A survey of failure prediction models offered by vendors with an application to Belgian data

Janet Mitchell
Patrick Van Roy

Introduction

Failure prediction models are defined as models that assign a probability of failure or a credit score to firms.(1) The development of the Basel II framework for regulatory capital requirements has stimulated vendors to offer such models to banks opting to use the internal ratings-based approach for calculating their regulatory capital requirement. Indeed, one of the inputs that banks adopting the internal ratings-based approach must provide is an estimate of the probability of default (PD) for each firm borrower.

Models offered by vendors to calculate PDs, or credit scores which can be mapped into PDs, are often used by banks as an off-the-shelf product or, alternatively, as a basis for development and benchmarking of their own internal rating systems. Although there exists a vast academic literature on failure prediction models (see, e.g., Balcaen and Ooghe, 2006 for a review), much less is known about failure prediction models offered by vendors.

The purpose of this article is twofold. First, it provides an overview of differences in vendor models for private (i.e. non-listed) firms, with respect to model inputs, output, methodology and field of application. Second, it uses data for 40,000 small and medium-size Belgian firms to compare model output. One of the questions addressed is whether different models yield significantly different output, or "rankings" for the same firm. In other words, is there much disagreement between models? This question is important because if banks use the PD or score produced by a failure prediction model to price the loan of a borrower and if different failure prediction models result in significantly different PDs or scores for the same borrower, then the bank's choice of model can have a significant impact on loan origination and pricing decisions. Observing the extent of disagreement between two models, however, does not permit a judgement about the relative performance of the models; therefore, a second question addressed by the article is whether some models better identify failing firms than others. These issues are investigated by comparing the output of two models offered by vendors with the output of a failure prediction model developed by the National Bank of Belgium (see Vivet, 2004).

The analysis reveals that models do appear to frequently disagree in firm rankings, and the level of disagreement can be quite important. At the same time, all three of the models studied here perform similarly, and quite well, in distinguishing sample firms that entered bankruptcy within one year from firms that did not enter bankruptcy during that period (understandably, the models perform somewhat less well in distinguishing firms that entered bankruptcy within five years from firms that did not). This excellent performance of the three models, which is considered to be reassuring from a financial stability viewpoint, is also important, given that the statistical methods

(1) The term "failure" here may represent either bankruptcy or default. The use of this term is motivated by the fact that vendor models assessing the credit risk of a firm may be calibrated on either bankruptcy or on default data. Bankruptcy data is normally available on a public basis, while default is normally non-public information which is generated and maintained within financial institutions.

as well as the measure of failure (bankruptcy vs. default) used to estimate and calibrate the models are different.

1. Characteristics of failure prediction models for private firms offered by vendors

The general principle behind failure prediction models consists of aggregating different types of information (inputs) through a given rule (methodology) in order to produce a credit "score" or a PD (output) for a specific obligor (field of application). Although this framework applies to every failure prediction model, there are notable differences in the specific inputs, methodology, output and field of application across vendors' models.

INPUTS

Vendors typically develop failure prediction models for private companies on very large databases that include hundreds of thousands (if not millions) of financial statements. The selection of input variables occurs in different ways. Some vendors rely on expert judgement to identify inputs, where the expert judgement may be complemented with statistical analysis. Many vendors rely on more statistical approaches, e.g. a "forward-selection" process, which consists of first identifying a set of variables that are correlated with business failure, then evaluating the performance of successive estimation of a statistical model, where the variables are tried out one by one and included in the model if they are statistically significant.

The variables used as inputs in failure prediction models are primarily quantitative; however, qualitative variables are sometimes used to supplement the quantitative information. The quantitative inputs include firm-level variables from financial statements (these variables can be grouped in categories such as leverage, liquidity, profitability, size, etc.). Sometimes industry-specific variables (such as industry dummies or market-based variables constructed from listed companies), as well as macro-economic variables (e.g. an indicator of industrial production), are also included. Qualitative information may include variables such as the legal form of the firm, its age, its geographical location, the layout of its annual accounts (full or abbreviated), etc.

Most vendor models generate an output even for firms for which some inputs are missing. In this case, the missing values are generally replaced with the average (or median) values of the relevant population. However, some models do not generate an output when too many inputs for a firm are missing.

METHODOLOGY

Failure prediction models for private firms are developed using statistical methods that include multiple discriminant analysis, probit (or logit) regressions, neural networks, decision trees, maximum expected utility and hybrid approaches that integrate two or more of these techniques (see Box 1).[1] Vendors usually perform refinements of these methods; e.g., they may not use a "pure" probit model but rather a probit model that involves some transformations of inputs and outputs.

OUTPUT

The output for a failure prediction model can be either a credit score or a PD. These measures are generated for different time horizons, most frequently between one and five years (although some models are restricted to shorter horizons).

In some cases, the vendors will translate the output into rating estimates expressed using the standard rating symbols, where the rating estimates are designed to exhibit default rates which, in aggregate, are similar to those published by the main credit rating agencies (however, the highest rating estimate attainable for private firms may fall below AAA). Some vendor models also allow users to map their own internal rating systems to rating estimates. Finally, most models offer the opportunity for the user to run a scenario analysis on a company by modifying some of the inputs and observing the impact on output.

FIELD OF APPLICATION

Some vendors offer different country-specific and, more rarely, industry-specific versions of their model. There are three main reasons for this. First, data availability may prevent using the same input variables for all countries and industries. Second, input variables may differ across countries or industries because determinants of failure are not the same. Third, some countries or industries are intrinsically riskier than others and experience therefore higher failure rates, even if they share the same determinants of failure as other countries or industries. As a result, country and industry-specific versions of the same model tend to differ with respect to their inputs and

(1) Failure prediction models developed by academics and banks also use these techniques, although with some differences. For instance, academic studies seem more frequently to use data mining techniques such as neural networks or support vector machines, while bank internal models tend also to rely on heuristic methods built on expert judgement.

parameter estimates, although they usually share the same methodology. In addition, some outputs (e.g. rating estimates) may not necessarily be available for all country or industry-specific versions of a model. However, when available, outputs have the same format and are designed to be comparable across model versions.

Box 1 – Some statistical methodologies used by vendors

MULTIPLE DISCRIMINANT ANALYSIS

Multiple Discriminant Analysis (MDA) is a classification technique that allows finding a function of variables that best distinguishes between two groups of firms, failing and non-failing, by maximizing the between-group variance while minimizing the within-group variance. In the case of linear MDA (the approach predominantly used by vendors), this function is a weighted linear combination of firm characteristics *X*:

$$D = a_1 . X_1 + a_2 . X_2 + ... + a_n . X_n ,$$

where *D* is the discriminant score (an indicator of the firms' financial health), *n* refers to the number of firm characteristics used in the discriminant function and *a* stands for each variable's coefficient.

Since the convention is to have low values of the discriminant score indicating a high credit risk, firms are classified in the failing group if *D* is lower than a certain cut-off *C* and in the non-failing group otherwise.

PROBIT AND LOGIT MODELS

Probit and logit models are regression models that allow estimating the probability of firm failure *p* conditional on a set of firm characteristics *X* by a non-linear maximum-likelihood estimation procedure. A general formulation for both models is given by:

$$p = F (a_1 . X_1 + a_2 . X_2 + ... + a_n . X_n) ,$$

where *F* is either the standard normal distribution (probit model) or the logistic distribution (logit model) and *a* and *n* have the same interpretation as above.

Contrary to multiple discriminant analysis and probit (logit) regressions, which model the relationship of company failure with a number of variables, machine learning techniques such as neural networks, decision trees and maximum expected utility attempt to "learn" this relationship from the data.

NEURAL NETWORKS

Neural networks (NN) are mathematical models that are inspired by the architecture of the human brain. Formally, NN consist of an input layer, which takes in firm information and passes it to downstream "neurons"; inner layers, which transform this information using both linear and nonlinear (e.g. logit) functions; and an output layer which receives the calculated output (failure prediction).

In the estimation process, neural networks are trained in order to minimize the deviation between the calculated and the actual output. The most commonly used method consists in adjusting the relevant weights in the linear transformation performed in the inner layers.

DECISION TREES

Decision trees (DT) are classification algorithms used to partition the data by employing variables to identify the subgroups that contribute most to explaining the dependent variable. Formally, DT consist of internal nodes with associated splitting predicates based on firm characteristics and final leafs with associated values (failure prediction).

The concept of entropy is most commonly used to determine on which firm variables the tree should be split.

MAXIMUM EXPECTED UTILITY

Maximum expected utility (MEU) is a technique which can be used to model conditional default probabilities among other risk parameters. MEU allows finding a probability measure (failure prediction) that maximizes the out-of-sample expected utility of an investor who chooses his investment strategy so as to maximize this expected utility under the model he believes to be efficient.

Note that vendors may also combine some of the above-mentioned methodologies (e.g. multiple discriminant analysis and decision trees) in a hybrid approach.

2. Failure prediction models offered by vendors: an application to Belgian data

This section compares the output of the National Bank of Belgium's bankruptcy prediction model (NBB) and two vendor models (Model 1 and Model 2) when estimated using a set of non-financial Belgian firms with at least 5 employees. The three models are compared on the basis of predictions made using input data from 2001 and 2004 and bankruptcy data between 2002 and 2006. More precisely, bankruptcies in 2002 and between 2002 and 2006 are used to assess the one-year and five-year predictions made on the basis of 2001 data, while bankruptcies in 2005 are used to assess the one-year predictions made on the basis of 2004 data. The database contains almost 30,000 firms in 2001 and nearly 40,000 in 2004.

The NBB model (see Vivet, 2004) is a logit model that predicts the risk of bankruptcy of Belgian firms using seven financial variables constructed directly from the firms' annual accounts and an additional variable, the time (in days) taken by firms to submit their annual accounts. The financial variables include one variable for capitalisation, one measure of earnings, two measures of liquidity and three variables reflecting solvency. The two vendor models, Model 1 and Model 2, use different statistical methods and are based on different inputs. All three models were calibrated on a set of failing and non-failing firms, where failure may represent default for some models and entry into bankruptcy for others.

Because not all of the models under consideration produce PDs, it is not possible to directly compare model outputs by examining PD estimates for each firm. However, it is possible to compare the ordinal rankings of firms across the different models. A convenient way to make this comparison is as follows. First, for each model, order firms from lowest to highest credit risk based on their credit score or PD. Then, allocate the firms into a predefined number of classes, or "buckets", according to a pre-defined distribution.[1] It is then possible to compare rates of model disagreement across classes, as well as the frequency of bankruptcy of the different classes.

There are differing motivations for defining internal rating systems with a higher versus a lower number of classes. Whereas a greater number of classes allows finer distinctions to be made between firms, a system with a high number of classes may lead to anomalies whereby the observed frequency of failure for higher-risk classes is lower than for lower-risk classes. For illustrative purposes, the analysis presented here makes use of seven risk classes. While the particular distribution was chosen to ensure that similar percentages of firms could be allocated from each model into each of the seven classes and to guarantee that, with only a small number of exceptions, bankruptcy frequencies of higher-risk classes are higher than frequencies for lower-risk classes, none of the article's qualitative results or conclusions depends upon the specific number

(1) The distribution used for illustrative purposes in this article is based loosely on the output of one of the three models.

of classes. Very similar results were obtained when using higher or lower numbers of classes.[1]

Table 1 reports the one-year and five-year bankruptcy frequencies for each of the seven classes for each model. One-year frequencies are reported for the 2001 and 2004 data, while five-year frequencies are reported for the 2001 data.[2] For each model, class 1 contains (roughly) the 1.4 percent of firms that are the least risky; i.e., with the lowest PDs or highest credit scores. Class 7 contains (roughly) the 3.3 percent of firms that are the riskiest; i.e., with the highest PDs or lowest credit scores. Classes 2 to 5 represent intermediate levels of risk and contain around 20 p.c. of firms each, while class 6 contains around 10 p.c. of firms.

Table 1 reveals that the one-year and five-year bankruptcy rates are generally increasing across classes and comparable across the three models, both in the 2001 and 2004 samples, which suggests that the seven classes reflect increasing degrees of credit risk. Table 1 also shows changing bankruptcy frequency rates over time. Due to the cyclical downturn in 2001, the percentage of bankruptcies in the entire sample was higher in 2002 (0.80) than in 2005 (0.56). The percentage of bankruptcies occurring between 2002 and 2006 was 3.50 (0.70 on annual basis) in the entire sample.

(1) Whereas Basel II requires having a minimum of seven buckets for non-defaulting borrowers and one for those that have defaulted, banks may work with internal ratings systems based on more than twenty-five buckets.

(2) The distribution of firms in the 2001 sample is slightly different for one-year and five-year bankruptcy frequencies. This is because a number of firms have exited the database between 2002 and 2006 for reasons other than bankruptcy (e.g. mergers, acquisitions, etc.).

TABLE 1 1-YEAR AND 5-YEAR BANKRUPTCY RATES ACROSS CLASSES FOR THE THREE MODELS

Class	Percentages of firms	Bankruptcy rates (percentages)		
		NBB	Model 1	Model 2
1-year bankruptcy rates (2001 sample)				
1	1.3	0.21	0.00	0.00
2	21.7	0.05	0.03	0.18
3	21.2	0.09	0.07	0.18
4	18.6	0.31	0.26	0.33
5	21.8	0.74	0.77	0.91
6	11.9	1.80	2.83	2.04
7	3.5	9.45	6.66	6.25
Total	**100.0**	**0.80**	**0.80**	**0.80**
1-year bankruptcy rates (2004 sample)				
1	1.4	0.00	0.00	0.00
2	21.3	0.00	0.01	0.06
3	21.7	0.10	0.07	0.11
4	18.9	0.23	0.17	0.21
5	22.0	0.33	0.43	0.56
6	11.6	1.46	2.00	1.11
7	3.3	7.74	5.64	7.00
Total	**100.0**	**0.56**	**0.56**	**0.56**
5-year bankruptcy rates (2001 sample)				
1	1.6	0.96	0.23	1.20
2	21.7	0.78	0.49	1.15
3	22.9	1.33	1.38	1.59
4	19.4	2.87	2.57	3.03
5	19.9	4.68	5.76	4.42
6	11.4	8.24	8.51	6.85
7	3.2	18.24	14.61	19.38
Total	**100.0**	**3.50**	**3.50**	**3.50**

2.1 Analysis of model disagreement

One way to use the mapping presented in Table 1 is to identify firms which are classified very differently by two models. Table 2a presents disagreement rates for high-risk firms (calculated as the percentage of class-7 firms of a given model which are classified in the median risk class, i.e., class 4, or below by another model). Table 2b shows disagreements for low-risk firms (calculated as the percentage of class-1 firms of a given model which are classified in the median risk-class or above by another model). Results are illustrated for the year 2004 (1-year predictions); however, results for 2001 (1-year and 5-year predictions) are quite similar.

A number of observations can be drawn from these tables. First, disagreement rates between the NBB model and Model 1 are much higher for high-risk firms than for low-risk firms.[1] However, comparisons of the NBB model with Model 2 and of Model 1 with Model 2 reveal that disagreement rates between these pairs of models are high for both high-risk and low-risk firms. The high disagreement rates between models suggests that if the pricing and origination of loans is a function of the class in which the firm is assigned, model choice can have a significant impact on loan decisions and pricing.

Of particular interest is whether disagreement between models is higher for firms in some industries than in others. This question is investigated for disagreements relating to the high-risk firms in the following way. For a given model, firms classified in class 7 are identified. Then the number of these firms that were classified by another model in classes 1, 2, 3 or 4 is calculated. These firms are called "outliers". Outliers are added for each pair of models in order to obtain the total number of outliers for that model pair. Finally, the distribution of outliers across industries is examined using a classification system based on NACE at 2-digit level (52 sectors represented). Table 3

(1) Note that this observation would still hold even if the disagreement rates for low-risk firms were doubled, to account for the smaller percentage of firms in class 1 than class 7.

TABLE 2a DISAGREEMENT FOR HIGH-RISK FIRMS (2004)

(percentage of class-7 firms of a given model classified as 1, 2, 3 or 4 by another model)

Class 7	Class 1, 2, 3 or 4	Percentages of class-7 firms
NBB	Model 1	16.7
	Model 2	13.8
Model 1	NBB	8.5
	Model 2	15.4
Model 2	NBB	18.0
	Model 1	18.9

TABLE 2b DISAGREEMENT FOR LOW-RISK FIRMS (2004)

(percentage of class-1 firms of a given model classified as 4, 5, 6 or 7 by another model)

Class 1	Class 4, 5, 6 or 7	Percentages of class-1 firms
NBB	Model 1	0.8
	Model 2	36.8
Model 1	NBB	0.7
	Model 2	21.0
Model 2	NBB	34.9
	Model 1	29.3

TABLE 3 INDUSTRY DISTRIBUTION OF OUTLIERS AND SAMPLE FIRMS, AND 1-YEAR BANKRUPTCY RATE OF SAMPLE FIRMS (2004)

(outliers = firms classified as 7 by one model and as 1, 2, 3 or 4 by the other model; industry classification system based on 52 sectors)

Industry	Outliers distribution (percentages)			Sample firms	
	NBB and Model 1	NBB and Model 2	Model 1 and Model 2	Distribution (percentages)	Bankruptcy rate (percentages)
Construction	25.6	16.8	36.5	16.2	0.71
Wholesale trade and commission trade	9.2	9.6	5.5	15.0	0.45
Retail trade	7.5	7.3	6.7	9.8	0.46
Hotels and restaurants	11.1	9.6	9.1	5.2	0.91
Land transport	4.9	6.5	4.1	5.9	0.77
Miscellaneous business activities	9.5	13.2	8.9	8.8	0.54
All other sectors	32.1	37.0	29.1	38.8	0.64

reports the distribution of outliers and sample firms across industrial sectors accounting for at least 5 p.c. of the total number of outliers, as well as one-year bankruptcy rates of sample firms.

Several observations emerge from the table. First, the share of outliers in the construction industry is significantly higher than the share of sample observations in this sector. Indeed, the share of sample firms that are in the construction industry is only 16.2 p.c., whereas the percentage of outliers in this sector runs as high as 36.5 p.c. for certain model comparisons. Second, the share of outliers in the hotels and restaurants and miscellaneous business activities sectors is also higher than the share of these sectors' firms in the sample (5.2 p.c. and 8.8 p.c., respectively). Third, the share of outliers in the retail trade and wholesale trade sectors is systematically lower than these sectors' shares of sample observations (9.8 p.c. and 15 p.c., respectively). Finally, bankruptcy rates are among the highest for the construction and hotels and restaurants sectors, i.e. two of the sectors whose share of outliers is higher than their share of sample firms. This suggests that it may be more difficult to identify the "true" financial condition of firms in industries with higher bankruptcy rates.

2.2 Analysis of model power

The above analysis is interesting in terms of its implications for loan origination and pricing decisions; however, it suggests nothing about the "power" of the three models; i.e., their ability to distinguish between failing and non-failing firms. Analysing the power of a model requires comparing its output with actual bankruptcy data. In this section, the power of failure prediction models is compared using Receiver Operating Characteristic (ROC) curves (see Box 2 for a description of ROC curves). Charts 1 and 2 present the ROC curves for one-year and five-year bankruptcy rates based, respectively, on 2004 and 2001 data. Table 4 reports the area below each of the curves.[1]

(1) The ROC curves for which the areas have been calculated are ROC curves for the 7-class rating system, rather than the curves for the raw output of the three models (credit scores or PDs). The area under the ROC curve varies very little (around .02) when using the classes of a rating system instead of the actual output values of the model.

Box 2 – ROC curves

The ROC (Receiver Operating Characteristic) curve is frequently used when comparing the accuracy of credit risk models. It is constructed by first ordering the non-failing firms from worst (highest risk) to best (lowest risk) from left to right on the horizontal axis. The vertical axis represents the percentage of all failing firms that would be captured at each percentile of non-failing firms on the horizontal axis. In other words, if *x* p.c. of non-failing firms (starting from the riskiest firm) were excluded from the sample, the vertical axis of the ROC curve gives the percentage of failing firms that would also be excluded (because they are ranked as equally risky or riskier than the least risky excluded non-failing firm).

ROC curves allow calculation of Type-1 and Type-2 errors at each point on the curve. The Type-1 error, or the error of labelling a non-failing firm as failing, corresponds to the percentage of non-failing firms excluded. The Type-2 error, or the error of labelling a failing firm as non-failing, equals the percentage of failing firms that is not excluded from the sample.

When the ROC curve of one model lies strictly above the ROC curve of another model (i.e., to the northwest), the former has unambiguously a lower Type-2 error rate for any given Type-1 error rate. When the ROC curves for two models cross, neither strictly dominates the other. In this situation, which model would be preferred would depend on the specific application one is interested in.

A convenient measure for summarizing the graph of the ROC curve is the area under the curve, which is calculated as the proportion of the area below the curve relative to the total area of the unit square. The area under the ROC curve may range from 0.5 (model with random classification) to 1.0 (model with perfect discrimination). The area may be interpreted as the probability that a randomly chosen failing firm is classified in a riskier class than a randomly chosen non-failing firm (Stein, 2002). Models with an area under the ROC curve between .7 and .8 are

often regarded as having "acceptable" discriminatory power, while models with an area between .8 and .9 can be considered as having "excellent" discriminatory power (see Hosmer and Lemeshow, 2000).

CHART ROC CURVE ILLUSTRATION

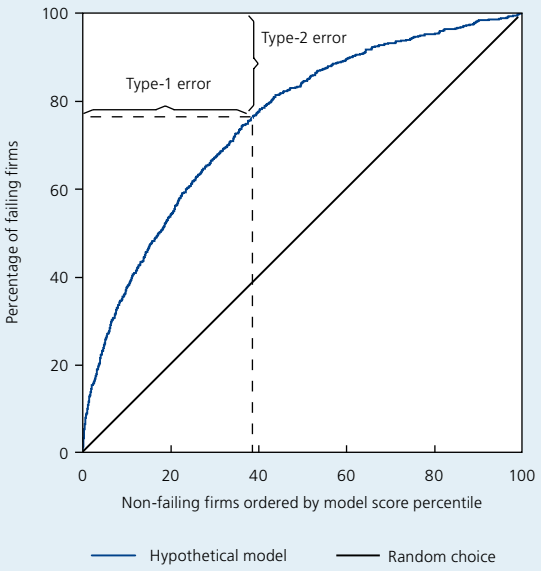

CHART 1 1-YEAR ROC CURVE OF THE THREE MODELS BASED ON THE 7-CLASS SYSTEM (2004)

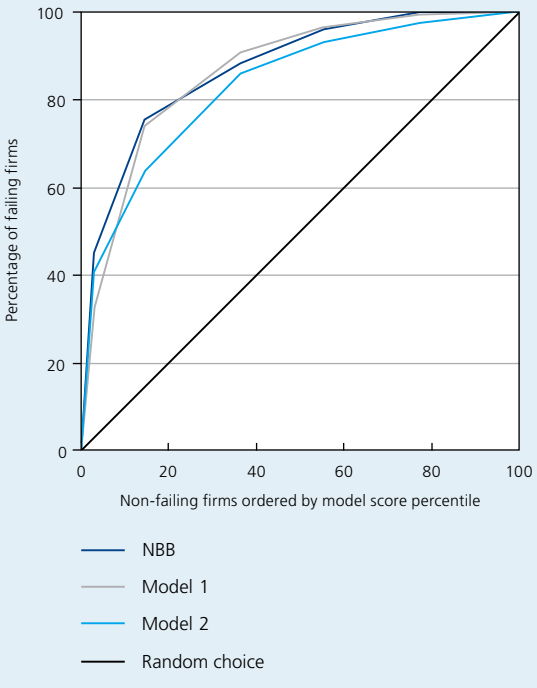

CHART 2 5-YEAR ROC CURVE OF THE THREE MODELS BASED ON THE 7-CLASS SYSTEM (2001)

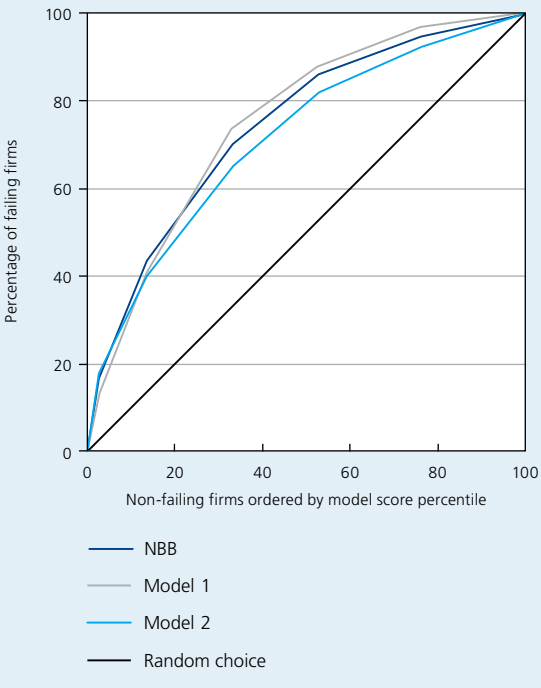

TABLE 4 AREA UNDER THE 1-YEAR AND 5-YEAR ROC CURVES

Model	1-year ROC (2004)	5-year ROC (2001)
NBB	0.873	0.743
Model 1	0.866	0.753
Model 2	0.834	0.714

Charts 1 and 2 illustrate that all three models perform considerably better than would a random selection of firms, which is depicted by the 45 degree line. In fact, the values of the area under the ROC curves given in Table 4 for each model would suggest that all three models exhibit excellent performance with respect to prediction of one-year bankruptcy rates. Not surprisingly, the three models perform considerably better at the one-year than at the five-year horizon. Despite this excellent performance, Model 2 appears to perform somewhat less well than the NBB model and Model 1, both at the one-year and five-year horizons. For instance, with respect to the one-year ROC curves, we see that if the failure/non-failure cut-off was placed at the level of the twenty percent of non-failing firms with the lowest scores or highest PDs, the NBB model and Model 1 would pick up roughly 80 p.c. of all failing firms, compared to 70 p.c. for Model 2.

Chart 1 also raises the question as to whether the relatively high levels of disagreement between models reported in Tables 2a and 2b might be related to the shape of and area under the ROC curves. For instance, the NBB model seems to be slightly better than Model 1 at differentiating failing from non-failing firms among the riskiest firms, i.e. firms falling below the 15th percentile of the distribution of non-failing firms, whereas Model 1 seems to perform slightly better than the NBB model for medium and low-risk firms. As a result, the overall performance of both models (as measured by the area under the ROC curve) is very close.

The excellent, and relatively similar, performance of the three models considered here is important, given that the statistical methods as well as the measure of failure (bankruptcy vs. default) used to estimate and calibrate the models are different. The latter result may suggest that the definition of failure used for model estimation and calibration does not matter as much as one might have expected. This observation, however, may be more true for Europe than the US. In fact, a much higher percentage of defaulting firms in Europe ultimately enter bankruptcy than in the US (Korablev, 2005).

Conclusion

The development of Basel II has stimulated a number of vendors to develop a range of products including failure prediction models, which assess the risk of failure of obligors. These models are often used by banks as an off-the-shelf product or, alternatively, as a basis for development and benchmarking their own internal rating systems.

This article reviews the main characteristics of failure prediction models for private firms offered by vendors and compares the output of two vendor models with a failure prediction model developed by the NBB. The analysis, which makes use of bankruptcy data for Belgian firms, focuses on the extent of disagreement in firm rankings across models and on differences across models in the ability to predict firm failure.

A first finding is that there is considerable disagreement in firm rankings among the three models studied, and the extent of disagreement (e.g., firms classified as very high risk by one model but low risk by another) can be quite important. Since banks use failure predictions not only as inputs in the calculation of their regulatory capital but also in several areas of credit risk management (e.g. loan pricing and loan origination), this result suggests that a bank's model choice may have a significant impact on its loan granting and pricing decisions.

A second finding is that, overall, the three models under consideration perform similarly, and quite well, in distinguishing bankrupt and non-bankrupt firms. This result, which can be considered reassuring from a financial stability viewpoint, is interesting in that the statistical methods as well as the measure of failure (bankruptcy vs. default) used to estimate and calibrate the models are different.

Finally, the high rates of disagreement observed among the three models studied, together with the excellent performance of each, suggests that there may be some benefits for banks in combining failure assessments of different models. The idea would be that the output of a failure prediction model represents a "signal" about the credit worthiness of a firm and, given that the signals produced by different models are not perfectly correlated, performance should be improved by making use of the combined information from multiple signals. One avenue for future research is to investigate the extent to which there is a trade-off between combining the output of more models versus using fewer models with superior performance.[1]

(1) For further investigation of this and other issues, see Mitchell and Van Roy (2007).

References

Balcaen S. and H. Ooghe (2006), "35 Years of Studies on Business Failure: an Overview of the Classic Statistical Methodologies and their Related Problems", *British Accounting Review*, Vol. 38 (1), 63-93.

Hosmer D. W. and S. Lemeshow (2000), "Applied Logistic Regression", John Wiley and Sons, New York.

Korablev I. (2005), "Power and Level Validation of the EDF Credit Measure in the European Market", Moody's KMV.

Mitchell J. and P. Van Roy (2007), "Failure Prediction Models: Disagreements, Performance, and Credit Quality", National Bank of Belgium working paper (forthcoming).

Stein R.M. (2002), "Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation", Moody's KMV Technical Report # 020305.

Vivet D. (2004), "Corporate Default Prediction Model", *Economic Review,* National Bank of Belgium, December, 49-54.