Journal of Risk Model Validation (77–113) Volume 1/Number 1, Spring 2007

Benchmarking default prediction models: pitfalls and remedies in model validation

Roger M. Stein
Moody's Investors Service, 99 Church Street, New York, NY 10007, USA;
email: roger.stein@moodys.com

We discuss the components of validating credit default models with a focus on potential challenges to making inferences from validation under real-world conditions. We structure the discussion in terms of: (a) the quantities of interest that may be measured (calibration and power) and how they can result in misleading conclusions if not taken in context; (b) a methodology for measuring these quantities that is robust to non-stationarity both in terms of historical time periods and in terms of sample firm composition; and (c) techniques that aid in the interpretation of the results of such tests. The approaches we advocate provide means for controlling for and understanding sample selection and variability. These effects can in some cases be severe and we present some empirical examples that highlight instances where they are and can thus compromise conclusions drawn from validation tests.

1 INTRODUCTION

A model without sufficient validation is only a hypothesis. Without adequate objective validation criteria and processes, the benefits of implementing and using quantitative risk models cannot be fully realized. This makes reliable validation techniques crucial for both commercial and regulatory purposes.

As financial professionals and regulators become more focused on credit risk measurement, they have become similarly interested in credit model validation methodologies. In the case of default models, validation involves examining the goodness of the model along two broad dimensions: model *power* and model *calibration*. In practice, measuring these quantities can be challenging.

Power describes how well a model discriminates between defaulting ("Bad") and non-defaulting ("Good") firms. For example, if two models produced ratings of "Good" and "Bad", the more powerful model would be the one that had a higher percentage of defaults (and a lower percentage of non-defaults) in its "Bad"

This article was originally released as a working paper (Stein (2002)). The current version has been updated to reflect both comments the author has received since 2002 and new results that have emerged since the first draft was circulated. I wish to thank Jeff Bohn, Greg Gupton, Ahmet Kocagil, Matt Kurbat, Jody Rasch, Richard Cantor, Chris Mann, Douglas Lucas, Neil Reid and Phil Escott for the valuable input I received in writing the first draft. I had very useful conversations with David Lando, Alan White and Halina Frydman as well. Any errors are, of course, my own.

category and had a higher percentage of non-defaults in its "Good" category. This type of analysis can be summarized with *contingency tables* and *power curves*.

Calibration describes how well a model's predicted probabilities agree with actual outcomes. It describes how close the model's predicted probabilities match actual default rates. For example, if there were two models A and B that each predicted the two rating classes "Good" and "Bad", and the predicted probability of default for A's "Bad" class were 5% while the predicted probability of default for B's "Bad" class were 20%, we might examine these probabilities to determine how well they matched actual default rates. If we looked at the actual default rates of the portfolios and found that 20% of B's "Bad" rated loans defaulted while 1% of A's did, B would have the more accurate probabilities since its predicted default rate of 20% closely matches the observed default rate of 20%, while A's predicted default rate of 5% was very different to the observed rate of 1%. (Of course, testing on identical portfolios, were that possible, would be even better.) This type of analysis can be summarized by means of *likelihood* measures.

The statistical and econometrics literature on model selection and model validation is broad[1] and a detailed discussion is beyond the scope of this article. However, measures of "goodness of fit" are in-sample measures and as such, while useful during model construction and factor selection, do not provide sufficient information about the application of a model to real-world business problems. These more traditional measures are necessary, although not sufficient, for models to be useful.

The methodology described here brings together several streams of the validation literature that we have found useful in evaluating quantitative default models.

In this article we discuss various methods for validating default risk models during the development process and discuss cases in which they may break down. We discuss our approach along three lines:

1. what to measure;
2. how to measure it; and
3. how to interpret the results.

We then make some recommendations about how these factors can be used to inform model selection.

These techniques are most useful in the context of model development where ample data is available. Although they also have implications for the validation of third-party models and other internal rating systems by banks and others, in some cases data limitations may make their application difficult in these situations and other approaches may be more suitable. That said, we feel that these techniques provide several validation components that represent best practices in model development.

The remainder of this article proceeds as follows: Section 2 describes some common measures of model performance and discusses their applicability to

[1] See, for example, Burnham and Anderson (1998) or Greene (2000).

various aspects of model validation; Section 3 describes the validation approach called "walk-forward" testing that controls for model overfitting and permits an analysis of the robustness of the modeling method through time; in Section 4 we discuss some of the practical concerns relating to sample selection and bias that can arise when testing models on empirical data, we also discuss how the definition of validation can be broadened to answer a variety of other questions; and in the concluding section we provide a summary and suggest some general guidelines for validation work.

2 WHAT TO MEASURE: SOME METRICS OF MODEL PERFORMANCE

This section reviews a variety of measures that assess the goodness of a model with respect to its ability to discriminate between defaulting and non-defaulting firms (its power), as well as the goodness of a model with respect to predicting the actual probabilities of default accurately (its calibration).

We begin by looking at simple tools for examining power. These are useful for providing an intuitive foundation but tend to be limited in the information they provide. We go on to show how these measures can be extended into more general measures of model power.

We next look at tools for comparing the calibration of competing models in terms of a model's ability to accurately predict default rates. This is done by comparing conditional predictions of default (model predicted probabilities) to actual default outcomes. We begin with a simple example that compares the predicted average default rates of two models given a test data set, and then extend this to a more general case of the analysis of individual predicted probabilities, based on firm-specific input data.

We also discuss the potential limitations of power and calibration tests.

2.1 Ranking and power analysis

2.1.1 Contingency tables: simple measures of model power

Perhaps the most basic tool for understanding the performance of a default prediction model is the "percentage right". A more formal way of understanding this measure is to consider the number of predicted defaults (non-defaults) and compare this to the actual number of defaults (non-defaults) experienced. A common means of representing this is a simple *contingency table* or *confusion matrix*.

	Actual default	**Actual non-default**
Bad	*TP*	*FP*
Good	*FN*	*TN*

In the simplest case, the model produces only two ratings (Bad/Good). These are typically shown along the y-axis of the table with the actual outcomes (default/non-default) along the x-axis. The cells in the table indicate the number

of true positives (*TP*), true negatives (*TN*), false positives (*FP*) and false negatives (*FN*), respectively. A *TP* is a predicted default that actually occurs; a *TN* is a predicted non-default that actually occurs (the company does not default). This *FP* is a predicted default that does not occur and a *FN* is a predicted non-default where the company actually defaults. The errors of the model are *FN* and *FP* shown on the off diagonal. *FN* represents a Type I error and *FP* represents a Type II error. A "perfect" model would have zeros for both the *FN* and *FP* cells, and the total number of defaults and non-defaults in the *TP* and *TN* cells, respectively, indicating that it perfectly discriminated between the defaulters and non-defaulters.

There have been a number of proposed metrics for summarizing contingency tables using a single quantity that can be used as indices of model performance. A common metric is to evaluate the true positive rate as a percentage of the *TP* and *FN*, $TP/(TP + FN)$, although many others can be used.[2] It turns out that in many cases the entire table can be derived from such statistics through algebraic manipulation, due to the complimentary nature of the cells.

Note that in the case of default models that produce more than two ratings or that produce continuous outputs, such as probabilities, a particular contingency table is only valid for a specific model *cutoff point*. For example, a bank might have a model that produces scores from 1 to 10. The bank might decide that it will only underwrite loans to firms with model scores better than 5 on the 1 to 10 scale. In this case, the *TP* cell would represent the number of defaulters whose internal ratings were worse than 5 and the *FP* would represent the number of non-defaulters worse than 5. Similarly, the *FN* would be all defaulters with internal ratings better than 5 and the *TN* would be the non-defaulters better than 5.

It is the case that for many models, different cutoffs will imply different relative performances. Cutoff "*a*" might result in a contingency table that favors model A, while cutoff "*b*" might favor model B, and so on. Thus, using contingency tables, or indices derived from them, can be challenging due to the relatively arbitrary nature of cutoff definition. This also makes it difficult to assess the relative performance of two models when a user is interested not only in strict cutoffs, but in relative ranking, for example when evaluating a portfolio of credits or a universe of investment opportunities.

2.1.2 CAP plots, ROC curves and power statistics: more general measures of predictive power

ROC (*relative* or *receiver operating characteristic*) curves (Green and Swets (1966); Hanley (1989); Pepe (2002); and Swets (1988; 1996)), generalize contingency table analysis by providing information on the performance of a model at *any* cutoff that might be chosen. They plot the *FP* rate against the *TP* rate for all credits in a portfolio.

[2]For example, Swets (1996) lists 10 of these and provides an analysis of their correspondence to each other.

ROCs are constructed by scoring all credits and ordering the *non-defaulters* from worst to best on the x-axis and then plotting the *percentage of defaults excluded* at each level on the y-axis. So the y-axis is formed by associating every score on the x-axis with the cumulative percentage of defaults with a score equal to or worse than that score in the test data. In other words, the y-axis gives the percentage of defaults excluded as a function of the number of non-defaults excluded.

A similar measure, a CAP (Cumulative Accuracy Profile) plot (Sobehart *et al* 2000), is constructed by plotting *all* of the test data from "worst" to "best" on the x-axis. Thus, a CAP plot provides information on the percentage of defaulters that are excluded from a sample (*TP* rate), given that we exclude all credits, Good and Bad, below a certain score.

CAP plots and ROC curves convey the same information in slightly different ways. This is because they are geared to answering slightly different questions.

CAP plots answer the question[3]

> *"How much of an* entire portfolio *would a model have to exclude to avoid a specific percentage of defaulters?"*

ROC curves use the same information to answer the question

> *"What percentage of* non-defaulters *would a model have to exclude to exclude a specific percentage of defaulters?"*

The first question tends to be of more interest to business people, while the second is somewhat more useful for an analysis of error rates.[4] In cases where default rates are low (ie, 1–2%), the difference can be slight and it can be convenient to favor one or the other in different contexts. The Type I and Type II error rates for the two are related through an identity involving the sample average default probability and sample size (see Appendix A).

Although CAP plots are the representation typically used in practice by finance professionals, more has been written on ROC curves, primarily from the statistics and medical research communities. As a result, for the remainder of this section, we will refer to ROC curves, because they represent the same information as CAPs but are more consistent with the existing literature.

ROC curves generalize the contingency table representation of model performance across *all possible* cutoff points (see Figure 1).

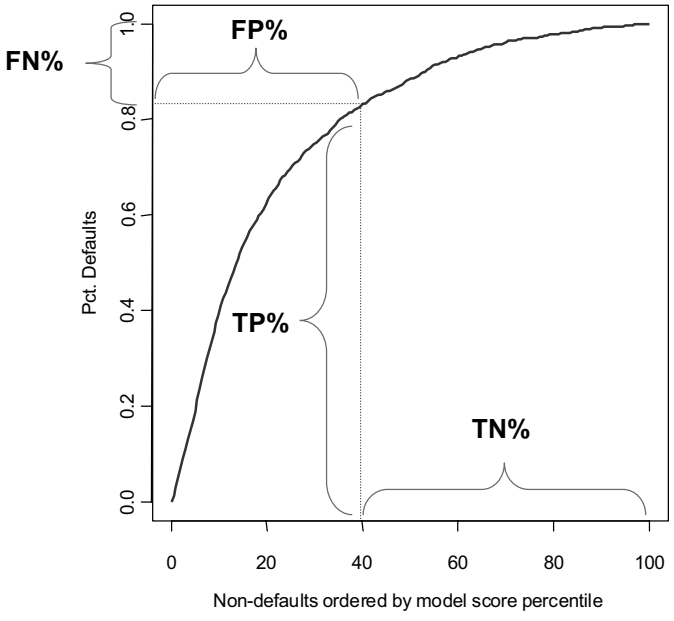
With knowledge of the sample default rate and sample size, the cells of the table can be filled in directly. If there are *ND* non-defaulting companies in the data set and *D* defaulting companies then the cells are given as follows.[5]

[3] In statistical terms, the CAP curve represents the cumulative probability distribution of default events for different percentiles of the risk score scale.

[4] CAP plots can also be more easily used for direct calibration by taking the marginal, rather than cumulative, distribution of defaults and adjusting for the true prior probability of default.

[5] Note that D can be trivially calculated as rP, where r is the sample default rate and P is the number of observations in the sample. Similarly ND can be calculated as $P(1-r)$.

FIGURE 1 Schematic of a ROC showing how all four quantities of a contingency table can be identified on a ROC curve. Each region of the x- and y-axes have an interpretation with respect to error and success rates for defaulting (*FP* and *TP*) and non-defaulting (*FN* and *TN*) firms.



	Actual default	Actual non-default
Bad	$TP\%*D$	$FP\%*ND$
Good	$FN\%*D$	$TN\%*ND$

Every cutoff point on the ROC curve gives a measure of Type I and Type II error as shown above. In addition, the slope of the ROC curve at each point on the curve is also a *likelihood ratio* of the probability of default to non-default for a specific model's score (Green and Swets (1966)).

Importantly, for validation in cases where the ROC curve of one model is strictly dominated by the ROC of a second (the second lies above the first at all points), the second model will have unambiguously lower error for any cutoff.[6]

[6]While often used in practice, cutoff criteria are most appropriate for relatively simple independent decisions. The determination of appropriate cutoff points can be based on business constraints (eg, there are only enough analysts to follow $x\%$ of the total universe), but these are typically sub-optimal from a profit maximization perspective. A more rigorous criterion can be derived with knowledge of the prior probabilities and the cost function, defined in terms of the costs (and benefits) of *FN* and *FP* (and *TN* and *TP*). For example, the point at which the line with slope S, defined below, forms a tangent to the ROC for a model, defines the optimal

A convenient measure for summarizing the ROC curve is the *area under the ROC* (A), which is calculated as the integral of the ROC curve: the proportion of the area below the ROC relative to the total area of the unit square. A value of 0.5 indicates a random model, and a value of 1.0 indicates perfect discrimination. A similar measure, the Accuracy Ratio (AR), can also be calculated for CAP plots (Sobehart *et al* (2000)).[7]

If the ROC curves for two models cross, neither dominates the other in all cases. In this situation, the ROC curve with the highest value of A may not be preferred for a specific application defined by a particular cutoff. Two ROC curves may have the same value of A, but have very different shapes. Depending on the application, even though the area under the curve may be the same for two models, one model may be favored over the other.

For example, Figure 2 shows the ROC curves for two models. The ROC curves each produce the same value for A. However, they have different characteristics. In this example, model users interested in identifying defaults among the worst quality firms (according to the model) might favor model A because it offers better discrimination in this region, while those interested in better differentiation among medium and high-quality firms might choose model B, which has better discrimination among these firms. For some types of applications it may be possible to use the two models whose ROC curves cross to achieve higher power than either might independently (Provost and Fawcett (2001)), although such approaches may not be suitable for credit problems.[8]

The quantity A also has a convenient interpretation. It is equivalent to *the probability that a randomly chosen defaulting loan will be ranked worse than a randomly chosen non-defaulting loan* by a specific model (Green and Swets (1966)). This is a useful quantity that is equivalent to a version of the Wilcoxon (Mann–Whitney) statistic (see, for example, Hanley and McNeil, (1982)).

cutoff (Swets (1996)). In this case, S is defined as follows:

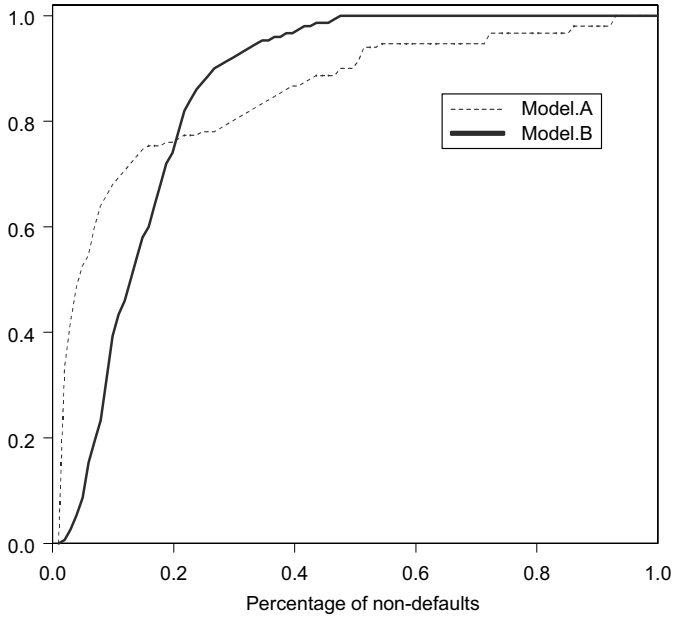
$$S = \frac{\mathrm{d}\,\mathrm{ROC}(k)}{\mathrm{d}k} = \frac{p(ND)}{p(D)} \frac{[c(FP) + b(TN)]}{[c(FN) + b(TP)]}$$

where $b(\cdot)$ and $c(\cdot)$ are the benefit and cost functions, $p(\cdot)$ is the unconditional probability of an event, k is the cutoff and D and ND are defined as before. Stein (2005) develops this approach further and eliminates the need for a single cutoff by calculating appropriate fees for any loan ranked along the ROC curve.

[7]Engelmann *et al* (2003) also provide an identity relating the area under the curve (A) to the accuracy ratio (AR): $AR = 2(A - 0.5)$ (see also Appendix A). Alternative versions of accuracy ratios have also been developed to tailor the analysis to problem-specific features that sometimes arise in practice (cf Cantor and Mann (2003)).

[8]The approach proposed by Provost and Fawcett (2001) requires one to choose probabilistically between two models for a particular loan application, where the probabilities are proportional to the location of an optimal cutoff on the tangent to the two ROC curves of the models. Although this strategy can be shown to achieve higher overall power than either model would on its own, the choosing process may be inappropriate for credit because the same loan could get different scores if run through the process twice. For applications involving direct marketing, fraud detection, etc, this is not generally an issue; however, in credit-related applications the drawbacks of this non-determinism may be sufficient to outweigh the benefits of added power.

FIGURE 2 Two ROCs with the same area but different shapes. The area under the ROC (A) for each is the same, but the shapes of the ROCs are different. Users interested in identifying defaults among the worst firms might favor model A, while those interested in better differentiation among medium and high-quality firms might choose model B.



ROC curves provide a good deal of information regarding the ability of a credit model to distinguish between defaulting and non-defaulting firms. The meaning of an ROC curve (or a CAP plot) is intuitive and the associated statistic A (or AR) has a direct interpretation with respect to a model's discriminatory power. For credit model evaluation, where we are typically most concerned with a model's ability to rank firms correctly, an ROC curve or CAP plot is more useful than simple contingency tables because the graphical measures avoid the need to define strict cutoffs and provide more specific information about the regions in which model power is highest. As such, they provide much more general measures of model power.

It is helpful to think not only in statistical terms, but also in economic terms. Stein (2005) demonstrated that the same machinery to that used to determine optimal cutoffs in ROC analysis (see footnote 6) can also be used to determine the fees that should be charged for risky loans *at any point on the* ROC curve, implying that no loans need be refused. In this way, the ROC directly relates to loan pricing. Furthermore, this relationship naturally provides a mechanism to determine the economic value of a more powerful model. Stein and Jordão (2003) provided a methodology for evaluating this through data-based simulation

and showed that for a typical bank the value of even a somewhat more powerful model can be millions of dollars. Blöchlinger and Leippold (2006) provided an alternative approach that does not require data (given a number of simplifying assumptions).

ROC curves and CAP plots and their related diagnostics are power statistics and, as such, do not provide information on the appropriateness of the *levels* of predicted probabilities. For example, using ROC analysis, model scores do not need to be stated as probabilities of default, but could be letter grades, points on a scale of 1–25, and so on. Scores from 1 to 7 or 10 are common in internal bank rating systems. Models that produce good rankings ordinarily will be selected by power tests over those that have poorer power even though the *meaning* of, say, a "6" may not be well defined in terms of probabilities.

Powerful models, without probabilistic interpretations, however, cannot be used in certain risk management applications such as CVAR and capital allocation because these applications require probabilities of default. Fortunately, it is usually possible to calibrate a model to historical default probabilities. In the next section we examine measures of the accuracy of such calibration.

2.2 Likelihood-based measures of calibration

As used here, likelihood measures[9] provide evidence of the plausibility of a particular model's probability estimates given a set of empirical data. A likelihood estimate can be calculated for each model in a set of candidate models and will be highest for the model in which the predicted probabilities match most closely with the actual observed data.

The likelihood paradigm can be used to evaluate which hypothesis (or model) among several has the highest support from the data. The agreement of predicted probabilities with subsequent credit events is important for risk management, pricing and other financial applications, and as a result, modelers often spend a fair amount of time *calibrating* models to true probabilities.

Calibration typically involves two steps. The first requires mapping a model score to an empirical probability of default using historical data. For example, the model outputs might be bucketed by score and the default rates for each score calculated using an historical database. The second step entails adjusting for the difference between the default rate in the historical database and the actual default rate.[10] For example, if the default rate in the database were 2%, but the

[9] A detailed discussion of likelihood approaches is beyond the scope of this paper. For an introduction, see Reid (2002), which provides a brief discussion, Edwards (1992), which provides more insight and some historical background, or (Royall (1997)), which provides a more technical introduction along with examples and a discussion of numerical issues. Friedman and Cangemi (2001) discuss these measures in the context of credit model validation.

[10] It is not uncommon for these rates to be different due to data gathering and data processing constraints, particularly when dealing with middle market data from banks data in which financial statements and loan performance data are stored in different systems. In technical terms, it is necessary to adjust the model prediction for the true prior probability.

actual default rate were known (through some external source) to be 4%, it would be necessary to adjust the predicted default rates to reflect the true default rate.[11]

Note that tests of calibration are tests of *levels* rather than tests of power and as such can be affected if the data used contains highly correlated default events or if the data represent only a portion of an economic cycle. This issue can arise in other contexts as well. Although we do not discuss it in detail in this article, it is important to consider the degree to which these may have impact on results (cf, Stein (2006)).

Likelihood measures are most useful in the context of evaluating the calibration of two competing models on a new data set. As a simple example of using likelihood to evaluate two models, consider evaluating two models using a portfolio with 100 loans in it, four of which have defaulted.

For exposition, we address the simpler problem of assessing the accuracy of the *average* predicted default rate with respect to the actual average observed default rate. In this case our "model" is just a prediction of the mean default rate. We wish to determine which of two proposed average probabilities of default, 1% or 5%, is more consistent with the observed default behavior in the portfolio. The likelihoods, based on a binomial assumption, are given in the following table.[12]

Proposed mean *PD*	**Likelihood of proposed mean *PD***
$\mu = 0.01$	$p(\mu = 0.01) = \binom{100}{4} (0.01^4)(0.99^{96}) = 0.0149$
$\mu = 0.05$	$p(\mu = 0.05) = \binom{100}{4} (0.05^4)(0.95^{96}) = 0.1781$

Using the likelihood paradigm, we would seek the model with the maximum likelihood given the actual data. In other words, we seek the model whose probability of default predictions are most consistent with the empirical data.

The default rate of 5% has a higher likelihood, given the data used for testing, than the default rate of 1% and, thus, the model proposing $\mu = 0.05$ would be favored over that proposing $\mu = 0.01$, by virtue of this higher likelihood ($0.1781 > 0.0149$).

In this simple case, most analysts would be able to perform this evaluation by simply choosing the default rate closest to the actual observed rate. However, in most real-world settings, models' estimates of probabilities of default are not

[11] An earlier draft of this paper noted simple approximation to this adjustment (true baseline/sample baseline) which works reasonably well for low probabilities. A more exact, although less intuitive, adjustment is possible as well (cf, Bohn *et al* 2007).

[12] Recall that the probability associated with a binomial distribution is given as $b(k; n, p) = \binom{n}{k} p^k [(1-p)^{n-k}]$, where p is the probability of an event, k is the number of default events observed and n is the total number of firms. Note that in the context of model validation, the constant (combinatorial) term can be dropped if convenient as it does not depend on the model and all models are evaluated on the same data.

constant as in the example above, but are conditional on a vector of input variables, x, that describes the company. Appendix B provides a review of the mathematics that extend the simple case to the more general case.

For technical reasons, it is convenient to work with the log of the likelihood, $\ell(\cdot)$, which is a monotonic transformation of the likelihood and thus the largest value of the likelihood will be associated with the model that produces the largest value for ℓ (model). (Note that this amounts to the value with the *least negative* number.) The model with the highest log likelihood would be the best calibrated. Importantly, note that in this context, the likelihood measure is calculated given only a vector of model outputs and a vector of default outcomes. It is *not* the likelihood that may have been calculated during the model estimation process and it does not admit information outside of the model predictions and actual outcomes. This formulation allows it to accommodate out-of-sample testing.

While calibration can be challenging, *it is generally far easier to calibrate a powerful model to true default rates than it is to make a weak but well calibrated model more powerful*.

This fact highlights a drawback of using likelihood measures alone for validation. The focus on calibration exclusively can lead likelihood methods to incorrectly reject powerful, but imperfectly calibrated models in favor of weaker but better calibrated models. This is true even when the poorly calibrated model can be "fixed" with a simple adjustment.

To see this, consider an example of two models W and P that are being applied to an identical portfolio of 11,000 loans with a true default rate of 5%. Assume that neither model was developed using the test set, so the test is an "out-of-sample" test, such as might be performed by a bank evaluating two credit models on its own portfolio. The models themselves are very simple, with only two ratings: Good and Bad. The models have been calibrated so each rating has an associated probability of default.

The following table gives the log likelihoods for these two example models under this data set. Using the likelihood criterion, we would choose model W over model P because a difference of 132 in log likelihood is large.[13]

Model	Log likelihood
W	−2,184
P	−2,316
(W) − (P)	132

Now consider how the models that produced these likelihoods performed. The following tables show the number of defaults predicted by each model and the predicted probabilities under each model for each class, based on the stated calibration of the two models.

[13] See Appendix B for details.

Model	Model *PD* prediction		Actual outcome: Default	Actual outcome: Non-default
W	5.10%	Bad	50	950
	4.90%	Good	500	9,500
P	1.50%	Bad	549	1
	0.01%	Good	1	10,449

Recall that the likelihood approach in this context is being used to determine the effectiveness of model calibration. Now note that the preferred (under the likelihood selection paradigm) model W (the weaker model) does a far worse job separating defaulting firms from healthy ones. It turns out that model W's actual performance is *no different than a random model* in that both Good and Bad ratings have a 5% probability of default on the test set, and this is the same as the actual central tendency of the test set. Thus, it does not matter which rating the model gives, the actual default rate will be the same: the mean for the population ($50/1{,}000 = 500/10{,}000 = 5\%$).

In contrast, model P (the more powerful model) demonstrates very good power, discriminating almost perfectly between defaulters and non-defaulters. Most lenders, if asked to choose, would select model P because, using this out-of-sample test data set, model P gives high confidence that the bank will identify future defaulters.

Why is model W selected over model P under the likelihood paradigm? It is because its probabilities more closely match the observed probabilities of the data set. Thus, even though it is a random model that results in the same realized probability for both Good and Bad classes, this probability is very close to what is actually observed on average, and thus the model probabilities are more likely, given the data, than those of model P.

On the other hand, model P was miscalibrated and therefore its very powerful predictions do not yield the number of defaulters in each class that would be expected by its predicted probabilities of 1.5% and 0.01%. It is not as well calibrated to its actual performance.

Now suppose that the modeler learns that the true prior probabilities of the sample were different than the probabilities in the real population and adjusts the model output by a (naive) constant factor of two. In this case, the log likelihood of model P would change to $-1{,}936$ and, after this adjustment, model P would be preferred. Note that the model is still badly miscalibrated, but even so, under the likelihood paradigm, it would now be preferred over model W. Thus, a simple (constant) adjustment to the prior probability results in the more powerful model being better calibrated than model W.

It is reasonable to ask whether there is a similar simple adjustment we can make to model W to make it more powerful. Unfortunately, there is generally no straightforward way to do this without introducing new or different variables into the model or changing the model structure in some other way.

Interestingly, if an analyst was evaluating models using tests of model power rather than calibration, almost any model would have been chosen over model W.

To see this, consider that model W makes the same prediction for every credit and thus provides no ranking beyond a random one. Thus, model W would be identical to the random model in a power test. Even a very weak model that still provided slight discrimination would be chosen over model W if the criterion were power.

Note that likelihood measures are designed for making relative comparisons between competing models, not for evaluating whether a specific model is "close" to being correctly calibrated or not. So the best model of a series of candidate models might still be poorly calibrated (just less so than the other models). Owing to this, many analysts find it useful to perform other types of tests of calibration as well. For example, it is often useful to bin a test data set by predicted probabilities of default and to then calculate the average probability of default for each bin to determine whether the predicted default rate is reasonably close to the observed default rate. In this case, the analyst can get more direct intuition about the agreement and direction of default probabilities than might be available from the likelihood paradigm, but may have a less precise quantitative interpretation of these results. Both analyses can be useful.

As a whole, the likelihood measures we have discussed here focus on the agreement of predicted probabilities with actual observed probabilities, not on a model's ability to discriminate between Goods and Bads. In contrast, a CAP plot or an ROC curve measures the ability of a model to discriminate between Goods and Bads, but not to accurately produce probabilities of default.

If the goal is to have the most accurate probability estimate, irrespective of the ability to discriminate between Good and Bad credits, the maximum likelihood paradigm will always provide an optimal model selection criterion. However, there is no guarantee that the model selected by the likelihood approach will be the most powerful or, as the example shows, even moderately powerful.

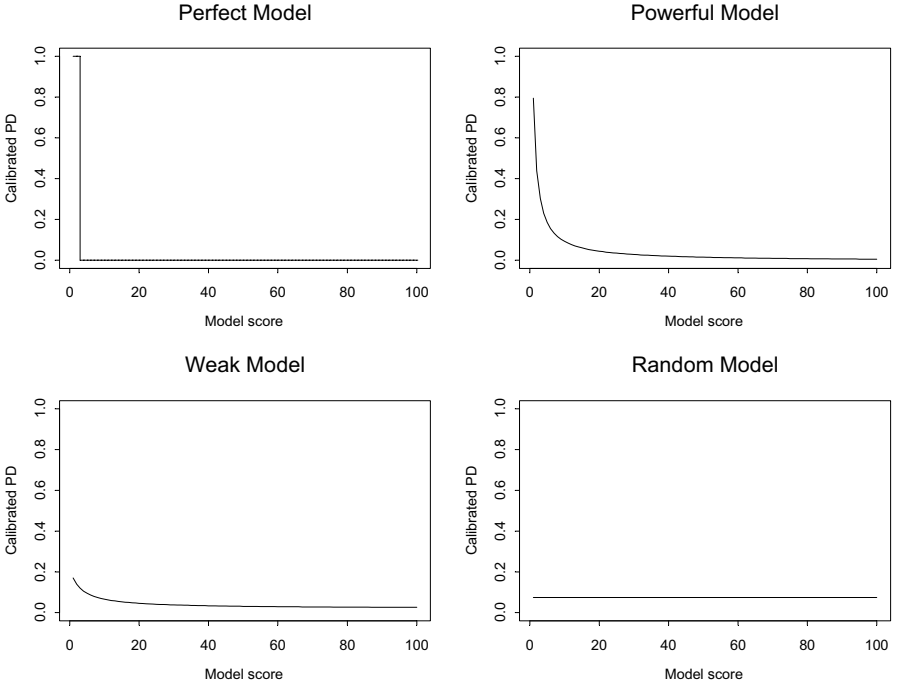
In contrast, if the goal is to determine which model discriminates best between defaulting and non-defaulting firms, tools such as ROC curves CAP plots provide a well-established means for doing so. In this case, there are no guarantees on the appropriateness of the probability estimates produced by the model, but we do get an unambiguous measure of the model's power.[14]

2.3 The relationship between power and calibration

Importantly, although they measure different things, power and calibration are related: the power of a model is a limiting factor on how high a resolution may be achieved through calibration, even when the calibration is done appropriately. To see this, consider the following example of four models each with different power: a perfect default predictor, a random default predictor, a powerful, but not perfect

[14]Both calibration and power analysis can be extended to more involved contexts. For example, likelihood measures can be used to evaluate long run profitability using the Kelly criteria (Kelly (1956)). While exceptions do exist (eg, Bell and Cover (1980)), in general, these criteria are based on atypical assumptions of a series of sequential single-period decisions rather than the existence of a portfolio. It is more common for an investor to own a portfolio of assets (eg, loans) and ask whether adding another is desirable.

FIGURE 3 Calibration curves for four hypothetical models with different power. The more powerful a model is, the better the resolution it can achieve in probability estimation. Each of these models is perfectly calibrated, given the data, but they are able to achieve very different levels because of their differing power. The random model can never generate very high or very low probabilities while the perfect model can generate probabilities that range from zero to one (although none in between). The more powerful model can give broad ranges, while the weaker model can only give narrower ranges.



model and a weak, but not random one. Assume that the data used for calibration is a representative sample of the true population.

Figure 3 shows the calibration curve for these four hypothetical models. In this example, we discuss only on the calibration of the very worst score (far left values in each graph) that the model produces because the other scores follow the same pattern.

How does the calibration look for the perfect model? In the case of the perfect model, the default probability for the worst score will be 1.00 since the model segregates perfectly the defaulting and non-defaulting firms, assigning bad scores to the defaulters. On the other hand, how does this calibration look for the random model? In this case, calibrated default probability is equal to the mean default rate for the population because each score will contain a random sample of defaults and non-defaults. The other two models lie somewhere between these extremes.

For the powerful model, the calibration can reach levels close to the perfect model, but because it does not perfectly segregate all of the defaults in the lowest score, there are also some non-defaulters assigned the worst score. As a result, even if the model is perfectly calibrated to the data (ie, the calibrated probabilities for each bucket exactly match those in the data set), it cannot achieve probabilities of 1.00. Similarly, the weak model performs better than the random model, but because it is not very powerful, it gives a relatively flat curve that has a much smaller range of values than the powerful model.

This implies that a more powerful model will be able to generate probabilities that are more accurate than a weaker model, even if the two are calibrated perfectly on the same data set. This is because the more powerful model will generate higher probabilities for the defaulting firms and lower probabilities for non-defaulting firms due to its ability to discriminate better between the two groups and thus concentrate more of the defaulters (non-defaulters) in the Bad (Good) scores.

There is no simple adjustment to the calibration of these models to improve the accuracy of the probability estimates because they are calibrated as well as possible, given their power. This does not mean that the probabilities of the weaker and more powerful model are equally accurate, only that they cannot be improved beyond their accuracy level through simple calibration unless the power is also improved.[15]

3 HOW TO MEASURE: A WALK-FORWARD APPROACH TO MODEL TESTING

Performance statistics for credit risk models are typically very sensitive to the data sample used for validation. We have found that tests are most reliable when models are developed and validated using some type of out-of-sample and out-of-time testing.[16]

Indeed, many models are developed and tested using some form of "hold-out testing" which can range from simple approaches such as saving some fraction of the data (a "hold-out sample") for testing after the model is fit, to more sophisticated cross-validation approaches. However, with time varying processes such as credit,[17] hold out testing can miss important model weaknesses not obvious when fitting the model across time because simple hold-out tests do not provide information on performance through time.

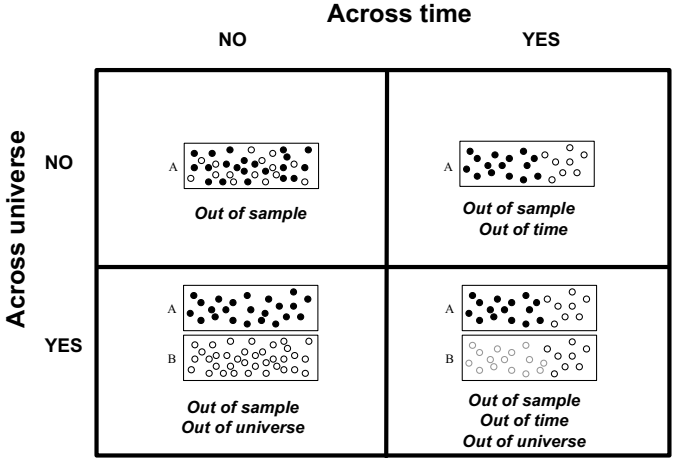
In the following section, we describe a validation framework that accounts for variations across time and across the population of obligors. It can provide

[15]Note that the relationship between power and calibration is direct provided that the calibration method involves monotonic transformations of the model output. However, calibration techniques that change the ordering of observations (cf, Dwyer and Stein (2005)) do not maintain the power–calibration relationship described above and can also improve model power.

[16]Out-of-sample refers to observations for firms that are not included in the sample used to build the model. Out-of-time refers to observations that are not contemporanious with the training sample.

[17]See, for example, Mensah (1984) or Gupton and Stein (2002).

FIGURE 4 Schematic of out-of-sample validation techniques. Testing strategies are split based on whether they account for variances across time (horizontal axis) and across the data universe (vertical axis). Dark circles represent training data and white circles represent testing data. Gray circles represent data that may or may not be used for testing. (Adapted from Dhar and Stein (1998).)



important information about the performance of a model across a range of economic environments.[18]

This approach is most useful in validating models during development and less easily applied where a third-party model is being evaluated or when data is more limited. That said, the technique is extremely useful for ensuring that a model has not been overfit.

3.1 Forms of out-of-sample testing

A schematic of the various forms of out-of-sample testing is shown in Figure 4. The figure splits the model testing procedure along two dimensions: (a) time (along the horizontal axis) and (b) the population of obligors (along the vertical axis). The least restrictive out-of-sample validation procedure is represented by the upper-left quadrant and the most stringent by the lower-right quadrant. The other two quadrants represent procedures that are more stringent with respect to one dimension than the other.

The upper-left quadrant describes the approach in which the testing data for model validation is chosen completely at random from the full model fitting data set. This approach to model validation assumes that the properties of the data remain stable over time (stationary process). As the data is drawn at random, this

[18]The presentation in this section follows closely that of Dhar and Stein (1998), Stein (1999) and Sobehart *et al* (2000).

approach validates the model across the population of obligors but does not test for variability across time.

The upper-right quadrant describes one of the most common testing procedures. In this case, data for model fitting is chosen from any time period prior to a certain date and testing data is selected from time periods only after that date. As model validation is performed with out-of-time samples, the testing assumptions are less restrictive than in the previous case and time dependence can be detected using different validation sub-samples. Here it is assumed that the characteristics of firms do not vary across the population.

The lower-left quadrant represents the case in which the data is segmented into two sets containing no firms in common, one set for building the model and the other for validation. In this general situation the testing set is out-of-sample. The assumption of this procedure is that the relevant characteristics of the population do not vary with time but that they may vary across the companies in the portfolio.

Finally, the most robust procedure is shown in the lower-right quadrant and should be the preferred sampling method for credit risk models. In addition to being segmented in time, the data is also segmented across the population of obligors. Non-overlapping sets can be selected according to the peculiarities of the population of obligors and their importance (out-of-sample and out-of-time sampling).

Out-of-sample out-of-time testing is beneficial because it prevents overfitting of the development data set, but also prevents information about future states of the world that would not have been available when developing the model from being included. For example, default models built before a market crisis may or may not have predicted default well during and after the crisis, but this cannot be tested if the data used to build the model was drawn from periods before and after the crisis. Rather, such testing can only be done if the model was developed using data prior to the crisis and tested on data from subsequent periods.

As default events are rare, it is often impractical to create a model using one data set and then test it on a separate "hold-out" data set composed of completely independent data. While such out-of-sample and out-of-time tests would unquestionably be the best way to compare models' performance if default data was widely available, it is rarely possible in practice. As a result, most institutions face the following dilemma:

- *If too many defaulters are left out of the in-sample data set, estimation of the model parameters will be seriously impaired and overfitting becomes likely.*
- *If too many defaulters are left out of the hold-out sample, it becomes exceedingly difficult to evaluate the true model performance due to severe reductions in statistical power.*

3.2 The walk-forward approach

In light of these problems, an effective approach is to "rationalize" the default experience of the sample at hand by combining out-of-time and out-of-sample tests. A testing approach that focuses on this last quadrant and is designed to test

models in a realistic setting, emulating closely the manner in which the models are used in practice, is often referred to in the trading model literature as "walk-forward" testing.

It is important to make clear that there is a difference between walk-forward testing and other more common econometric tests of stationary or goodness of fit. For example, when testing for goodness of fit of a linear model in economics, it is common to look at an R^2 statistic. For more general models, a likelihood related (eg, Akaike's information criterion (AIC), etc) measure might be used. When testing for the stationarity of an economic process, researchers will often use a Chow test.

The key difference between these statistics and tests and statistics derived from walk-forward tests is that statistics such as R^2 and AIC and tests such as Chow tests are all *in-sample* measures. They are designed to test the agreement of the parameters of the model or the errors of the model with the data used to fit the model during some time period(s). In contrast, walk-forward testing provides a framework for generating statistics that allow researchers to test the *predictive power* of a model on data not used to fit it.

The walk-forward procedure works as follows:

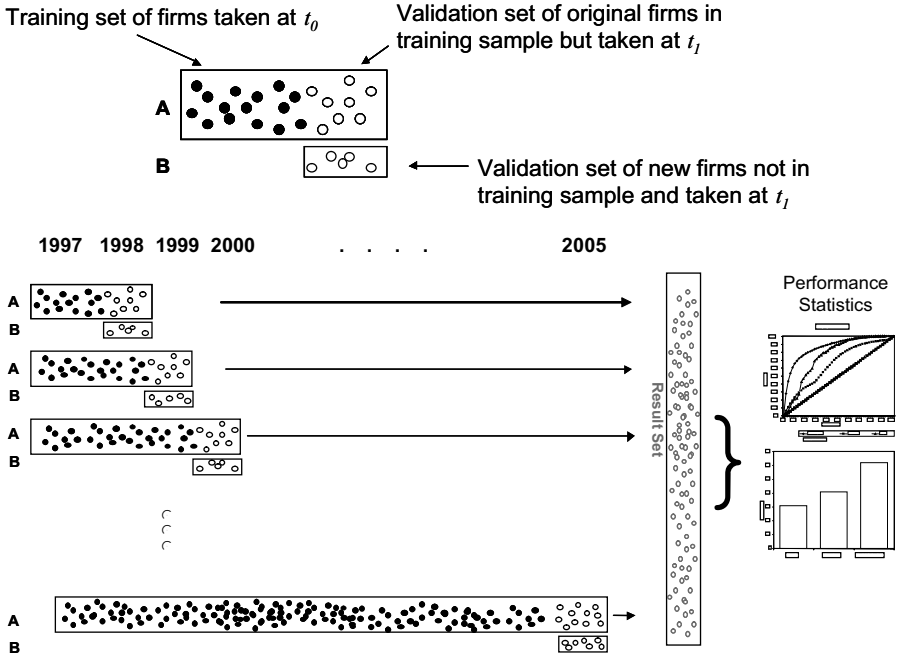
(1) Select a year, for example 1997.
(2) Fit the model using all the data available on or before the selected year.
(3) Once the model's form and parameters are established for the selected time period, generate the model outputs for all of the firms available during the following year (in this example, 1998).
(4) Save the prediction as part of a result set.
(5) Now move the window up (eg, to 1998) so that all of the data through that year can be used for fitting and the data for the next year can be used for testing.
(6) Repeat steps (2) to (5) adding the new predictions to the result set for every year.

Collecting all of the out-of-sample and out-of-time model predictions produces a set of model performances. This *result set* can then be used to analyze the performance of the model in more detail. Note that this approach simulates, as closely as possible given the limitations of the data, the process by which the model will actually be used. Each year, the model is refit and used to predict the credit quality of firms one year hence. The process is outlined in the lower left of Figure 5.

In the example below, we used a one-year window. In practice the window length is often longer and may be determined by a number of factors including data density and the likely update frequency of the model itself, once it is online.

The walk-forward approach has two significant benefits. First, it gives a realistic view of how a particular model would perform over time. Second, it gives analysts the ability to leverage to a higher degree the availability of data for validating models. In fact, the validation methodology not only tests a

FIGURE 5 Schematic of the walk-forward testing approach. In walk-forward testing, a model is fit using a sample of historical data on firms and tested using both data on those firms one year later and data on new firms one year later (upper portion of exhibit). Dark circles represent in-sample data and white circles represent testing data. This approach results in "walk-forward testing" (bottom left) when it is repeated in each year of the data by fitting the parameters of a model using data through a particular year, and testing on data from the following year, and then moving the process forward one year. The results of the testing for each validation year are aggregated and then may be resampled (lower left) to calculate particular statistics of interest.



particular model, but it tests the entire modeling approach. As models are typically reparameterized periodically as new data come in and as the economy changes, it is important to understand how the approach of fitting a model, say once a year, and using it in real-time for the subsequent year, will perform. By employing walk-forward testing, analysts can get a clearer picture of how the entire modeling approach will hold up through various economic cycles.

Two issues can complicate the application of the walk-forward approach. The first is the misapplication of the technique through the repeated use of the walk-forward approach while *developing* the model (as opposed to testing a single final model). In the case where the same "out-of-sample" data is used repeatedly to garner feedback on the form of a candidate model as it is being developed, the principle of "out-of-time" is being violated.

The second complication can arise when testing models that have continuously evolved over the test period. For example, banks often adjust internal rating models to improve them continuously as they use them. As a result it is often impossible to recreate the model, as it would have existed at various points in the historical test period, so it can be difficult to compare the model with others on the same test data set. In such situations, it is sometimes feasible to test the model only in a period of time after the last change was made. Such situations are reflected in the upper-right quadrant of Figure 4. This approach tends to limit the number of defaults available and as a result it can be difficult to draw strong conclusions (see Section 4).

In light of the complications that can result in limitations on test samples, it is of interest to develop techniques for interpreting the results of validation exercises. In the next section we discuss some of these.

4 HOW TO INTERPRET PERFORMANCE: SAMPLE SIZE, SAMPLE SELECTION AND MODEL RISK

Once a result set of the type discussed in Section 3 is produced and performance measures as described in Section 2 are calculated, the question remains of how to *interpret* these performance measures. Performance measures are sensitive to the data set used to test a model, and it is important to consider how they are influenced by the characteristics of the particular data set used to evaluate the model.

For problems involving "rare" events (such as credit default), it is the *number of occurrences of the rare event* more than the total number of observations that tends to drive the stability of performance measures. For example, if the probability of default in a population is of the order of 2% and we test a model using a sample of 1,000 firms, only about 20 defaults will be in the data set. In general, had the model been tested using a different sample that included a different 20 defaults, or had the model used only 15 of the 20 defaults, we could often observe quite different results.

In addition to the general variability in a particular data set, a sample drawn from one *universe* (eg, public companies) may give quite different results than would have been observed on a sample drawn from a different universe (eg, private companies). Recall again the discussion of Figure 4.

In most cases, it is impossible to know how a model will perform on a different sample than the one at hand (if another sample were available, that sample would be used as data for testing as well). The best an analyst can do is size the magnitude of the variability that arises due to sampling effects.

We next focus on two types of uncertainty that can arise in tests using empirical data.

- There is a natural variability in any sample chosen from the universe of defaulting and non-defaulting firms. Depending on the number of defaulting and non-defaulting firms, this variability can be relatively small or quite large. Understanding this variability is essential to determining whether

differences in the performance of two models are spurious or significant. One approach to sizing this variability is through the use of *resampling* techniques, which are discussed, along with examples, in Section 4.1. As expected, the smaller the number of defaults, the higher the variability.

- There can be an artificial variability that is created when samples are drawn from different populations or universes, as in the case when a model is tested on a sample drawn from a population that is different than the population that the model will be applied to. For example, a bank may wish to evaluate a private firm default model but only have access to default data for public companies. In this case, analysts will have less certainty and it will be harder to size the variably of testing results. We briefly discuss this type of variability in Section 4.2 where we try to give some indication for the differences in model performance when samples are drawn from different populations. We provide an example of the performance of the same model on samples drawn from different populations.

It is often the case that competing models must be evaluated using only the data that an institution has on hand and this data may be limited. This limitation can make it difficult to differentiate between good and mediocre models. The challenge faced by many institutions is to understand these limitations and to use this understanding to design tests that lead to informed decisions.

4.1 Sample size and selection confidence

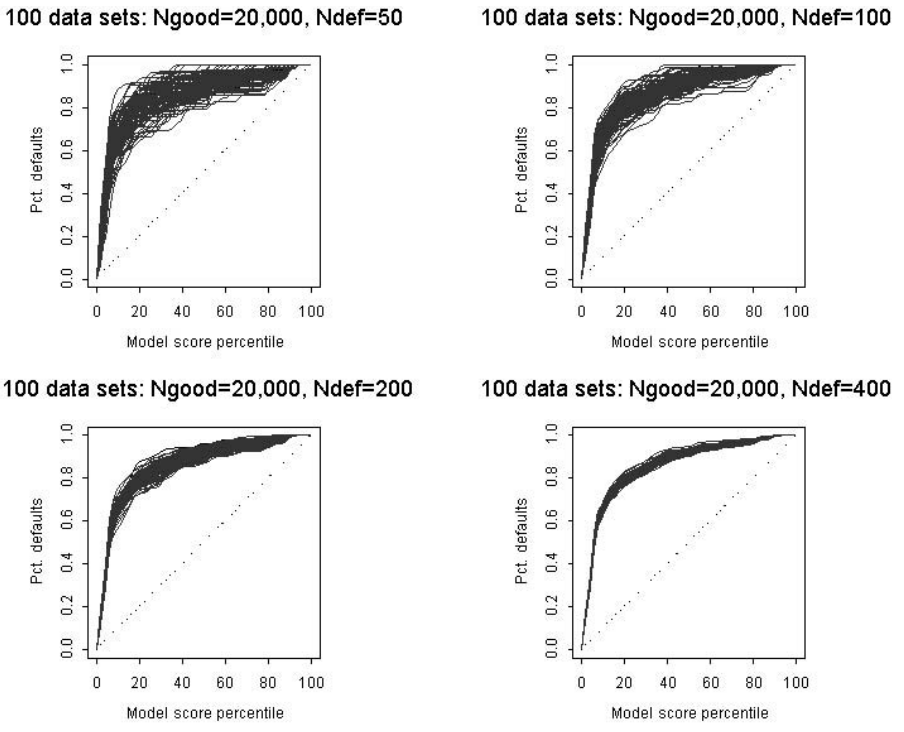
Many analysts are surprised to learn of the high degree of variability in test outcomes that can result from the composition of a particular test sample.

To demonstrate this more concretely we present the results of evaluating a simple model using randomly selected samples of 50, 100, 200 and 400 defaults, respectively. The samples are drawn (without replacement) from a large database of defaulting and non-defaulting firms. We do this 100 times for each sample size. These results are shown graphically in Figure 6. In the figure, for each sample a separate CAP plot is graphed. Thus, each figure shows 100 CAP pots, each for the same model, but each of which is based on a different sample of 20,000 non-defaulting observations and 50, 100, 200 and 400 defaulting observations.

In examining the figures, note the (very) high degree of variability present. This variability decreases dramatically as the number of defaults is increased and thus is primarily driven by the *number of defaults*, rather than the total *number of observations*. Despite the large differences in variability in the graphs, the sample size for the first and last set of tests (upper-left and lower-right graphs) differ by less than 2%. Note also the wide variability of results, even at relatively high numbers of defaults.

To show that it is the number of defaults that drives the variability, we present the CAP plots of for another group of data sets. These show the results of evaluating the same model using randomly selected samples of 10,000, 15,000, 20,000 and 50,000 non-defaulting financial statements but keeping the number of defaults constant at 100. We do this 100 times for each sample size.

FIGURE 6 Variability of test results when different numbers of defaults are used for evaluation. These graphs show the results of evaluating a simple model using randomly selected samples of varying numbers of defaults. Each figure shows 100 CAP pots, each of which is based on evaluating the same model using a different sample of 20,000 non-defaulting observations and 50, 100, 200 and 400 defaulting observations.

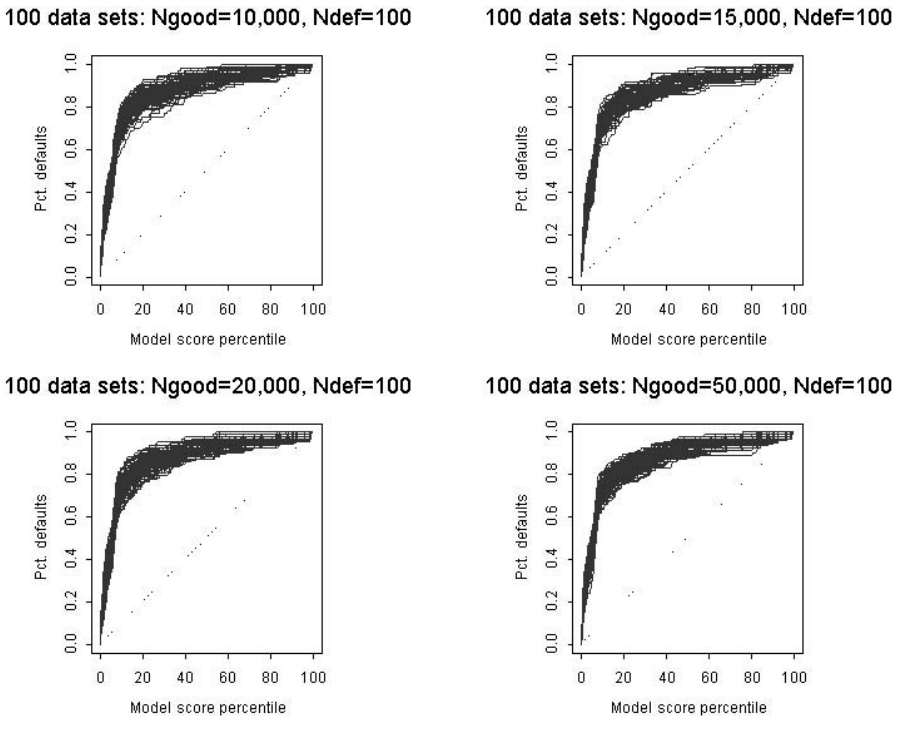


These results are shown graphically in Figure 7 similar to the presentation in Figure 6. In this case, it is clear that the effect increasing the number of non-defaulting records, even by fivefold, does not materially affect the variability of the results. This outcome stands in contrast to that shown in Figure 6 and supports the observation that the key factor in reducing the variability of test results is the number of defaults used to test a model.

We can make this intuition more precise. For example, we calculated an accuracy ratio for each of the 100 CAP plots in each graph in Figure 7. For each graph we then calculated the standard deviation (and interquartile ranges) of the accuracy ratios. The SD (and IQR) for the four cases differed only at the third decimal place. Further, the differences showed no consistent pattern and statistical tests showed no significant difference in distribution.

This is not unexpected. Most of the closed-form solutions that have been suggested for the variance of the area under the ROC curve involve terms that

FIGURE 7 Variability of test results when different numbers of non-defaults are used for evaluation. These graphs show the results of evaluating the same model using randomly selected samples of 10,000, 15,000, 20,000 and 50,000 non-defaulting financial statements but keeping the number of defaults constant at 100. It is clear that the effect increasing the number of non-defaulting records does not appear to affect the variability of the results greatly.



scale as the inverse of the product of the number of defaults and non-defaults. The intuition here is that the addition of another defaulting firm to a sample reduces the variance by increasing the denominator by the (much larger) number of good records, while the addition of a non-defaulting firm only increases the denominator by the (much smaller) number of bad firms.

In the extreme, Bamber (1975) describes an analytic proof (attributed to Van Danzig (1951)) that, in the most general case, the maximum variance of the area under the curve is given by

$$A(1-A)/D$$

assuming there are fewer defaults than non-defaults, where A is the area under the curve and D is the number of defaults. Another formulation, with stricter assumptions, gives the maximum variance as

$$(1/3ND)[(2N+1)A(1-A)-(N-D)(1-A)^2]$$

where the notation is as above and N is the number of non-default records.

From inspection, the maximum variance in the first case is inversely proportional to number of defaults only, and in the second case it is inversely proportional to the product of the number of defaults and non-defaults, due to the first term, so when the number of non-defaults is large, it decreases much more quickly as the number of defaults is increased.

Note that it is convenient to speak in terms of the maximum variance here because there are many possible model–data combinations that yield a specific A, each producing its own unique ROC and each with its own variance. The maximum variance is the upper bound on these variances for a given A. Upper bounds for the variance in special cases (those with different assumptions) can be derived as well. However, it is generally difficult to determine which assumptions are met in realistic settings and therefore closed-form approaches to defining confidence intervals for A based on either variance or maximum variance derivations can lead to over- or underestimates. We discuss this below.

If the number of defaults was greater relative to the number of non-defaults, the relationship would be reversed, with non-defaults influencing the variance more dramatically. Similarly, if the number of defaults and non-defaults were about even, they would influence the result to approximately the same degree. In general, it is the *minority class* (the class with fewer observations) of a sample that will influence the variance most dramatically.

Most institutions face the challenge of testing models without access to a large data set of defaults. Given such a situation, there is relatively little that can be done to *decrease* the variability of the test results that they can expect from testing. A far more reasonable goal is to simply understand the variability in the samples and use this knowledge to inform the interpretation of any results they do obtain to determine whether these are consistent with their expectations or with other reported results.

Sometimes we wish to examine not just the area under the curve, but a variety of metrics and a variety of potential hypotheses about these statistics. To do this, we often need a more general approach to quantifying the variability of a particular test statistic. A common approach to sizing the variability of a particular statistic given an empirical sample is to use one of a variety of *resampling* techniques to leverage the available data and reduce the dependency on the particular sample.[19]

A typical resampling technique proceeds as follows. From the result set, a sub-sample is selected at random. The performance measure of interest (eg, A or the number of correct predictions for a specified cutoff, or other performance statistics of interest) is calculated for this sub-sample and recorded. Another sub-sample is then drawn and the process is repeated. This continues for many repetitions until a distribution of the performance measure is established. The sampling distribution is used to calculate statistics of interest (standard error, percentiles of the distribution, etc).

[19]The bootstrap (eg, Efron and Tibshirani (1993)), randomization testing and cross-validation (eg, Sprent (1998)) are all examples of resampling tests.

Under some fairly general assumptions, these techniques can recover estimates of the underlying variability of the overall population from the variability of the sample.

For example, Figure 8 shows the results of two separate experiments. In the first, we took 1,000 samples from a data set of defaulted and non-defaulted firms. We structured each set to contain 20,000 non-defaulted observations and 100 defaulted observations. Conceptually, we tested the model using 1,000 different data sets to assess the distribution of accuracy ratios. This distribution is shown in the top graph of Figure 8. In the second experiment, we took a single sample of 100 defaults and 20,000 non-defaults and bootstrapped this data set (sampled with replacement from the single data set) 1,000 times and again plotted the distribution of accuracy ratios. The results are shown in the bottom graph (adjusted to a mean of zero to facilitate direct comparison[20]).

Of interest here is whether we are able to approximate the variability of the 1,000 different data sets from the full universe by looking at the variability of a single data set from that universe, resampled 1,000 times. In this case, the results are encouraging. It turns out that the quantiles of the distribution of the bootstrap sample slightly overestimate the variability,[21] but give a fairly good approximation to the variability.

This is useful because in most settings we would only have the bottom frame of this figure and would be trying to approximate the distribution in the top frame. For example, given the estimated variability of this particular sample, we can determine that the 90% confidence interval spans about 14 points of *AR*, which is very wide.

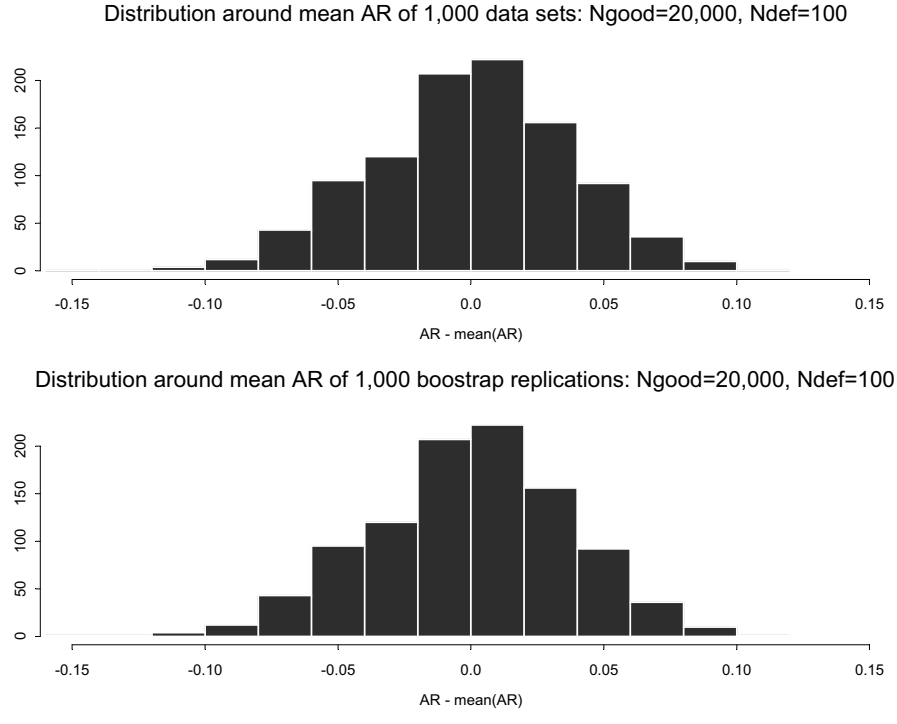
Resampling approaches provide two related benefits. First, they give a more accurate estimate of the variability around the actual reported model performance. Second, because of typically low numbers of defaults, resampling approaches decrease the likelihood that individual defaults (or non-defaults) will overly influence a particular model's chances of being ranked higher or lower than another model. For example, if model A and model B were otherwise identical in performance, but model B *by chance* predicted a default where none actually occurred on company XYZ, we might be tempted to consider model B as inferior to model A. However, a resampling technique would likely show that the models were really not significantly different in performance, given the sample.

This example worked out particularly well because the sample we chose to bootstrap was drawn at random from the full population. Importantly, in cases

[20]We need to do this since the mean of the bootstrap sample will approach the value of the AR for the original single sample, but the mean of the 1,000 individual samples drawn from the full population will approach the mean of the overall population. Due to the sampling variability these may not be the same.

[21]See Efron and Tibshirani (1993) for a discussion of adjustments to non-parametric and parametric estimates of confidence intervals that correct for sample bias, etc. Note that under most analytic formulations, the variance of the estimate of the area under the ROC is itself affected by the value of the area. As a result, some of the discrepancy here may also be due to the difference in the level of the *AR*, which is related to the area under the curve.

FIGURE 8 Two similar distributions of *AR*s: 1,000 samples and a bootstrap distribution for a single sample. The top frame shows the distribution of accuracy ratios for 1,000 samples from a data set of defaulted and non-defaulted firms. Each set contains 20,000 non-defaulted observations and 100 defaulted observations. The bottom frame shows the distribution of accuracy ratios for 1,000 bootstrap replications of a single sample of 100 defaults and 20,000 non-defaults. Results are adjusted to a mean of zero to facilitate direct comparison.



where a particular sample may not be drawn at random from the population, where there may be some selection bias, resampling techniques are less able to recover the true population variability. In such cases, resampling can still be used to size the variability of a set of models' performance on that sample, and as such can be useful in comparing models, but the results of that testing cannot easily be extended to the overall population.

For comparisons between competing models, more involved resampling techniques can be used that take advantage of matched samples. In such a setting, it is more natural to look at the distribution of differences between models on matched bootstrap replications.

For example, in comparing the performance of two models we are typically interested in which model performs better along some dimension. One way to understand this is to resample the data many times and, for each sample, calculate

the statistic for each model. The result will be two sets of matched statistics, one set containing all of the statistics for the first model on each resampled data set, and one set containing all of the statistics for the second model on each set. The question then becomes whether one model consistently outperforms the other on most samples.

Bamber (1975) also provides a (semi)closed-form solution for determining the variance of A, although this relies on assumptions of asymptotic normality. Bamber's estimator derives from the correspondence between the area under the ROC curve, A, and the Mann–Whitney statistic. A confidence bound for the Mann–Whitney statistic can often be calculated directly using standard statistical software.

Engelmann *et al* (2003) discuss this approach and test the validity of the assumption of asymptotic normality, particularly as relates to smaller samples. The sample of defaulted firms in their experiments lead them to conclude that the normal approximation is generally not too misleading, particularly for large (default) samples. That said, they present evidence that as the number of defaults decreases, the approximations become less reliable. For (very) low numbers of defaults, confidence bounds can differ by several percentage points, however, such differences can be economically significant. As shown in Stein and Jordão (2003) a single percentage point of difference between two models can represent an average of 0.97 bps or 2.25 bps additional profit per dollar loaned depending on whether a bank was using a cutoff pricing approach, respectively. In 2002, for a medium sized US bank these translated into additional profit of about $2 million and $5 million, respectively.

While smaller samples can lead to unstable estimates of model performance when using closed form approaches, Engleman *et al* (2003) show that the closed form results variance estimation approaches appear reasonably reliable.

It should be noted that a great advantage of Bamber's (1975) closed-form confidence bound is the relatively lower computational cost associated with calculating confidence intervals based on this measure which enables fast estimation. One reasonable strategy is thus to use the approximations during early work in developing and testing models, but to confirm results using bootstrap in the later phases. To the extent that these results differ greatly, more extensive analysis of the sample would be required.

4.2 Performance levels for different populations

In the previous section, we discussed ways in which resampling approaches can provide some indication of the sensitivity of the models being tested to the particular sample at hand, and thus provide information that is valuable for comparing models. However, this approach does not provide detailed information about how the models would perform on *very different* populations.

In this section, we give a brief example of how the performance observed on one set of data can be different to that observed on another set of data, if the data sets are drawn from fundamentally different populations. The goal of this section

is not to present strong results about the different universes that might be sampled, but rather to highlight the potential difficulties in drawing inferences about one population based on testing done on another.

The example we choose comes up often in practice. We show that performance statistics can vary widely when a model is tested on rated public companies unrated public companies and private companies. In such cases, an analyst might wish to understand a default model's performance on a bank portfolio of middle-market loans, but incorrectly test that model on a sample of rated firms. Inferences about the future performance of the model on middle-market companies, based on an analysis of rated firms, could result in misleading conclusions.[22]

Earlier research has provided evidence that models *fit* on public or rated companies may be misspecified if applied to private firms (eg, Falkenstein *et al* (2000)). A similar phenomenon can be observed in test results if models are *tested* on a different population than that on which they will be applied.

As an example, consider the CAP plot shown in Figure 9. The figure provides evidence of differences in performance statistics as the same model is applied to rated, unrated and private firms, respectively.

This example serves to highlight the degree to which testing on the appropriate population can affect decisions about models. It gives an indication of the sensitivity of models to different populations, particularly if the data are drawn from a population that is not representative of the population to which it will be applied.

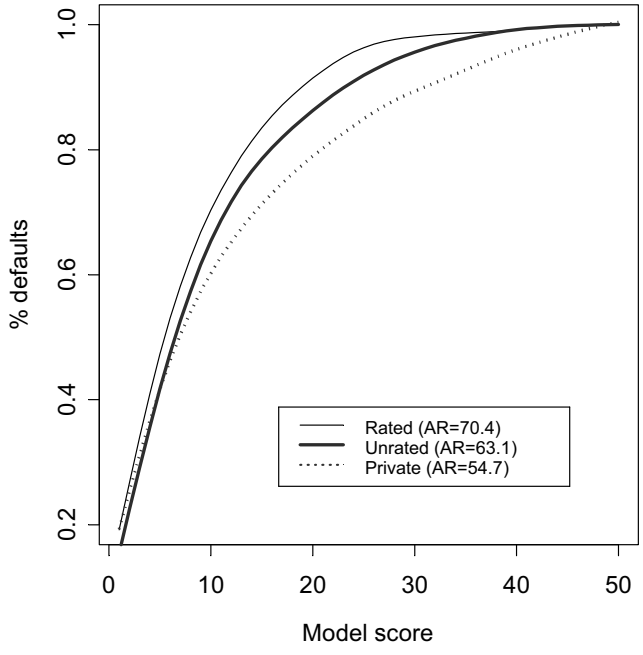
From this figure, we can see that the performance statistics generated from tests run on one population are clearly different than those that were generated from tests run on a different population.

What this implies is that it is very difficult to assess the performance of a model on, say, private companies by testing on publicly rated firms. It would be even more difficult to compare two models intended for use on private firms if the test results were generated using publicly rated firms. The reasons for these differences could be manifold, from the design of the model to the nature of the data used to develop it, and to the nature of the data used to test it. For these reasons, tests must be designed carefully. Given the differences in sample composition and sample size, it is clearly inappropriate to compare these statistics directly with each other. In fact, it is not even clear how one would directly compare the statistics on these three samples.

For example, users of private firm models sometimes ask whether the model might perform well on public firms. This question can be answered, but it is only

[22] Note that this common-sense principle also shows up in some surprising settings, for example, in the rules of evidence that govern US court proceedings. In its landmark 1997 ruling, *General Electric versus Robert K Joiner*, the United States Supreme Court upheld an appellate court ruling that the court may exclude scientific evidence when there is simply too great an analytic gap between the data and the opinion it is used to support. The Court ruled that, "The studies [offered as evidence] were so dissimilar to the facts presented. . . that it was not an abuse of discretion for the District Court to have rejected the experts reliance on them." (Opinion of Judge Renquist, Supreme Court of the United States (1997)).

FIGURE 9 CAP plots of the same model applied to rated, unrated publicly traded and private firms.



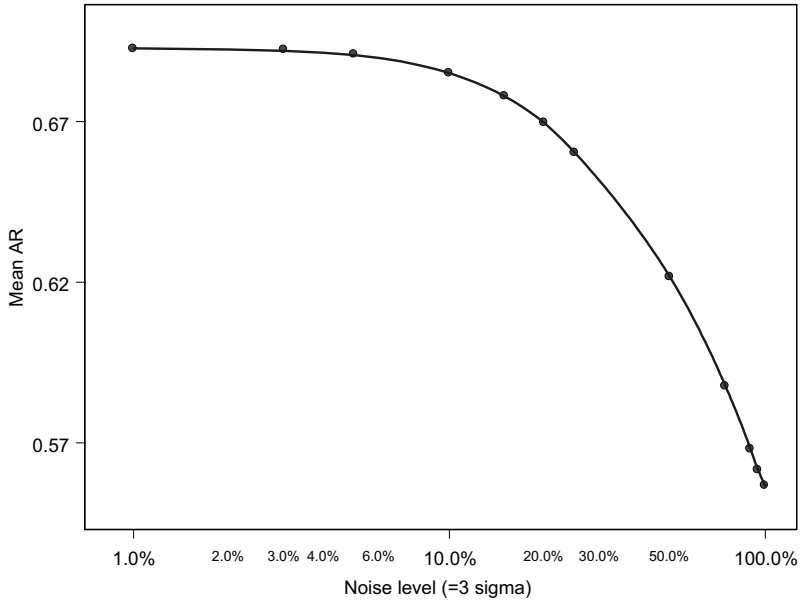
meaningful to ask how the model performs relative to other public firm alternatives on public firms, not relative to the model's performance on private firms, because no direct comparison can be made in this case. It can be particularly misleading to draw inferences about the performance of a model on one population if it is tested on a sample drawn from different model.

4.3 Extensions

Validation testing need not be limited to tests of model performance alone. Consider the evaluation of a model for use within a bank located where accounting data is known to be of poor quality. Validation testing can be used to evaluate the impact of this noisy accounting data on model performance by observing the degradation in model performance as noise is added to test data and it is run through the model.

Figure 10, shows the results of adding random noise at various levels to accounting data and running that data through a default model that was designed to be tolerant to noisy data. The noise levels indicate the distribution of the noise, in multiplicative terms, that is used to shock the accounting data and are plotted against the average accuracy ratio for the simulations at that noise level. So, for

FIGURE 10 The impact of noise levels in accounting data on accuracy ratios for a default model. The results of adding uncorrelated random errors to accounting data and running that data through a default model that was designed to be tolerant to noisy data are shown. As the noise levels increase the model performance degrades, but it does so gradually. The noise levels on the x-axis indicate the distribution of the noise, in multiplicative terms, that is used to shock the accounting data and are plotted against the average accuracy ratio for the simulations at that noise level.



example, a noise level of 30% indicates that about 99% of the time an accounting variable will be randomly increased or decreased by between 0 and 30%.

Figure 10 shows that for this particular model, the average performance degrades as the noise levels increase, but does this fairly gradually. The noise levels can be fairly high before the average model's performance begins to degrade significantly. In a fuller experiment, this would give some confidence in the robustness of the model and could provide some guidelines for its usage.

In a different study, Stein *et al* (2003), used validation testing approaches to determine which risk factors, systematic or idiosyncratic, better explain private firm defaults. Through a series of validation experiments, they test the predictive power of two different modeling approaches – one that relies primarily on systematic factors and one that relies primarily on idiosyncratic factors. By segmenting the population along dimensions that isolate these factors, they conclude that the majority, but not all, of the risk associated with private firms is attributable to idiosyncratic risk components.

5 CONCLUSION

In this article we have discussed several strategies for testing model performance. We can test for both power and calibration and these two properties are related. As a result, we find it useful to first identify the most powerful model available and then to calibrate this powerful model. While this calibration may be flawed, by choosing a powerful model we increase the probability that the model chosen, when calibrated, will have acceptable performance. While it is never possible to calibrate a model to *future* default experience, we can attempt to simulate this through appropriately designed experiments.

We have discussed a framework called walk-forward testing that allows developers to test models and modeling methodologies while controlling for sample and time dependence. Where its use is feasible, the technique reduces the chances that models will be overfit because it never uses data in the testing that was used to fit model parameters. At the same time, this approach allows modelers to take full advantage of the data by using as much of the data as possible to fit and to test the models.

The results of model testing, whether using walk-forward tests on a newly developed model or whether testing a third-party model for use within a financial institution, are subject to sample variability. This variability is driven more by the number of defaults than by the total number of observations in a data set. Small numbers of defaults lead to very high variability results.

In general, it is not possible to reduce this uncertainty, so often the best we can do is size and understand it. We have suggested strategies, based on both analytic and resampling techniques, that permit the recovery of information about sample variability from the sample of data at hand. We have also provided a suggestive example of cases in which a model could have very different performance on a test data set than on the population to which it was being applied if that population was structurally different from the test set. We have shown an example of extending the application of validation work beyond model comparison, as in the case of testing for model robustness in the presence of noise.

Taken as a whole, this suggests a strategy of:

- testing models against reasonable benchmarks to understand better the relative scale of performance statistics on the data set on hand;
- taking care to ensure that the population from which the test sample is drawn is similar in its characteristics to the population on which it will be applied;
- evaluating the variability of performance results through appropriate resampling experiments; and
- evaluating models first with respect to power and, upon selecting the most powerful model, testing the calibration and perhaps recalibrating.

We feel that it is of fundamental importance to understand the behavior of the metrics being calculated, to recognize the limitations of the data used to test them and to design experiments that provide good measures of the performance

of default models along the dimensions that are important for the business and research applications at hand.

In this paper, we have attempted to bring together material from a number of disparate research streams and to discuss aspects of each that can have an impact on the interpretation of benchmarking results. We have found that the approaches we describe here, while not appropriate for all models, form an effective suite of tools for benchmarking internal and external credit models where data permits and provide guidance on interpreting results where data is sparse.

At the time that the first draft of this paper was circulated in 2002, formal model validation research was still evolving. However, as evidenced by the additional references and topics discussed in this updated version of the article, the past five years have enjoyed increased interest both among academics and practitioners on the issues of validation and on the limitations of various validation approaches. These results, in addition to being of keen interest intellectually, ultimately benefit practitioners directly. They serve to highlight the importance of conducting careful experiments on credit model performance and, most importantly, the economic consequences of failing to do so.

APPENDIX A TYPE I AND TYPE II ERROR: CONVERTING CAP PLOTS INTO CONTINGENCY TABLES

A common tool for evaluating default prediction models is the use of power curves and CAP plots which graphically show the power of various models on a set of test data. These diagnostics are often summarized though power statistics and Accuracy Ratios.

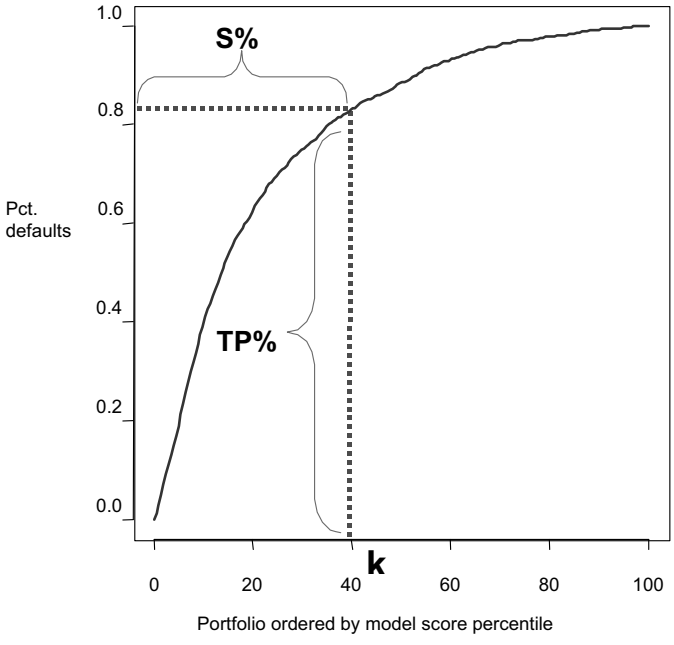
However, for some purposes, analysts and researchers are interested in understanding Type I and Type II error rates for specific cutoffs. With knowledge of the sample default rates, these error rates are fully described by the CAP plots. In this appendix, we describe a methodology for inferring Type I and Type II error rates from these diagnostic plots.

For reference, we repeat that the CAP plot is constructed similarly to the ROC curve. Like a ROC curve, the CAP plot describes the percentage of defaults excluded by a particular model for a given cutoff criteria on the y-axis. Unlike ROC curves, CAP plots display the percentage of the entire sample excluded by a cutoff criteria on the x-axis (as opposed to only showing the non-defaulting credits on the x-axis, as in the ROC curve).

Thus, the x and y coordinates display the percentage of the sample excluded and the true positive rate respectively, as shown in Figure A.1.

With knowledge of the sample default rate, there is a straightforward transformation that can be used to construct a table of Type I and Type II errors (and false negative and false positive results) from the CAP plot for any cutoff criteria of interest. This, in turn, permits an analyst to determine what percentage of a portfolio would need to be excluded in order to capture a desired percentage of defaults or, conversely, determine the percentage of defaults that would be excluded for a given exclusion of a portfolio.

FIGURE A.1 A schematic representation of a CAP plot.



To derive the transformation, first note that the true positive and false positive rates are defined directly from the y-axis of the CAP plot. The true negative and true positive rates are not directly observable from the x-axis, owing to the fact that the x-axis of a CAP plot contains both non-defaulted and defaulted firms: the entire sample ($S\%(k)$).

As we are interested in percentages in terms of the total number of non-defaults, rather than the percentage of all firms in the sample, we need to back out the defaulted firms. To do this, we need to know how many of them are contained in the region between zero and the cutoff, k.

This is just the true positive rate times the total default rate, $r\%$, since $TP\%(k)$ of the total defaults were excluded and $r\%$ of the total sample represents the total percentage of the sample that defaulted. This quantity,

$$S\%(k) - (TP\%(k) * r\%)$$

gives the total proportion of the entire sample that were non-defaulters, below the cutoff.

Finally, the above expression is still given in units of the total sample, rather than in units of the non-defaulters only. For true and false negatives, because we are interested in the proportion of the non-defaulters that are excluded for a particular cutoff criterion, rather than the proportion of the total sample, we need

to normalize the expression above to put it in terms of total non-defaults only. To do this, we adjust the above percentage so that it is a percentage not of the total sample, but of the total sample minus the defaulters, or $1 - r\%$. This yields the following expression for false positives:

$$FP\% = \frac{S\%(k) - (TP\%(k) * r\%)}{1 - r\%}$$

The remaining terms can be calculated directly.

A contingency table, incorporating the above transformation, is shown in the following table.

	Actual default	Actual non-default
Predicted default	$TP\%(k)$	$\frac{S\%(k) - (TP\%(k)^* r\%)}{1 - r\%}$
Predicted non-default	$1 - TP\%(k)$	$1 - \frac{S\%(k) - (TP\%(k)^* r\%)}{1 - r\%}$

In this table, k is a specific cutoff criterion, $TP\%(k)$ is the true positive rate (y-axis: percentage defaults captured) associated with criterion k, $S\%(k)$ is the percentage of the full sample (x-axis: percentage of sample excluded) with criterion k and $r\%$ is the sample default rate.

APPENDIX B THE LIKELIHOOD FOR THE GENERAL CASE OF A DEFAULT MODEL

In the case of a model predicting a binary event (default/no default), the model's estimate of the probability of a single event y happening given data x is

$$\text{prob}(y \mid x) = p(x)^y [1 - p(x)]^{(1-y)}$$

where $p(x)$ is the probability (of default) predicted by the model, conditional on the input x, and the event y is defined as follows:

$$y = \begin{cases} 1 & \text{if the firm defaults} \\ 0 & \text{if the firm remains solvent} \end{cases}$$

Note that as either y or $(1 - y)$ will always be 0, this reduces to the simple predicted probability of the outcome, according to the model, conditioned on x.

For a given set of data, the *likelihood* of the model, $\mathcal{L}(\text{model})$, is calculated by computing the model's predicted probabilities, given the data inputs to the model, and calculating the appropriateness of these predictions for each observation, given the *actual* outcomes. Here we assume that the model is evaluated using a specific data set and that the likelihood is thus with respect to this data set.

To do this we generalize the above probability by taking advantage of the fact that the overall likelihood of a model for a pooled data set is the product of the

individual likelihoods of any disjoint subsets of the data:

$$\mathcal{L}(\text{model}) = \prod_{i=1}^{n} \text{prob}(y_i \mid x_i) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{(1-y_i)}$$

In general, it is easier to work with summations than products, particularly for more complicated likelihood functions, so by convention we work with the log of the likelihood, $\ell(\text{model})$. This is also convenient computationally because for large datasets and small probabilities, the likelihoods can become tiny in their raw form causing underflows. The log likelihood is[23]

$$\ell(\text{model}) = \sum_{1}^{n} y_i \ln[p(x_i)] + (1 - y_i) \ln[1 - p(x_i)]$$

REFERENCES

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12**, 381–417.

Bell, R. M., and Cover, T. M. (1980). Competitive optimality of logarithmic investment. *Mathematics of Operations Research* **5**(2), 161–166.

Bohn, J. R., Crosbie, P., and Stein, R. M. (2007). *Active Credit Portfolio Management in Practice*. Wiley, New York (forthcoming).

Blöchlinger, A., and Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking and Finance* **30**(3), 851–873.

Burnham, K. P., and Anderson, D. R. (1998). *Model Selection and Inference*. Springer: New York.

Cantor, R., and Mann, C. (2003). Measuring the performance of Moody's corporate bond ratings. Moody's Special Comment, Moody's Investors Service, New York.

Dhar, V., and Stein, R. (1998). Finding robust and usable models with data mining: examples from finance. *PCAI*, December.

Dwyer, D., and Stein, R. (2005). Moody's KMV RiskCalc$^{\text{TM}}$ 3.1. Technical Document, Moody's KMV.

Edwards, A. W. F. (1992). *Likelihood*. Johns Hopkins University Press, London.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability, No. 57)*. Chapman and Hall/CRC Press.

Engelmann, B., Hayden, E., and Tasche, D. (2003). Testing rating accuracy. *RISK*, January.

Falkenstein, E., Boral, A., and Carty, L. V. (2000). RiskCalc for private companies: Moody's default model. Special Comment, Moody's Investors Service.

Friedman, C., and Cangemi, R. (2001). Selecting and evaluating credit risk models. *ERisk iConference series*.

[23]Note that because y_i is always either zero or one, this can also be written as $\ln[y_i\, p(x_i) + (1 - y_i)(1 - p(x_i))]$ which is computationally slightly more convenient.

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Peninsula Publishing, Los Altos, CA.

Greene, W. (2000). *Econometric Analysis*. Prentice Hall, Upper Saddle River, NJ.

Gupton, G. M., and Stein, R. M. (2002). LossCalc(TM): Moody's model for predicting Loss Given Default (LGD). *Moody's Rating Methodology*. Moody's Investors Service, New York.

Hanley, A., and McNeil, B. (1982). The meaning and use of the area under a Receiver Operating Characteristics (ROC) curve. *Diagnostic Radiology*, **143**(1), 29–36.

Hanley, J. A. (1989). Receiver Operating Characteristic (ROC) methodology: the state of the art. *Critical Reviews in Diagnostic Imaging* **29**(3), 307–335.

Kelly, J. L. Jr. (1956). A new interpretation of information rate. *The Bell System Technical Journal* **35**, 185–189.

Mensah, Y. M. (1984). An examination of the stationarity of multivariate bankruptcy prediction models: a methodological study. *Journal of Accounting Research* **22**(1), 380–395.

Pepe, M. S. (2002). Receiver operating characteristic methodology. *Statistics in the 21st Century*, Raftery, E. A., Tanner, M. A. and Wells, M. T. (eds). Chapman and Hall/CRC Press.

Provost, F., and Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning* **42**, 203–231.

Reid, N. (2002). Likelihood. *Statistics in the 21st Century*, Raftery, E. A., Tanner, M. A., and Wells, M. T. (eds). Chapman and Hall/CRC Press.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm (Monographs on Statistics and Applied Probability, Vol. 71)*. Chapman & Hall/CRC Press.

Sobehart, J. R., Keenan, S. C., and Stein, R. M. (2000). Benchmarking quantitative default risk models: a validation methodology. Moody's Special Comment, Moody's Investors Service.

Sprent, P. (1998). *Data Driven Statistical Methods*. Chapman & Hall, London.

Stein, R. M. (1999). An almost assumption free methodology for evaluating financial trading models using large scale simulation with applications to risk control. Working Paper, Stern School of Business, New York University, New York.

Stein, R. M. (2002). Benchmarking default prediction models: pitfalls and remedies in model validation. Technical Report #020305, Moody's KMV, New York.

Stein, R. M., Kocagil, A. E., Bohn, J., and Akhavein, J. (2003). Systematic and idiosyncratic risk in middle-market default prediction: a study of the performance of the RiskCalc™ and PFM™ models. Technical Report, Moody's KMV, New York.

Stein, R. M., and Jordão, F. (2003). What is a more powerful model worth? Technical Report #030124, Moody's KMV, New York.

Stein, R. M. (2005). The relationship between default prediction and lending profits: integrating ROC analysis and loan pricing. *Journal of Banking and Finance* **29**, 1213–1236.

Stein, R. M. (2006). Are the probabilities right? Dependent defaults and the number of observations required to test for default rate accuracy. *Journal of Investment Management* **4**(2), 61–71.

Supreme Court of the United States (1997). General Electric Company, *et al*, Robert K. Joiner *et ux*.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science* **240**, 1285–1293.

Swets, J. A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics*. Laurence Erlbaum Associates, Mahwah, NJ.

Van Danzig, D. (1951). On the consistency and power of Wilcoxon's two sample test. *Koninklijike Nederlanse Akademie van Wetenschappen, Proceedings, Series A* **54**.