapped into the log-odds of default that are associated with this variable. We perform such a mapping for each bootstrap sample and each variable defined in Table 1 by (i) defining 20 quantiles of each default predictor variable stated in Table 1, (ii) computing the default rate for each quantile, (iii) converting these default rates into log-odds of default, and (iv) smoothing the log-odds across quantiles by means of a Hodrick-Prescott filter, as described by Falkenstein, Boral, and Carty (2000). The smoothened log-odds are then used as inputs of our BMA and stepwise model selection procedures.

Table 3 presents the output of a re-run of our prior analysis based on the same bootstrap samples, after all variables have been mapped into log-odds of default as described above. Comparing these results to those in Table 2 shows that the mapping improves the performance of models obtained from stepwise model selection procedures relative to that of models constructed by means of of BMA. The stepwise procedures now even tend to outperform BMA, but only to a small extent. BMA does however appear to be more robust than stepwise model selection procedures with respect to violations of linearity assumptions of the class of logistic models. If no care is taken to avoid such violations, stepwise model selection will yield models that perform substantially worse than models constructed by means of BMA.

4 Conclusion

Our analysis suggests that BMA should be considered as an alternative to stepwise model selection procedures. We find that BMA can yield credit risk models with superior performance. Stepwise model selection methods do seem to work as well as BMA if model performance is merely measured in terms of a model's capacity to rank debtors. BMA is however commendable for constructing models that are supposed to measure debtors' absolute default probabilities.

We also obtain findings that hold a lesson for proponents of stepwise model selection procedures. We find that such procedures should not be used to select logistic credit risk models unless the input data has first been transformed in order to avoid violations of linearity assumptions

of the class of logistic models. If such data transformations are deemed to be too cumbersome or too abstract for the typical loan officer, BMA is preferable to stepwise model selection as a strategy for logistic credit risk modeling.

Appendix 1: Bayesian model averaging of m models:

This section extends the formulation of BMA from two to m models. As discussed above, BMA is a way to construct an average model, the coefficient vector of which are weighted averages of the coefficient vectors of all candidate models. In this weighted average, any model i's weight is computed as follows:

$$w_i = \frac{\text{pr}(D|M_i)\text{pr}(M_i)}{\sum_{j=1}^{m} \text{pr}(D|M_j)\text{pr}(M_j)}, \tag{4}$$

where D denotes the data, $\text{pr}(M_i)$ denotes the prior probability of model i, and $\text{pr}(D|M_i)$ is the marginal probability of the data given model M_i. The latter is the integral of the likelihoods of different parameterizations θ_i of the model, given a prior $\text{pr}(\theta_i|M_i)$:

$$\text{pr}(D|M_i) = \int \text{pr}(D|\theta_i, M_i)\text{pr}(\theta_i|M_i)\text{d}\theta_i. \tag{5}$$

Appendix 2: Measuring and testing model performance

The accuracy ratio: The accuracy ratio (AR) is a statistic derived from the cumulative accuracy profile (CAP), a curve which visually represents the discriminative power of a risk model. We therefore first explain the concept of the CAP. To obtain the CAP, all debtors are first ordered by their respective credit scores (i.e. the predictions of a credit risk model) from riskiest to safest. For any given x-quantile of the debtors' credit scores, the CAP depicts the percentage $d(x)$ of debtors with scores worse than x who have defaulted on their debts.

The AR measures the quality of a credit risk model in terms of the model's CAP relative to two extremes: (i) a hypothetical perfect model, and (ii) a "null" model without any power to discriminate between debtors. A perfect rating model will assign the worst scores to the defaulters. In this case, the CAP increases linearly until it reaches $d(x) = 1$ and then flattens out. For the null model without any discriminative power, $d(x) = x$. Given these two extremes, the AR of a credit risk model is defined as the ratio of (i) the area between the model's CAP and the CAP of the null model, and (ii) the area between the perfect model's CAP and the CAP of the null model. Given this definition, the AR is a number between zero (the worst case) and one (the best case).

The Brier score: In contrast to the AR, which is an ordinal validation measure, the Brier score and the logarithmic score are cardinal model performance measures. The Brier score (BS) is calculated as follows:

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}(\pi_i - y_i)^2, \tag{6}$$

where π_i is the predicted probability of a default of debtor i and y_i is a default dummy that equals one if the debtor defaults and otherwise equals zero. Hence, the Brier score employs a quadratic loss function and is therefore the dichotomous analogue to the mean squared error for a continuous outcome. The expected Brier score can take values between zero for a perfect match between prediction and outcome and one in the worst case. Sensible Brier scores lie between 0 and 0.25.

The logarithmic score: The logarithmic score (LS) is calculated as follows:

$$\text{LS} = \sum_{i=1}^{n} \text{LS}_i \text{ , where } \quad \text{LS}_i = \begin{cases} \ln \pi_i & \text{if } y_i = 1, \\ \ln(1-\pi_i) & \text{if } y_i = 0. \end{cases} \tag{7}$$

The logarithmic score is always negative with a maximum value of zero when all events are predicted perfectly. While Winkler and Murphy (1968) note that the LS and BS are empirically often highly correlated, Bickel (2007) and Johnstone et al. (2011) stress that LS can represent non-linear utility functions better than BS. The reason is that compared to the quadratic loss function of BS, the effect of the logarithmic loss function is to give relatively more weight to the non-defaulting observations.

Test statistics for model performance measures: For the accuracy ratio, the test builds on work of DeLong, DeLong, and Clarke-Pearson (1988). Two ARs are deemed to be equal if the following test statistic is below a critical value of the chi-square distribution:

$$T = \frac{(AR_1 - AR_2)^2}{\text{Var}[AR_1] + \text{Var}[AR_2] - 2\,\text{Covar}[AR_1, AR_2]}, \tag{8}$$

where the variances and covariances are estimated as described in Engelmann, Hayden, and Tasche (2003a) and implemented in standard statistical software packages.

For the Brier score, Redelmeier, Bloch, and Hickam (1991) introduce a formal test for the equality of two Brier scores, provided that they both satisfy the calibration test proposed by Spiegelhalter (1986). We hence first perform the latter test, and then compute Redelmeier's test statistic for all pairs which pass Spiegelhalter's test with a p-value of more than 5%. Redelmeier's test statistic is defined as follows:

$$Z_R = \frac{\sum_{i=1}^{N} \left[\left(\pi_{i,m1}^2 - \pi_{i,m2}^2 \right) - 2 \left(\pi_{i,m1} - \pi_{i,m2} \right) y_i \right]}{\left[\sum_{i=1}^{N} \left[\left(\pi_{i,m1} - \pi_{i,m2} \right)^2 \left(\pi_{i,m1} + \pi_{i,m2} \right) \left(2 - \pi_{i,m1} - \pi_{i,m2} \right) \right] \right]^{0.5}}, \tag{9}$$

where the subscripts 1 and 2 relate to the two different models under considerations, π denotes the predicted default probability, and $y = 0$ for non-defaults and $y = 1$ for defaults.

References

Altman, E. I. and G. Sabato (2007). Modelling credit risk for SMEs: Evidence from the US market. *Abacus 43*(3), 332–357.

Bickel, J. E. (2007). Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis 4*(2), 49–65.

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics 44*(3), 837–845.

Derksen, S. and H. J. Keselman (1992). Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical & Statistical Psychology 45*(2), 265–282.

Engelmann, B., E. Hayden, and D. Tasche (2003a). Measuring the discriminative power of rating systems. *Discussion Paper 01/03 on Banking and Financial Supervision, Deutsche Bundesbank*.

Engelmann, B., E. Hayden, and D. Tasche (2003b). Testing rating accuracy. *Risk 16*(1), 82–86.

Falkenstein, E. G., A. Boral, and L. Carty (2000). RiskCalc for private companies: Moody's default model. *Moody's Investors Service - Global Credit Research*.

Fildes, R. and P. Goodwin (November/December 2007). Against your better judgment? how organizations can improve their use of management judgment in forecasting. *Interfaces 37*(6), 570–576.

George, E. I. (1999). Comment – Bayesian model averaging: A tutorial. *Statistical Science 14*(4), 409–412.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*(494), 746–762.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *The Journal of the American Statistical Association 102*(477), 359–378.

Johnstone, D. J., V. R. R. Jose, and R. L. Winkler (December 2011). Tailored scoring rules for probabilities. *Decision Analysis 8*(4), 256–268.

Li, J., Z. Zhang, J. Rosenzweig, Y. Y. Wang, and D. W. Chan (2002). Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry 48*(8), 1296–1304.

Raftery, A. E. and Y. Zheng (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association 98*(464), 931–938.

Redelmeier, D. A., D. A. Bloch, and D. H. Hickam (1991). Assessing predictive accuracy: How to compare Brier scores. *Journal of Clinical Epidemiology 44*(11), 1141–1146.

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine 5*(5), 421–433.

Viallefont, V., A. E. Raftery, and S. Richardson (2001). Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine 20*(21), 3215–3230.

Wang, D., W. Zhang, and A. Bakhai (2004). Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. *Statistics in Medicine 22*(22), 3451–3467.

Winkler, R. and A. Murphy (1968, October). 'good' probability assessors. *Journal of Applied Meteorology 7*, 751–758.

Winkler, R. L. and V. R. R. Jose (2011). Scoring rules. In *Encyclopedia of Operations Research and Management Science*. Wiley.

Winkler, R. L. and A. H. Murphy (1992, June). On seeking a best performance measure or a best forecasting method. *International Journal of Forecasting 8*(1), 104–107.

Zhang, J. L. and W. K. Härdle (2010). The Bayesian Additive Classification Tree applied to credit risk modelling. *Computational Statistics & Data Analysis 54*(5), 1197–1205.