Moody's Investors Service
Global Credit Research

March 2000

Rating Methodology

Contact	Phone
New York	
Jorge R. Sobehart	1.212.553.1653
Sean C. Keenan	
Roger M. Stein	

RATING METHODOLOGY

Benchmarking Quantitative Default Risk Models: A Validation Methodology

continued on page 3

Authors

Jorge R. Sobehart
Sean Keenan
Roger Stein

Production Associate

Don Linares

Overview

Many of the world's largest financial institutions have developed advanced quantitative credit risk models that help to measure, monitor and manage credit risk across their business lines. However, the Basel Committee on Banking Supervision recently identified credit model validation as one of the most challenging issues in quantitative credit model development[1]. In particular, issues of data sufficiency and model sensitivity analysis were highlighted as was the lack of a consistent and formalized *validation* methodology in many institutions.

Because of Moody's leading role in credit risk assessment, Moody's has also been active in developing and testing quantitative methods that can be used for credit risk management. This article presents a summary of the approach Moody's used to validate and benchmark a series of popular quantitative default risk models, including our own Public Firm model[2]. We discuss performance measurement and sampling techniques, as well as other practical considerations associated with performance evaluation for quantitative credit risk models. This framework specifically addresses issues of data sparseness and the sensitivity of models to changing economic conditions. Our model validation approach continues to evolve and is used extensively for evaluating internal and external quantitative models.

In summary:

1. We describe some of the techniques used at Moody's to benchmark the performance of a number of corporate default prediction models. This approach uses a combination of statistical and computational methods to address the data problems that often appear in credit model validation and to provide an indication of the stability of default models over time.
2. Because we have found that simple statistics (such as the number of defaults correctly predicted) are insufficient and often inappropriate in the domain of credit models, we have developed the use of several metrics for evaluating model performance: cumulative accuracy profiles (CAP) plots, accuracy ratios (AR), conditional information entropy ratios (CIER) and Mutual Information Entropy (MIE).
3. We demonstrate the validation techniques we describe by benchmarking a variety of popular credit risk models, including Moody's own Public Firm Default Model, using our proprietary databases. To our knowledge, it is the first time that such broad analysis has been undertaken using an extensive comprehensive data set and a consistent methodology based on the information content of the models.
4. Using Mutual Information Entropy, we are able to demonstrate the amount of additional predictive information contained in one credit model versus another, which often suggests situations in which two models can be profitably combined or in which an inferior model can be eliminated.

[1] Basel (1999).
[2] This work and the details of Moody's Public Firm model are described more fully in Sobehart, Stein, Mikityanskaya and Li (2000).

Table of Contents

1 Introduction

Credit risk can be defined as the potential that a borrower or counter-party will fail to meet its obligations in accordance with the terms of an obligation's loan agreement, contract or indenture. For most individual and institutional investors, bonds and other tradable debt instruments are the main source of credit risk. In contrast, for banks, loans are often the primary source of credit risk.

Since banks often lend to unrated firms, they often have need of supplemental credit assessments. However, since a bank's individual exposures to such firms are often relatively small, it is typically uneconomical for borrowers to obtain a Moody's rating or for banks to devote extensive internal resources to the analysis of a particular borrower's credit quality. Not surprisingly, these economic factors have caused banking institutions to be among the earliest adopters of quantitative credit risk models.

A major challenge in developing models that can effectively assess the credit risk of individual obligors is the limited availability of high-frequency objective information to use as model inputs. In cases where no historical data is available at all, both model development and validation must rely on heuristic methods and domain experts. However, when historical data are available, model validation can proceed in a more objective and rigorous context. The approach we present in this *Rating Methodology* is an example of such a validation strategy.

Most models estimate the creditworthiness over a period of one year or more, which often implies a need for several years of historical financial data for each borrower.[3] While reliable and timely financial data can usually be obtained for the largest corporate borrowers, they are difficult to obtain for smaller borrowers, and are particularly difficult to obtain for companies in financial distress or default, which are key to the construction of accurate credit risk models. The scarcity of reliable data required for building credit risk models stems from the highly infrequent nature of default events.

In addition to the difficulties associated with developing models, both the limited availability of data and the averaging effect (over multiple credit cycles) present challenges in assessing the accuracy and reliability of credit risk models. As institutions become more familiar with credit modeling technology, their focus is widening to include a much higher level concern with model validation. Unfortunately, due to data issues and the infrequent nature of defaults, many statistical tests of model accuracy are not sensitive enough to adequately distinguish gradations of effectiveness between models under these data-poor circumstances.

The Basel Committee on Banking Supervision, in its recent report on credit risk modeling, highlighted the relatively informal nature of the credit model validation approaches at many financial institutions. In particular, the Committee specifically emphasized data sufficiency and model sensitivity analysis as significant challenges to validation. While the Committee has identified validation as a key issue in the use of quantitative default models, they conclude that[4]

"*…the area of validation will prove to be a key challenge for banking institutions in the foreseeable future.*"

This *article* describes several of the techniques that Moody's has found valuable for quantitative default model validation and benchmarking. More precisely, we focus on (a) the segmentation of data for model validation and testing, and (b) several robust measures of model performance and inter-model comparison that we have found informative and currently use. The techniques we present are especially useful in domains where the sparseness of default data makes standard statistical approaches unreliable. We also address the two fundamental issues that arise in validating and determining the accuracy of a credit risk model under:

1) *what is measured*, or the metrics by which model "goodness" should be defined; and

2) *how it is measured*, or the framework that should be used to ensure that the observed performance can reasonably be expected to represent the behavior of the model in practice.

The structure of this *article* is as follows: in section 2 we briefly discuss model accuracy and its impact on credit risk assessment. Moody's validation methodology and model testing framework are discussed in section 3. In section 4 we describe some of the model performance measures used by Moody's to assess model performance. To demonstrate each performance measure, we apply it using our framework to a

[3] See, for example, Herrity, Keenan, Sobehart, Carty and Falkenstein (1999).
[4] Basel, op. cit., p. 50.

wide variety of popular credit models. In section 5 we present a summary of the approach. A mathematical appendix provides details on the exact form of the measures we describe.

2 Model Accuracy

Although accuracy is only one dimension of model quality,[5] it is often the most prominent one in discussions of credit risk models. Because credit risk models are often used to generate opinions of credit quality on which investment decisions are taken, it is important to understand each model's strengths and weaknesses.

When used as classification models, default risk models can err in one of two ways. First, the model can indicate low risk when, in fact, the risk is high. This is referred to as Type I error, and corresponds to the assignment of a high ranking (low credit risk) to issuers who nevertheless default or come close to defaulting in their obligations. The cost to the investor can be the loss of principle and interest that was promised, or a loss in the market value of the obligation. Second, the model can assign a low ranking (high credit risk) when, in fact, the risk is low. This case is referred to as Type II error. Potential losses resulting from Type II error include the loss of return and origination fees when loans are either turned down or lost through non-competitive bidding. In the case of tradable loans or securities, Type II error may result in the selling of obligations that could be held to maturity, at disadvantageous market prices.

These accuracy and cost scenarios are described schematically in Figure 1 and Figure 2, below.

Figure 1. Types of Errors

		Actual	
		Low Credit Quality	**High Credit Quality**
Model	Low Credit Quality	Correct Prediction	Type II Error
	High Credit Quality	Type I Error	Correct Prediction

Figure 2. Costs of Errors

		Actual	
		Low Credit Quality	**High Credit Quality**
Model	Low Credit Quality	Correct Assessment	Opportunity costs, and lost potential profits. Lost interest income and origination fees. Premeature selling at disadvantageous prices.
	High Credit Quality	Lost interest and principle through defaults. Recovery costs. Loss in market value.	Correct Assesment

Although it is possible for some risk models to commit less of one type of error than another, investors and financial institutions usually seek to keep the probability of making *either* type of error as small as possible. Unfortunately, minimizing one type of error usually comes at the expense of increasing the other type of error. That is, the probability of making a Type II error increases as the probability of a Type I error is reduced.

The issue of model *error cost* is a complex and important one. It is often the case, for example, that a particular model will out-perform another under one set of cost assumptions, but be disadvantaged under a different set of assumptions[6]. Since different institutions have different cost and payoff structures, it is difficult to present a single cost function that is appropriate across all firms. For this reason, in the tests described in this *article*, we use cost functions related only to the information content of the models.

[5] See Dhar and Stein (1997) for a discussion of factors affecting model quality.
[6] See, for example, Provost and Fawcett (1997) or Hoadley and Oliver (1998).

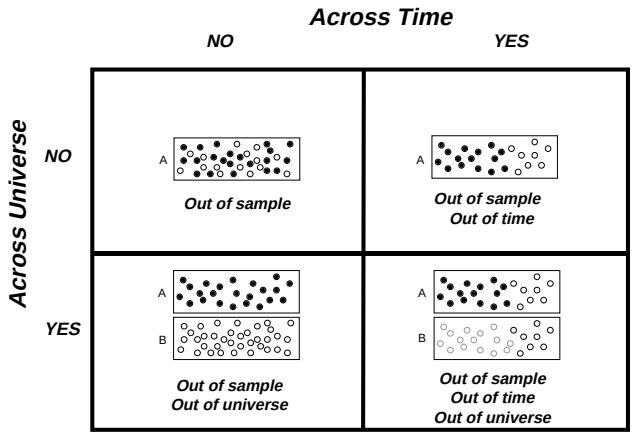
3 A Validation Framework For Quantitative Default Models

The performance statistics for credit risk models can be highly sensitive to the data sample used for validation. To avoid embedding unwanted sample dependency, quantitative models should be developed and validated using some type of out-of-sample[7], out-of-universe and out-of-time testing approach on panel or cross-sectional data sets[8]. However, even this seemingly rigorous approach can generate false impressions about a model's reliability if done incorrectly. Hold out testing can easily miss important model problems, particularly when processes vary over time, as credit risk does[9].

The statistical literature on model selection and model validation is quite broad. While we will not attempt to exhaustively cover this topic, the methodology described here brings together several streams of the validation literature that we have found useful in evaluating quantitative default models.

In the following section, we describe a validation framework that accounts for variations across both time and across the population of obligors. In doing so, it can provide important information about the performance of a model across a range of economic environments[10]. A schematic of the framework is shown in Figure 3. The figure breaks up 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 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 another.

Figure 3. Schematic of out of sample validation techniques



Testing strategies are broken out 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. (Reproduced from Dhar and Stein (1998).)

The upper left quadrant describes the approach in which the testing data for model validation is chosen completely randomly from the full training data set. This approach to model validation assumes that the properties of the data stays stable over time (stationary process). Because the data are drawn at random, this approach validates the model across the population of obligors preserving its original distribution.

The upper right quadrant describes one of the most common testing procedures. In this case, data for model training are chosen from any time period prior to a certain date and testing data are selected from time periods only after that date. A model constructed with data from 1990 through 1995 and tested on data from 1996 through 1999 is a simple example of this out-of-time procedure. Because model validation is performed with out-of-time samples, the testing assumptions are less restrictive than in the previous

[7] Out-of-sample refers to observations for firms that are not included in the sample used to build the model. Out-of-universe refers to observations whose distribution differs from the population used to build the model. Out-of-time refers to observations that are not contemporary with the training sample.

[8] A panel data set contains observations over time on many individuals. A cross sectional data set contains one observation on many individuals.

[9] See, for example, Mensah (1984).

[10] The presentation of the validation framework follows closely that of Dhar and Stein (1998) and Stein (1999). The performance measures and visualization tools we propose are described in Keenan and Sobehart (1999), with additional clarifications and enhancements.

case and time dependence can be detected using different validation sub-samples. However, since the sample of obligors is drawn from the population at random, this approach also validates the model preserving its original distribution.

The lower-left quadrant represents the case in which the data are segmented into training and testing sets containing no firms in common. In this general situation the testing set is out-of-sample. If the population of the testing set is different from that of the training set, the data set is out-of-universe. An example of out-of-universe would be a model that was trained on manufacturing firms but tested on other industry sectors. Because the temporal nature of the data is not used for constructing this type of out-of-sample test, this approach validates the model homogeneously in time and will not identify time dependence in the data. Thus, the assumption of this procedure is that the relevant characteristics of the population do not vary with time.

Finally, the most flexible procedure is shown in the lower-right quadrant and should be the preferred sampling method for credit models. In addition to being segmented in time, the data are 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-universe sampling). An example of this approach[11] would be a model constructed with data for all rated manufacturing firms from 1980 to 1989 and tested on a sample of all retail firms rated Ba1 or lower for 1990 to 1999.

It is common to validate a default model by using observed data on historical defaults. However, model validation based solely on predicted default events can be problematic because statistical tests for samples with low default rates often have extremely low power[12] and consequently, would require many (unavailable) default events to produce reliable results. A common fix-up is to use long time horizons (e.g.: ten or twenty years of data) to create large panel data sets.

Unfortunately, this approach may introduce bias in the testing procedure due the high temporal correlation of the model outputs and the low number of distress or default events. Temporal correlation in the data cannot be ignored since it violates the assumptions of many standard statistical tests of model performance (e.g., the Kolmogorov-Smirnoff test). On the other hand, if a hold-out sample is selected over a relatively short time frame (to avoid aggregation issues), tests based on this sample may incorrectly disqualify relatively accurate models and certify the accuracy of many relatively poor models due to insufficient data.

Because default events are rare and default model outputs for consecutive years are highly correlated, 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 cross-sectional data. While such out-of-sample and out-of-time tests would unquestionably be the best way to compare models' performance if default data were widely available, this is usually not the case. 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.

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. The procedure we describe is often referred to in the trading model literature as "walk-forward" testing.

The procedure works as follows. Select a year, for example, 1989. Then, fit the model using all the data available on or before the selected year. Once the model form and parameters are established, generate the model outputs for all the firms available during the following year (in this example 1990). Note that the predicted model outputs for 1990 are out-of-time for firms existing in previous years, and out-of-sample for all the firms whose data become available after 1989. Now move the window up one year, using all of the data through 1990 to fit the model and 1991 to validate it. The process is repeated using data for every year.

[11] This case is particularly important when one type of error is more serious that another, that is, there is cost structure associated to different errors. To illustrate these ideas, an error of two notches for an Aa-rated credit is generally less costly than a similar error for a B-rated credit, given the latter's relative proximity to default. The cost structure depends, among other things, on the action taken as a result of accepting or rejecting obligors based on the outputs of the models.

[12] Recall that statistical power refers to the probability that a statistical test at a particular significance level will unintentionally confirm the null hypothesis when in fact an effect is present. While significance gives information about Type II error, power gives information on Type I error. For an overview see Cohen (1988).

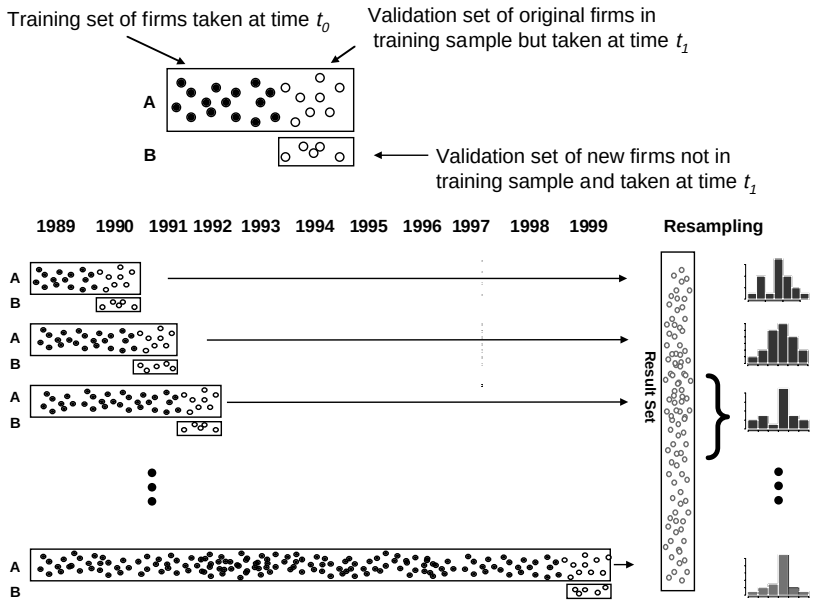
Collecting all the out-of-sample and out-of-time model predictions produces a set of model performances. This *validation result set* can then be used to analyze the performance of the model in more detail. Note that this approach simulates, as closely as possible given the limitations of the data, the process by which the model will actually be used in practice. Each year, the model is refit and used to predict the credit quality of all known credits, one year hence. The process is outlined in the lower left of Figure 4.

For example, for Moody's Public Firm Default Model, we selected 1989 as the first year for which to construct the validation result set (prior to 1989 we did not have enough data to build a sufficiently reliable model). Following the above procedure, we constructed a validation result data set containing over 54,000 observations (firm-years), representing about 9,000 different firms, and including over 530 default events from Moody's extensive database.

Once a result set of this type has been produced, a variety of performance measures of interest can be calculated, (we suggest several in Section 4). However, before turning to performance evaluation, it is important to note that the result set is itself a sub-sample of the population and, therefore, may yield spurious model performance differences based only on data anomalies. A common approach to addressing this problem, and one used extensively in our research, is to use one of a variety of *resampling* techniques to leverage the available data and reduce the dependency on the particular sample at hand.[13]

A typical resampling technique proceeds as follows.[14] From the result set, a sub-sample is selected at random. The performance measure of interest (e.g.: number of defaults correctly predicted) 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 mean (standard error, etc.) of this distribution then becomes the reported performance measure. A schematic of Moody's entire validation process is shown in Figure 4.

Figure 4. Moody's testing methodology: end-to-end



Moody's fits a model using a sample of historical data on firms and tests the model using both data on those firms one year later, and using data on new firms one year later (upper portion of exhibit). Dark circles represent training data and white circles represent testing data. We do "walk-forward testing" (bottom left) by fitting the parameters of a model using data through a particular year, and testing on data from the following year, and then inching the whole process forward one year. The results of the testing for each validation year are aggregated and then resampled (lower left) to calculate particular statistics of interest.

[13] The bootstrap (e.g., Efron, B. and R. J. Tibshirani (1993)), randomization testing (e.g., Sprent, P. (1998)), and cross-validation (ibid.) are all examples of resampling tests.

[14] A type of resampling was also used in Herrity, Keenan, Sobehart, Carty and Falkenstein (1999).

Resampling approaches provide two related benefits. First, they give an estimate of the variability around the actual reported model performance. In those cases in which the distribution of means converges to a known distribution, this variability can be used to determine whether differences in model performance are statistically significant using familiar statistical tests. In cases where the distributional properties are unknown, non-parametric permutation type tests can be used instead.

Second, because of the 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 inferior to model A. However, a resampling technique like the one we use might show that the models were virtually equivalent. In our testing, 85% of the result set was drawn at random (resampled) 100 times and the metric of interest (and its distribution) was calculated taking into account the correlation of the estimates.

In the next section, we use these approaches to compute the values of several specific performance measures that Moody's has found to be particularly valuable in evaluating quantitative credit models.

4 Model Performance And Benchmarking

In this section we introduce objective metrics for measuring and comparing the performance of credit risk models and analyzing information redundancy[15]:

(1) Cumulative Accuracy Profiles,

(2) Accuracy Ratios,

(3) Conditional Information Entropy Ratios, and

(4) Mutual Information Entropy.

These techniques are quite general and can be used to compare different types of models even when the model outputs differ and are difficult to compare directly. Furthermore, categorical outputs, such as the credit ratings produced by Moody's, can be evaluated side by side with continuous score values generated by a model.

In order to demonstrate the applicability of the methodology described here, we compared seven univariate and multivariate models of credit risk using Moody's proprietary databases including our default database and our credit modeling database. We compared the following models:

(1) a simple univariate model based on return on assets (ROA),

(2) reduced Z' score model[16] (1993),

(3) Z' score model (1993),

(4) a hazard model[17] (1998),

(5) a variant of the Merton model based on distance to default,[18] and

(6) Moody's Public Firm model, a model based on ratings, market and financial information (2000).

These models represent a wide range of modeling approaches listed in order of complexity.

Inter-model comparison is essentially the comparison of prediction errors for each model. Unfortunately, a large segment of the validation research found in the literature can be viewed as "residual error diagnostics" (e.g., t-statistics) which are of limited practical use for model comparison. Many of the assumptions that underlie residual diagnostics are frequently violated in practice.[19] Although it is not difficult to determine to what extent these assumptions are violated in each case, it is exceedingly difficult to determine *how to correct* the t-statistics figures or other statistics that authors cite in recommendation of their models.

---

[15] See Keenan and Sobehart (1999).

[16] For the definition of the original Z score and its various revisions Z' see Altman (1968) and Caouette, Altman, Narayanan (1998).

[17] For simplicity we selected the model based on Zmijweski's variables described in Shumway (1998).

[18] For this research, Moody's has adapted the Merton model (1973, 1974) in a similar fashion to which KMV has modified it to produce their public firm model. More specifically, we calculate a Distance to Default based on equity prices and firm's liabilities. See also Vasicek (1984) and McQuown (1993). For an exact definition of Moody's distance to default measure see Sobehart, Stein, Mikityanskaya and Li (2000).

[19] In particular, independence of samples or the Gaussian distribution of errors does not typically hold.

The techniques discussed below are useful not only because of their power and robustness, but because they can be used to compare default prediction models, even when data is correlated or otherwise "messy", or when its true statistical properties are unknown.

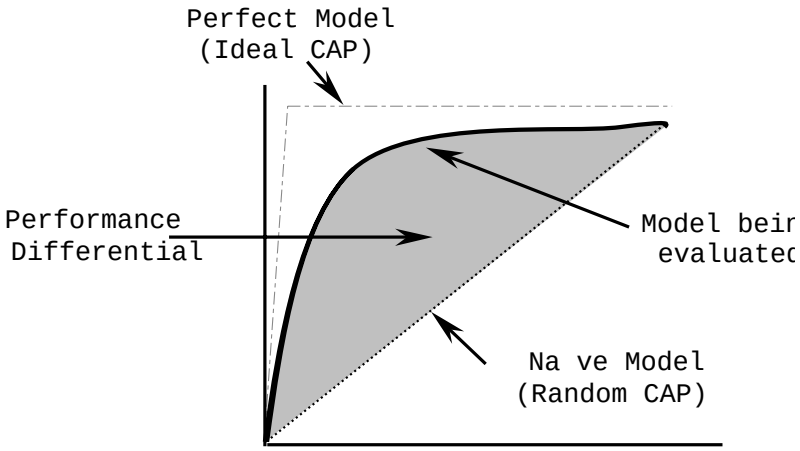
Comparing the performance across different default prediction models is challenging since the models themselves usually measure slightly different aspects of the default events and time horizons and may be expressing a quantification of credit risk using different types of outputs. For example, some models calculate an explicit probability of default, or expected default frequency, which is a number between zero and one and is usually reported to several decimal places. Others, such as agency ratings, rank risk in using a coarser scale, but incorporate other aspects of default, such as recovery and expected losses. Moreover, Moody's ratings are intended to endure normal economic cycles and, therefore, place a premium on stability over long time horizons. In contrast, some models are designed to react sharply to potential changes in short-term creditworthiness and market conditions[20].

4.1 Cumulative Accuracy Profiles (CAPs)

Moody's uses Cumulative Accuracy Profiles (CAP), to make visual, qualitative assessments of model performance. While similar tools exist under a variety of different names (lift-curves, dubbed-curves, receiver-operator curves, power curves, etc.). Moody's use of the term CAP refers specifically to the case where curve represents the cumulative probability over the *entire* population, as opposed to the non-defaulting population only[21]. This form of the plot is particularly useful in that it simultaneously measures Type I and Type II errors.

To plot cumulative accuracy profiles, companies are first ordered by model score, from riskiest to safest. For a given fraction *x%* of the total number of companies, a CAP curve is constructed by calculating the percentage *y(x)* of the defaulters whose risk score is equal to or lower than the one for fraction x. Figure 5 shows an example of a CAP plot.

Figure 5. Type I CAP curve



The dark curved line shows the performance of the model being evaluated. It depicts the percentage of defaults captured by the model (vertical axis) vs. the model score (horizontal axis). The heavy dotted line represents the naïve case of zero information (which is equivalent to a random assignment of scores). The gray dashed line, represents the case

[20] An attractive feature of these validation measures, not discussed in detail in this Rating Methodology, is that they can also provide estimates of a model's precision. Although model outputs are often given as "continuous" variables, in reality, due to data limitations and statistical significance, all models that are econometrically calibrated to historical default frequency will exhibit some underlying granularity in their outputs. This is true of most statistical models and also of structural models (e.g., contingent claims models) when they are adjusted to reflect historical default experience. A key issue in model comparison is to determine whether a higher degree of refinement in the scale of a given model's output represents any additional "precision" supported by statistical evidence, or whether small increments in estimated risk just reflect random noise. For example: is there a statistically meaningful difference between a model default prediction of 2 bp and 3 bp? For these tests, the minimum finite precision that produces a significant difference in the performance of the model determines the precision of the model output. See: Keenan and Sobehart (1999) for a more detailed discussion.

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

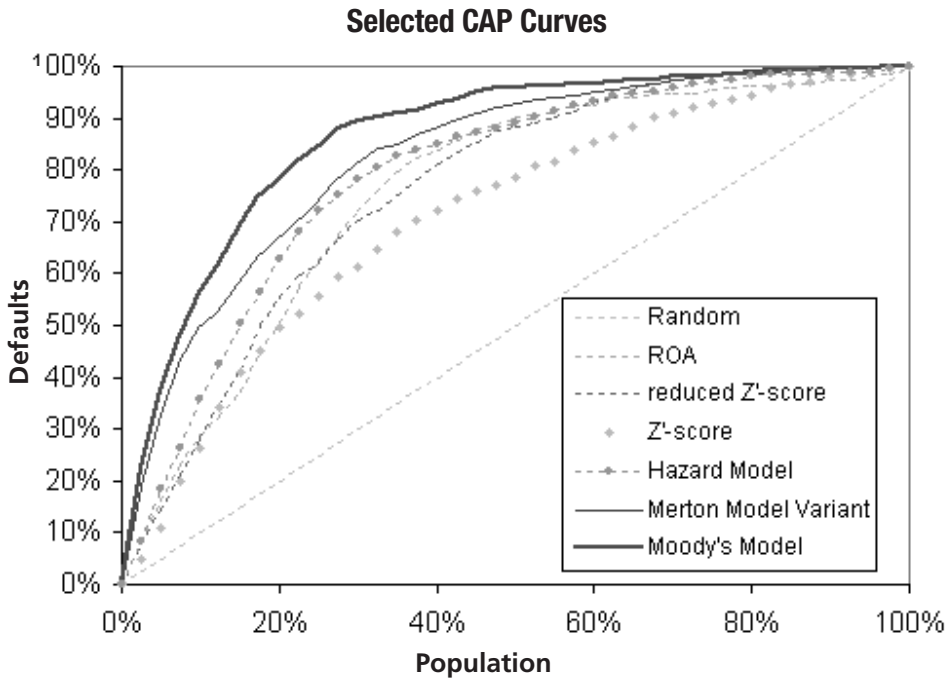
in which the model is able to discriminate perfectly and all defaults are caught at the lowest model output. The gray region represents the performance differential between the naïve model and the model being evaluated.

A good model concentrates the defaulters at the riskiest scores and so the percentage of all defaulters identified (the *y* axis in the figure above) increases quickly as one moves up the sorted sample (along the *x* axis). If the model were totally uninformative, if, for example, it assigned risk scores randomly, we would expect to capture a proportional fraction, i.e., x% of the defaulters with about *x*% of the observations, generating a straight line or *Random CAP* (the dotted line in Figure 5). A perfect model would produce the *Ideal CAP*, which is a straight line capturing 100% of the defaults within a fraction of the population equal to the default rate of the sample. Because the historical default rate is usually a small number, the ideal CAP would look like a vertical line at the point in the plot where the percentage of remaining firms was equal to the actual number of defaults.

A good model also concentrates the non-defaulters at the lowest riskiness. Therefore, the percentage of all non-defaulters (the *z(x)* variable) should increase slowly at first. One of the most useful properties of CAPs is that they reveal information about the predictive accuracy of the model over its entire range of risk scores for a particular time horizon.

Figure 6 shows the CAP curves for several models using the validation sample (out-of-sample and out-of-time). The values plotted represent the mean values of the resampling tests. Similar results are obtained for the in-sample tests.[22] Note that Moody's Public Firm model appears to outperform all of the benchmark models consistently.

Figure 6. CAP curves for the tested models



This composite figure shows the CAP curves for six models. All models were tested on the same data set. The 45° dashed gray line represents the naïve case (which is equivalent to a random assignment of scores). All models perform considerably better than random, however the nonlinear hybrid model clearly outperforms all others. Note that the second best model, the Merton model variant, performs almost as well as the nonlinear model in the case of extremely poor quality firms, but that the nonlinear model clearly performs better beyond about the bottom 10% of the populations and is much better at discriminating defaults in the middle ranges of credits.

[22] Here *in-sample* refers to the data set used to build Moody's nonlinear model.

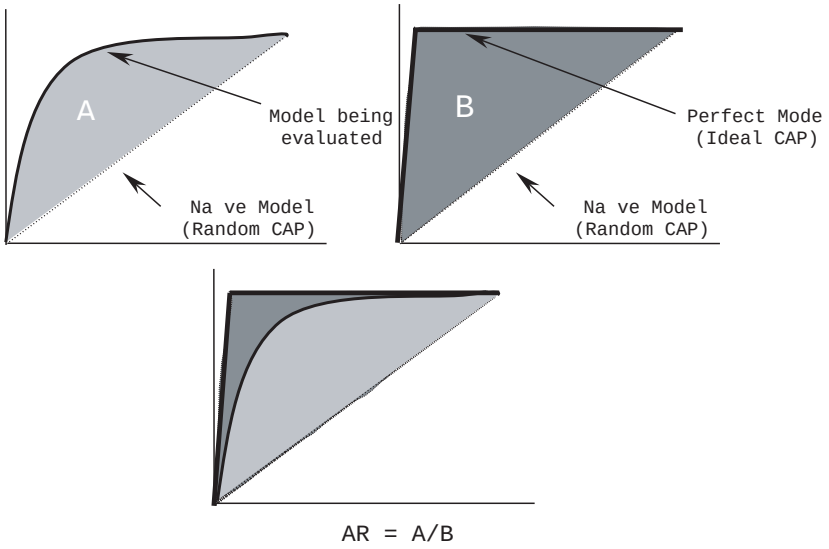
4.2 Accuracy Ratios (ARs)

While CAP plots are a convenient way to visualize model performance, it is often convenient to have a single measure that summarizes the predictive accuracy of each risk measure for both Type I and Type II errors into a single statistic. We obtain such a measure by comparing the CAP plot of any set of risk scores with the ideal CAP for the data set under consideration; the closer the CAP is to its ideal, the better the model performs. To calculate the summary statistic, we focus on the area that lies *above* the Random CAP and is *below* the model CAP. The more area there is below the model CAP and above the Random CAP, the better the model is doing overall (Figure 5).

The maximum area that can be enclosed above the Random CAP is identified by the Ideal CAP. Therefore, the ratio of the area between a model's CAP and the random CAP to the area between the ideal CAP and the random CAP summarizes the predictive power over the entire range of possible risk values. We refer to this measure as the Accuracy Ratio (AR), which is a fraction between 0 and 1. Risk measures with ARs close to 0 display little advantage over a random assignment of risk scores while those with ARs near 1 display almost perfect predictive power.

The accuracy ratio can be envisioned as the ratio of the shaded region in the graph on the left of Figure 7 to the shaded region on the right of Figure 7, shown in the bottom of Figure 7, below:

Figure 7. Heuristic representation of the Accuracy Ratio



The accuracy ratio is the ratio of (A) the performance improvement over the naïve model of the model being evaluated to (B) the performance improvement over the naïve model of the Perfect Model. It can be envisioned as the ratio of the shaded region in the graph on the left of to the shaded region on the right. The result is shown in the bottom of the figure.

Most of the models we tested had AR's in the range of 50% to 75% for out-of-sample and out-of-time tests. The results we report here are the product of the resampling approach described in the previous section. Thus, in addition to the reported value, we are also able to estimate an error bound of the statistic. We found that the maximum absolute deviation of the AR is of the order of 0.02 for most models.[23] Not surprisingly, we found that accuracy of the estimates deteriorates for small samples.

In a loose sense, AR is similar to the commonly used Kolmogorov-Smirnov (KS) test designed to determine if the model is better than a random assignment of credit quality. However, AR is a global measure of the discrepancy between the CAPs while the KS test focuses only on the maximum discrepancy. Since the K-S focuses only on a single maximum gap, it can be misleading in cases where two models behave quite differently as they cover more of the data space from low risk model outputs to high risk

[23] Due to the high levels of correlation in the resampling, the maximum absolute deviation gives a more robust estimate of an error range than a corrected standard error.

model outputs. Also notice that, because the comparison of ARs is relative to a data set, our definition of the AR is not restricted to having completely independent samples as in the KS test.[24]

Table 1 shows AR values for the tested models for in-sample and validation tests (out-of-sample and out-of-time). The typical error bound is 0.02. To confirm the validity of the AR figures, we also checked if a particular model differed significantly from the one ranked immediately above it by calculating a KS statistics tests using over 9,000 independent observations selected from the (out-of-sample/out-of-time) validation set. KS tests support the AR results on the validation sample. More precisely, KS tests showed that only the reduced Z'-score and ROA were not significantly different.

Table 1. Selected Accuracy Ratios

	In-sample AR	Validation AR
ROA	0.53	0.53
Reduced Z'-Score	0.56	0.53
Z'-Score	0.48	0.43
Hazard Model	0.59	0.58
Merton Model Variant	0.67	0.67
Moody's Model	0.76	0.73

4.3 Conditional Information Entropy Ratio (CIER)

Another measure used to determine the power of a model is based on the information about defaults contained in the distribution of model scores, or information entropy. The *information entropy* (IE) is attractive since it is applicable across all types of model outputs, requires no distributional assumptions and is a powerful way of objectively measuring how much real value is contained in a set of risk scores. In the same way we reduced the CAP plot to a single AR statistic to create a measure that lent itself to comparison across models, we can reduce information entropy measures into another useful summary statistic to summarize how well a given model predicts defaults.

This is done via the Conditional Information Entropy Ratio[25] (CIER). The CIER compares the amount of "uncertainty" regarding default in the case where we have no model (a state of more uncertainty about the possible outcomes) to the amount of "uncertainty" left over after we have introduced a model (presumably, a state of less ignorance), with a given accuracy δ. The association of the word information with the concept of entropy should be taken in a "loose" sense because it usually carries the wrong connotation to the casual reader. Intuitively, the entropy measures the overall "amount of uncertainty" represented by a probability distribution. Thus, the CIER can be used to measure of the amount of uncertainty about defaults contained in the different models *as long as all the models are evaluated on the same data set.*

To calculate the CIER, we first calculate the uncertainty (IE) associated with the event of default without introducing any model. This entropy reflects knowledge common to all models – that is, the likelihood of default given by the probability of default for the sample as a whole. We then calculate the uncertainty after having taken into account the predictive power of the model. The CIER is one minus the ratio of the latter to the former.[26] If the model held no predictive power, the CIER would be 0. In this case the model provides no additional information on the likelihood of the outcomes that is not already known. If it were perfectly predictive, the conditional information entropy ratio would be 1. In this case, there would be no uncertainty about the outcomes and, therefore, perfect default prediction. Because the information entropy measures the reduction of uncertainty, a higher CIER indicates a better model. Table 2 shows the CIER results. CIER errors are of the order of 0.02 and are obtained with a bootstrap scheme similar to the one described for the AR measure.

Table 2. Selected Entropy Ratios

	In-sample CIER	Validation CIER
ROA	0.06	0.06
Reduced Z'-Score	0.10	0.09
Z'-Score	0.07	0.06
Hazard Model	0.11	0.11
Merton Model Variant	0.14	0.14
Moody's Model	0.21	0.19

[24] In fact, AR based on panel data sets will provide aggregated information about the time correlation of the risk scores.

[25] This is similar to measures such as gain ratios used in the information theory and time series analysis literature (see, for example, Prichard and Theiler (1995). However, our definition measures explicitly the uncertainty to predict defaults instead of the overall uncertainty in the distribution of model outputs (see *Keenan and Sobehart* (1999)).

[26] CIER = 1- IER, where IER is the information entropy ratio defined in Herrity, Keenan, Sobehart, Carty and Falkenstein (1999). Here we introduce CIER for consistency with the concept of conditional entropy in Information Theory and Communication Theory.

4.4 Mutual Information Entropy (MIE)

To this point, we have been describing methods of comparing models to each other on the assumption that the best performing model would be adopted. However, it is not unreasonable to question whether a combination of models might perform better than any individual one. Two models may both predict 10 out of 20 defaulters in a sample of 1,000 obligors. Unfortunately, this information does not provide guidance on which model to choose. However, if each model predicted a different set of 10 defaulters, then using both models would be the obvious solution as this composite approach would have double the predictive accuracy of either model individually[27]. In this hypothetical case, the models are independent. But, as is usually the case, there is considerable overlap, or dependence, in what two models will predict for any given data sample.

To quantify the dependence between any two models[28] A and B, Moody's uses a measure called the mutual information entropy (MIE). The mutual information entropy is a measure of how much information can be predicted about model B given the output of model A with a given accuracy δ.

If models A and B are independent, the mutual information entropy is zero, while if model B is completely dependent on model A then MIE = 1-CIER(A). The additional uncertainty generated by model B can be estimated by comparing with the uncertainty generated by model A alone. Table 3 shows the difference D = MIE(A,B) – MIE(A,A), where A is Moody's model and B is any of the other selected models. In this example, we have compared all the benchmark models to Moody's model to determine if they contain redundant information.[29] Because the MIE is calculated with the joint conditional distribution of models A and B, this measure requires a large number of defaults to be accurate.[30]

Table 3.
Difference of Mutual Information Entropy with respect to Moody's Public Firm model

	In-sample MIE	In-sample D	Validation MIE	Validation D
ROA	0.96	0.17	0.97	0.16
Reduced Z'-Score	0.93	0.14	0.96	0.15
Z'-Score	0.95	0.16	0.98	0.17
Hazard Model	0.91	0.12	0.92	0.11
Merton Model Variant	0.87	0.08	0.87	0.06
Moody's Model	0.79	0	0.81	0

The additional uncertainty generated by a model can be estimated by comparing with the uncertainty generated by Moody's model alone. Table 3 shows the difference D = MIE(A,B) – MIE(A,A), where A is Moody's model and B is any of the other selected models.

5. Summary

The benefits of implementing and using quantitative risk models cannot be fully realized without an understanding of how accurately any given model represents the dynamics of credit risk. This makes reliable validation techniques crucial for both commercial and regulatory purposes. In this *article* we have presented a set of measures and a testing approach that we have found useful for benchmarking default models and validating their performance. This approach, which continues to evolve, is part of Moody's ongoing efforts in the area of quantitative risk modeling.

The framework uses a combination of statistical and computational approaches that addresses the severe data problems that often present themselves in credit model validation. The approach is flexible and permits the calculation of arbitrarily many performance measures of interest. It facilitates direct statistical comparisons of models that produce quite different outputs.

[27] Of course combining the models could also create ancillary trade-offs with respect to increased Type II error. These unwanted side effects would need to be evaluated in the context of the models' usage.

[28] Here A and B refer to two different models. They should not be confused with the areas in Figure 7.

[29] In this context, the statistic serves much the same function as a correlation coefficient in a classic regression sense. However, the MIE statistic is based on information content of the models.

[30] This requirement can be relaxed by including degrees of credit quality instead of defaults only.

In the course of our research into quantitative credit modeling, we have found that simple statistics[31] (such as the number of defaults correctly predicted) are often inappropriate in the domain of credit models. As a result, we have developed several useful metrics that give a sense of the value added by a quantitative risk model. In this *Rating Methodology* we described four such measures: Cumulative Accuracy Profiles (CAP), Accuracy Ratios (AR), Conditional Information Entropy Ratios (CIER) and Mutual Information Entropy (MIE).

This last measure is interesting in that it permits analysts to assess the amount of additional predictive information contained in one credit model versus another. In situations where a specific model contains no additional information relative to another, the less informative should be discarded in favor of the more informative. In the special case where *both* models contribute information to each other, users may wish to combine the two to garner additional insight.

Finally, we attempted to fill a gap in the default model literature by benchmarking a variety of popular credit risk models, including Moody's Public Firm model, using Moody's extensive proprietary default database. This allowed us, by way of example, to demonstrate the validation approaches discussed in this *article*.

We feel that the approach we describe here is a very effective way to benchmark internal and external credit models where data permit. In that regard, we believe that it begins to address several of the Basle Committee's key concerns regarding validation. However, Moody's efforts in credit modeling in general, and default modeling in particular, are ongoing, as is our research on model validation. This *Rating Methodology* describes one approach that we use for model validation and benchmarking of quantitative models when sufficient data are available. Future *articles* will address other issues in this evolving field.

[31] For an example of a more standard approach to validation see: Caouette, Altman and Narayanan (1998).

6. References

Altman, E. I., (1968), "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", *Journal of Finance*, September, 589-609.

Basel (1999), "Credit Risk Modeling Practices and Applications," *Basle Committee on Banking and Supervision*, Basle, April.

Caouette, J. B., Altman, E. I., Narayanan, P., (1998), *Managing Credit Risk*: *The Next Great Financial Challenge*, New York, Wiley, pp. 112 – 122.

Cohen, J., (1988), *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Dhar, V. and Stein, R. (1998), "Finding Robust & Usable Models with Data Mining: Examples from Finance," *PCAI*, September, 1998.

Dhar, V. and Stein, R., (1997), *Seven Methods for Transforming Corporate Data into Business Intelligence*, Upper Saddle River, Prentice-Hall.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York, Chapman & Hall.

Herrity, J., Keenan, S.C., Sobehart, J.R., Carty, L.V., Falkenstein, E., (1999) *Measuring Private Firm Default Risk*, Moody's Investors Service Special Comment (June).

Hoadley, B. and Oliver, R. M., (1998), "Business measures of scorecard benefit," *IMI Journal of Mathematics Applied in Business & Industry*, 9, pp. 55-64.

Keenan, S.C., Sobehart J.R., (1999), "Performance Measures for Credit Risk Models", *Moody's Risk Management Services*, *Research Report* 10-10-99.

Mensah, Y. M., (1984), "An Examination of the Stationarity of Multivariate Bankruptcy Prediction Models: A Methodological Study," *Journal of Accounting Research*, Vol. 22, No. 1(Spring).

Merton, R.C., (1973), "Theory of Rational Option Pricing", Bell Journal of Economics and Management Science 4, 141-183.

Merton, R.C., (1974), "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates", Journal of Finance 29, 449-470.

McQuown, J.A., (1993), "A Comment On Market vs. Accounting Based Measures of Default Risk", KMV Corporation.

Provost, F. and Fawcett, T., (1997), "Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distributions," Proceedings Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, CA, August 14-17.

Prichard D., Theiler, J., (1995), "Generalized Redundancies for Time Series Analysis", *Physica D* 84, 476-493.

Shumway, T., (1998), "Forecasting Bankruptcy More Accurately: A Simple Hazard Model", University of Michigan Business School working paper.

Sobehart, J. R., Stein, R. M., Mikityanskaya, V., Li, L, (2000), "Moody's Public Firm Risk Model: A Hybrid Approach to Modeling Short-Term Default Risk", Moody's Investors Service (February).

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," *Information Systems Working Paper Series*, Stern School of Business, New York University. Working Paper #IS-99-015.

Vasicek, O. A., (1984), "Credit Valuation", KMV Corporation.

7. Appendix: A Mathematical Description of the Performance Measures[32]

7.1 ACCURACY RATIO

Mathematically, the AR value is defined as

$$AR = \frac{2\int_0^1 y(x)\,dx - 1}{1 - f} = \frac{1 - 2\int_0^1 z(x)\,dx}{f} \tag{A.1}$$

Here *y(x)* and *z(x)* are the Type I and Type II CAP curves for a population *x* of ordered risk scores, and $f = D/(N+D)$ is the fraction of defaults, where *D* is the total number of defaulting obligors and *N* is the total number of non-defaulting obligors. Note that our definition of AR provides the same performance measure for Type I and Type II errors.

7.2 CONDITIONAL INFORMATION ENTROPY RATIO

Consider two mutually exclusive outcomes of a credit event, one of which must be true: outcome D, the issuer defaults, or outcome *N*, the issuer does not default. Given a set of risk scores $R = \{R_1, .., R_n\}$ produced by a model, the conditional information entropy which measures the information about the propositions *D* (issuer defaults), and *N* (issuer does not default) is

$$H_1(R) = -\sum_k P(R_k)\big(P(D \mid R_k)\log(P(D \mid R_k) + P(N \mid R_k)\log P(N \mid R_k)\big) \tag{A.2}$$

where $P(D \mid R_k)$ is the probability that the issuer defaults given that the risk score is R_k. This value quantifies the average information gained from observing which of the two events *D* and *N* actually occurred.

For models with continuous outputs, the most straightforward way to estimate the quantities defined in equation (A.2) is to use a bin counting approach to quantize the values. The range of the model output is divided into a number of bins of size δ, usually corresponding to the accuracy of the output. Because equation (A.2) requires estimating the conditional distributions of defaults and non-defaults, the bins of size δ have to be bigger than the resolution of some of the model outputs to provide a meaningful statistics.

To calculate the *CIER*, we first calculate the information entropy $H_0 = H_1(p)$, where *p* is the default rate of the sample. That is, without attempting to control for any knowledge that we might have about credit quality, we measure the uncertainty associated with the event of default. This entropy reflects knowledge common to all models – that is, the likelihood of the event given by the probability of default. We then calculate the information entropy $H_1(R,\delta)$ after having taken into account the predictive power of the model. The *CIER* is defined as

$$CIER(R,\delta) = \frac{H_0 - H_1(R,\delta)}{H_0} \tag{A.3}$$

7.3 MUTUAL INFORMATION ENTROPY

To quantify the dependence between any of two models A and B, Moody's uses the mutual information entropy *(also called information redundancy)*

$$MIE(r,R,\delta) = \frac{1}{H_0}\big(H_1(r,\delta) + H_1(R,\delta) - H_2(r,R,\delta)\big) \tag{A.4}$$

where

[32] See, Herrity, Keenan, Sobehart, Carty and Falkenstein (1999), and Keenan and Sobehart (1999).

$$H_2(r,R,\delta) = -\sum_{j,k} P(r_j,R_k)\Big(P(D\mid r_j,R_k)\log(P(D\mid r_j,R_k)+P(N\mid r_j,R_k)\log P(N\mid r_j,R_k)\Big) \qquad \text{(A.5)}$$

Here $r = \{r_1,..,r_n\}$ and $R = \{R_1,\ldots,R_m\}$ are the outputs for models *A* and *B*.

The mutual information entropy is a measure of how many bits one can predict about model *A* given the output of model *B* with accuracy δ. If models *A* and *B* are independent, the mutual information is zero, while if model *B* is completely dependent on model *A* then $MIE(r,R,\delta) = 1\text{-}CIER(r,\delta)$.

Here the entropy H_2 is also implemented with a bin counting approach. A partition size δ is chosen, corresponding to highest accuracy of the two models, and the outputs of the models are discretized into integers $j = 1,\ldots n$, $k = 1,2,\ldots m$ depending on which bin of size δ they fall into. Because MIE requires estimating the conditional distributions of defaults and non-defaults, the bins of size δ have to be bigger than the resolution of some of the model outputs to provide a meaningful statistics. Here we use δ = 5% of the model output range for each model.

Rating Methodology Benchmarking Quantitative Default Risk Models: A Validation Methodology

Report Number: 53621