8

Credit Scoring for Corporate Debt

Eric Falkenstein

Deephaven Capital Management

The term credit scoring refers to quantitative methods for evaluating the credit quality of companies. These methods are generally nonparametric, as opposed to the structured approaches of theoretical models such as Merton (1974) or Jarrow and Turnbull (1995). This tool is mainly used by lenders who are evaluating a portfolio of tens of thousands of loans, but they are also making headway into bond trading operations where portfolios include only a few hundred exposures. If you rank order companies by their scores, companies at one extreme should fail at higher rates than those at the other extreme (failure usually, but not always, means default).[1] Historically most scoring has been applied to consumer credits, such as credit cards and home loans – where it now dominates the underwriting process – but it is starting to be applied to businesses with revenues above US$10 million (eg, corporate or commercial loans). It is this latter objective that we focus upon here.

Introduction

Credit scoring is a quantitative exercise that is refreshingly productive. A good scoring system can save lenders money and time, and can be a first-order competitive advantage. The reason good scoring is helpful, of course, is that so many credit infrastructures are manifestly sub-optimal. The average credit department within a bank or bond trading desk generates a miasma of information in its combination of a book report, ratio analysis, projections and competitive conventions. Such unstructured analysis can be indulged indefinitely without producing results meaningful to outsiders, which is why commercial credits with only internal ratings either have to be re-underwritten or have existing agency ratings in order to be sold at virtually any institution.[2] In contrast, scoring done properly has credibility outside an institution. Scoring makes credit analysis succinct and unambiguous, and generally is most efficient at its main objective, distinguishing high risk companies from low prior to default.

Modelling credit risk is a deceptively simple problem. Unlike in equity markets where it is difficult to find a metric of risk that consistently relates to future returns, firms with high default risk have very measurable and theoretically straightforward characteristics: low profitability, high leverage, low liquidity, small size, high volatility, high inventories and extreme growth. The biggest problem that most models have is trying to do too much, not content with the patent incompleteness of, say, a 3% default probability based on five financial ratios (eg, income/assets, current liabilities/assets, etc.). There is an inevitable tendency to add many more nuances, not one but several profitability measures such as EBITDA, net income, and gross margin, or to focus on a credit cycle component that puts as much weight on GDP projections as on the firm's cross sectional risk. There is also a tendency to add more

complex functional relationships such as a neural network or fuzzy logic algorithm.[3] Smart people with the best intentions are wont to make deleterious refinements to them in order to add value: the road to poor out-of-sample performance is paved with good intentions. Specifically, predicting the future is hampered by the law of unintended consequences, which states that unanticipated indirect effects often counter and outweigh the intended direct effects. After a very short while, the more one does to improve a model, the greater the probability that one actually diminishes model performance.

This chapter will outline the major concepts involved in credit scoring. It reflects the author's biases, but perhaps it is more useful to hear strong opinions than equivocating judgments because although there are many ways to construct a scoring model, most are arguably roads to obscurity. Historically there have been many models that have failed to garner significant market share in spite of large contemporaneous advantages in credibility, data, or connection to rating agencies, proving that it is not easy to create a viable model. The author's perspective prioritises modeller's incentives, because understanding these incentives help us to understand why certain evaluative prejudices are justified.[4] Some of these insights have been presented in the author's previously published work while based at Moody's, and the interested reader should consult the Moody's RiskCalc collection as a useful set of papers on commercial credit issues.[5] This chapter is not a comprehensive discussion of scoring. Several important technical matters are addressed only briefly, such as correcting common data sample biases, calibration of outputs to default probabilities (DPs), the use of a multinomial rather than binomial estimation, integration with recovery rate estimates, and integration with business cycle dynamics. Further, the application of scores to pricing, incentive compensation, capital allocation and provisioning is nontrivial and important.

OVERFIT

The concept of being 'overfit' and the related concept of 'degrees of freedom' will be frequently mentioned below, and so they deserve an explicit definition here at the outset. 'Overfit' refers to any mathematical model that fits a set of data very well, but, because it is 'overfit', it models not just fundamental relationships but also random patterns in the sample data.[6] Since random patterns are not expected to recur, fitting these patterns gives spurious picture of power, and implies a worse out-of-sample fit than a more limited model for non-obvious reasons that are explained below. 'Degrees of freedom' is a statistical concept that refers to the number of observations in a data set minus one, and it also can be applied to models: the data give degrees of freedom and a model (or any explanation) uses them up.[7] To give an example, if Hans is taller than Franz (sample size of two), any one arbitrary metric (eg, age) for which the two differ can perfectly 'explain' – in a statistical sense – why Hans is taller than Franz. There is one degree of freedom, and so a model with one degree of freedom (ie, one variable) can perfectly explain the data. If we had three people, there are two degrees of freedom and it would take two variables to perfectly 'explain' or 'fit' the data, and so on. Having many degrees of freedom left over (on the data set minus that in the model) is helpful because it lessens the probability that the model fits the data by accident. When the data sample minus the model leaves few degrees of freedom, there is a greater chance one will find only a meaningless, parochial and ultimately ephemeral correlation.

Measuring degrees of freedom is complicated by two factors. First, interdependence in the data observations reduces degrees of freedom in a subtle way. For example, if all 100 firms in one sector default, this allows one to 'explain' or 'fit' the sample data by pointing to that one attribute. One can explain all the data with one attribute because there is effectively one degree of freedom in this case. This is extreme, but it explains why even 500 observations are not as informative as they appear – correlations lower the effective degrees of freedom considerably. Secondly,

the concept is further clouded by the fact that many degrees of freedom are used up in the selection process of the functional form and inputs. A model with only five inputs which were chosen after examining 100 different inputs, uses all 100 degrees of freedom even if 95 did not end up in the model. This makes it very difficult to judge the number of degrees of freedom in a final model because one does not know how many different inputs and functional forms were tried. Overfitting the data is perhaps the greatest modelling problem in commercial credit because there are so few degrees of freedom in most datasets, and it is not obvious which variables or functional forms to use.

QUALITIES OF A GOOD SCORE

A good model simply aggregates objective information in a statistically optimal way. Comparing one's model to extant models is humbling, requiring honesty and education – one must not only be aware of the best alternatives, one must model these alternatives with as much care as they apply to one's own. It is tempting to be less conscientious when constructing and testing a benchmark, thereby making one's innovation appear more valuable. Simplicity is helpful because it protects against overfitting, encourages transparency, and sublimates the ego of the modeller to the objective.[8] A good model is modest because modesty is implied by honesty and knowledge of the many other good models out there (see Shumway, 2001). In contrast, most models are overly-sophisticated and are presented using biased datasets, though in their defence these features are necessary when one is selling to users who can generally overestimate the efficacy and complexity of their own credit culture.

The most important reflection of a model's value is whether it is being used to make business decisions such as pricing or decisioning, as opposed to the less consequential expositions of risk to regulators or other outsiders. The following criteria are means to that end: transparency, power, calibration and validation. Each is necessary and none is sufficient.

Transparency

Transparency gives users confidence, which is especially important when trying to convince new users to adopt a model. Transparency also ensures against the possibility of a particularly inexcusable modelling vice, *viz*, those who use complexity to avoid having to defend the logic of the model; modellers know and sometimes exploit the fact that outsiders can be naturally wary of criticising what they do not understand. Most users prioritise transparency above all other attributes, and so modellers should be aware that unless they are producing tools for themselves, they should consider the importance of transparency first.

Power

Power is a technical term that implies the ability to separate good firms and bad firms in the following sense: firms with high scores should default at higher rates than those with low scores. The more extreme the difference in default rates between low and high scoring firms, the more powerful the model.

Validation

Validation moves models from interesting to actionable. While quantitative people think of validation in terms of statistical tests, most validation is actually implicit, based on a history of usage within the organisation and demonstration of the particular model's power. This is an important factor in why the rating agencies such as Moody's and Standard & Poor's are so pre-eminent, because they have credibility from the validation implicit in the history and usage of their ratings. If one wants to compress this otherwise multi-decade approach, there are statistical tests that demonstrate power, and these are especially convincing when used on out-of-sample datasets.

Calibration

Calibration means aligning model output not just with an ordinal ranking, but a cardinal number that allows for an exact proportional relationship. The most useful output is a default probability that is often stated in annualised terms, however, recent models have been developed using credit spreads or their expected volatility for a company, and these are also directly measurable for certain firms, as well as very relevant.[9] A grouping labelled one through ten can be powerful, but mapping these buckets to default probabilities (DP) or credit spreads, or replacing buckets altogether and using DPs directly, eliminates an otherwise arduous step in a model's application. Calibration is complicated by the highly cyclical behaviour of defaults, where recessionary periods have five times the default rate of non-recessionary periods. Since recent recessions in the US have occurred at 10-year intervals, this makes the testing of a 'through-the-cycle' default rate virtually impossible. This does not make validation of a DP impossible. It is just not as straightforward as might be expected when one has several years of data with thousands of defaults; time diversification is something one needs to adjust for explicitly (see Cantor and Falkenstein, 2001).

THE QUANTITATIVE REVOLUTION

Credit analysis is almost as old as civilisation itself, but there is a true paradigm shift that began around 1990 due to a more quantitative focus. Historically, credit analysis had experienced little progress due to a lack of data. Academic studies in the twentieth century generally used less than 100 defaults – too few to develop models better than the most naïve alternatives – and even within Moody's circa 1990 few analysts had a sense of the historical or prospective annual default rate for B-rated loans (ie, what we now know to be around 6% on average). The problem was nicely illustrated by the 'junk bond' debate in the 1980s: what were the expected returns, and therefore default rates, on these instruments? Financier Michael Milken argued that, historically, non-investment grade bonds had positive risk-adjusted returns, and the debate went nowhere because he could not be proved right or wrong (a bit like arguing the feasibility of socialism in 1900): almost all you had was theory and anecdote. Without data to test and evaluate differing theories, progress in analytics is arrested in a literary state.

In the last decade, however, data has become to define credit quality. Moody's and S&P have started to generate historical and monthly updates of default rates by credit rating, so that instead of thinking about risk as high or low, it is now, say, 0.15% annualised default rate (BBB-rated) versus a 1.5% annualised default rate (BB). This allows for confidence in creating and syndicating collateralised bond obligations (CBOs), where now the rating agencies are forced to take a stand on how ratings relate to default and recovery rates to avoid ratings inconsistency. The BB rating of a CBO has to have the same default rate as the BB loans that underlie it, and thereby necessitates an accurate forecast of those default rates.

Other pressures have been at work too. The new Basel capital accord focuses credit on measurement. KMV, before being bought by Moody's, argued that statistically their Merton model's expected default frequency (EDF) was statistically more powerful than agency ratings, a quantitative assertion that invites competition and measurement that did not happen before. Moody's RiskCalc now has over 3,000 defaults underlying their model parameters – this simply takes a model into a different league than the old models of academia that used less than 100 defaulting firms.

Data and quantified risk bucketing invites statistical analysis, whereas previous qualitative distinctions do not. A DP suggests that a credit opinion is not like a book review, but instead is objective and testable; it has precise empirical implications. Henceforth, ratings have clear empirical implications, and this allows one to compare to them directly. Moving from the literary review to an empirical issue takes credit analysis from a soft science to a hard one, where scoring methods can compete and thrive.

INDUSTRY, MARKET AND COUNTRY DIFFERENCES

The rating agencies have many different departments, such as sovereign, commercial mortgage backed securities and municipalities. These distinctions are made because each loan type looks at very different sets of information. While we are focusing here only on corporate non-financial credit risk, the general modelling approach applies to these other areas as well.[10]

Within the corporate lending category, credits trade in three regimes: investment-grade, high-yield and distressed. While technically, these refer to ratings above Ba1/BB+ and Caa1/CCC+, this author prefers the distinction that investment-grade loans are quoted at spreads to treasuries, non-investment grade (or high-yield) bonds are traded using bond prices or spread to Libor, while distressed bonds trade on a recovery rate basis (ie, independent of coupon and maturity).[11] It is the high-yield and unrated private companies that are mainly targeted with scoring models, even though investment grade companies can be mapped into similar-looking DP buckets through more qualitative techniques. Distressed companies are already in some stage of workout, and a DP is a second-order consideration relative to its recovery rate. Investment grade companies, on the other hand, provide a subtler problem.

Scoring generally is not useful for investment-grade companies, and they create misleading inferences for credit modelling. Correlations that exist for the hundreds of thousands of regular corporations disappear or go the wrong way for the thousand or so non-financial companies *worldwide* considered investment-grade. For example, because of commercial paper access higher quality investment-grade firms have lower liquidity, the exact opposite of what is observed for non-investment grade company. A naïve researcher would infer then that liquidity is either inversely related to financial strength or, given the opposite relation for non-investment grade companies, at best ambiguous. In fact, liquidity is a powerful model input for non-investment grade companies, and the lack of commercial paper access explains the apparent ambiguity rather neatly (this is just one example; there are other perverse issues). Thus, in spite of their high profile and ability to get much high quality data (including bond prices of differing maturities and subordination), they constitute a major selection bias, being highly unrepresentative of non-investment grade companies. What works best for companies whose bonds are quoted at spreads to treasuries is very different than what works well for the 99.9% of companies not in that elite class. Lastly, as of 2001 there have been only about 30 defaults within one year for these companies, too few to model these companies with much confidence.[12]

Different countries have different bankruptcy laws, tax rates and accounting conventions, which implies that the relation between financial ratios and default rates should vary significantly between countries. While ideally one should re-estimate and (more importantly) recalibrate based on different countries, the differences are not as considerable as one would think, and can mainly be addressed through a simple renormalisation of the inputs (ie, changing the input functions so that the 10th percentile of the profitability in, say, Italy, maps to the 10th percentile in the US). It shouldn't be surprising that regardless of industry or country, firms with high profitability and low leverage tend to default at lower rates. Also, the differences in private companies between countries tend to be much greater than for public companies.

Industry distinctions within the non-financial sector offer a dilemma. There simply is not enough data to allow one to separately estimate models for each industry group. If data limitations were not a problem, separately estimating a model for each industry would be a good idea, but in practice one loses too much information for this to work. While normalising financial ratios for an industry sector is usually a benign adjustment, it can be counterproductive if, for example, some industries have higher default rates because of their higher-than-average leverage (and thus risk). This is a real distinction between scores and traditional analysis,

because traditionally credit analysts focus on tightly defined peer groups and thus implicitly ignore this issue. A lot of this could be said to be simple posturing, people trying to manufacture the aura of specialised expertise by emphasising idiosyncratic conventions in particular sectors, as well as the sincere but misguided belief that more granular distinctions are always valuable.

THE MERTON MODEL, A SUFFICIENT INPUT?

A special case for inputs is whether or not a particular structural model is a sufficient input, and would make other inputs redundant.

Structural models make simplifications by definition, and therefore are potentially incomplete. While the Merton model makes a compelling case for its particular focus, there are equally compelling cases for other structural models; examples include the academic gambler's ruin problem (Wilcox, 1977), which is based on cashflow as opposed to equity value, but more common are the thousands of nuanced structural models, many based on cashflow, implied by an experienced practitioner's 'expert rule' system and targeted towards a particular industry sector (see Stentin, 1987; Casey and Bartczak, 1984). How else to reconcile these approaches than to examine, empirically, how well they or their parts work together?

The lack of precedence for a successful structural model might inform one's initial skepticism. Statistical tautologies and arbitrage relationships can work quite well, everything else might be seen as just a heuristic (eg, structural macro models, the quantity theory, or the capital asset pricing model (CAPM)). The data seen by the author simply confirms these prejudices and provides high confidence that a Merton model is an incomplete measure of cross-sectional default risk (see Boral and Falkenstein, 2001; Sobehart and Stein, 2000). Data used for such tests are generally proprietary, however, which makes it difficult for an impartial outsider (ie, most readers) to adjudicate. For any readers interested in testing this issue, the test is simple: group firms into sufficiently small Merton bands (small enough to consider each firm similarly risky, yet large enough to generate a sufficient number of defaults). Break these groups into high and low groupings of some additional variable, such as book leverage or profitability. See if the default rate trends across this new dimension.

In general sense, however, this author is a Merton advocate: the non-structural model that will be outlined in this chapter has much more in common with a Merton model than the qualitative approach that dominates most credit departments. Merton proponents and alternative models are defining performance the same way, making reconciliation (or submission) foreseeable. In contrast, many qualitative modellers actively discourage statistical comparisons, and there is comparatively little common ground. Further, one of the really nice things about Merton's model is its transparency and parsimonious structure. It may not be optimal, but the greatest modelling error is over-engineering; over-zealous allegiance to a simple structural model like Merton's is preferable. An incomplete Merton model is better than what most modellers would create because of modeller incentives to create overly sophisticated and overfit models.

THE SCORING ALGORITHM

There are four main steps in scoring:

1. choosing inputs;
2. transforming inputs;
3. combining transformed inputs; and finally
4. mapping the output into default probabilities.

We will consider them in turn. Scoring is engineering and not pure science, modellers should aspire to be like Wernher von Braun rather than Albert Einstein.[14]

PANEL 1

DRIVERS OF A DEFAULT RISK

There are literally hundreds of ratios and interactions that are potentially interesting, but seven factors seem crucial (Chen and Shimerda, 1981). Given the noise in financial data, the author favours the broadest concepts so that there is no need to be dependent upon meticulous and accurate accountants (eg, earnings before interest and taxes (EBIT) versus earnings before interest, taxes, depreciation and amortisation (EBITDA) minus capital expenditures plus leases).
The following are the most powerful inputs for predicting default:

1. *Volatility:* higher equity volatility implies higher probability of a firm's asset value falling below its level of debt, which implies insolvency. This is only measurable for public companies.

   Examples: historical volatility, option implied volatility.

2. *Size:* for non-traded companies, size proxies for much of equity volatility. Bigger companies are generally more diversified in their exposure to geographies, products, and people, and this lowers their prospective volatility. One could argue that size is truly a different factor, and there's some truth there, though this can easily get into hair-splitting.

   Examples: market cap, sales.

3. *Profitability:* higher profits lowers default probabilities. Combining profitability with interest expense makes it a combination of leverage and profitability.

   Examples: net income/assets, EBIT/interest, EBIT/Debt.

4. *Leverage/Gearing:* higher leverage implies higher default probabilities. Higher market valuation implies a greater distance between a firm's asset value and its level of debt

   Examples: debt/assets, market capitalisation/debt, total liabilities/total assets.

5. *Liquidity:* lower liquidity (current assets/current liabilities) implies higher default probabilities in all countries, though this effect is reversed for those 3,000 or so investment-grade companies.

   Examples: current assets/current liabilities, short-term debt/total debt.

6. *Growth:* both high and low growth rates are associated with higher default probabilities.

   Examples: year-over-year sales growth.

7. *Inventories:* higher inventory levels imply higher default probabilities.

   Examples: inventory/sales.

Scoring is a sequence of analytics, each step of which requires thoughtful consideration, as opposed to one elegant and unambiguous equation. The approach advocated here is a semi-parametric approach, using non-parametric estimation for univariate relationships and parametric estimation for the multivariate relationships. The approach is therefore a generalised linear model in the transformed inputs,

meaning that the algorithm has at its core an additively separable set of inputs; almost all the non-linearity is univariate.

Transforming inputs

This is an area where conventional statistics can learn from the neural net literature such as Golden (1996). Statistics books are relatively silent on this part of estimation, while the neural net literature takes great care in transforming inputs so that results are robust, primarily because they are engaged in an algorithm that is highly susceptible to overfitting and local maximums, whereas traditional statistical approaches assume that one 'knows' the functional form one is trying to test. Transformations of this sort are essential because financial ratios are highly skewed and fat-tailed, which causes a few observations to overly influence the output if not transformed. For example, one would never use 'assets' as a raw, untransformed input because they are lognormally distributed among firms. Transformation methods include:

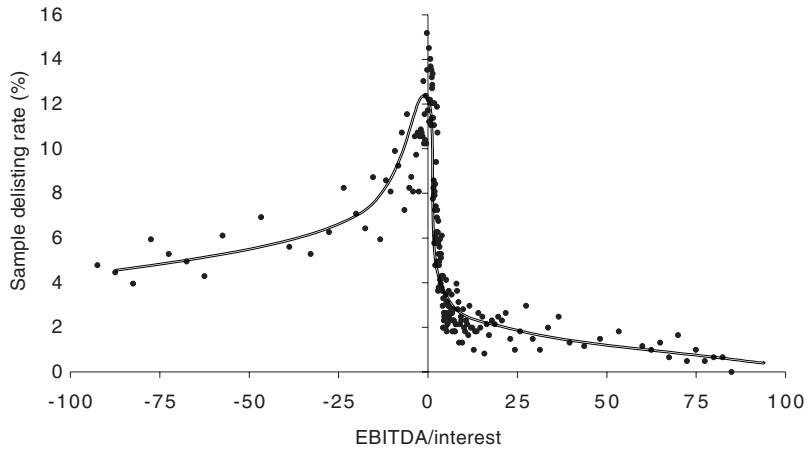
- ❑ replacing the ratio with its percentile;
- ❑ turning the ratio into a standardised Gaussian variable (eg, $\frac{x-\mu}{\sigma}$) or
- ❑ applying a variety of sigmoidal functions (eg, $\frac{1}{1+e^{-x}}$), to the ratio or its standardised gaussian variable;
- ❑ using the nonparametric univariate default estimate generated by each variable.

The idea with all of these is to generate more Gaussian-distributed input so that outliers do not dominate the estimation. The author's favourite transformation is the latter, to use the non-parametric relation between the input and default rates, since by construction each input is an unbiased estimator of the output, and then the next step involves finding the right weights for a set of unbiased but correlated estimators, a generally stable and intuitive approach. Also, by transforming in this way one can avoid problems caused by non-monotonic relationships. For example, firm growth rates are related to future default rates in a U-shape, extreme negative or positive growth correlated with higher default rates, and so although a linear univariate relationship with default is zero, this factor does show up with the univariate default-rate transformation. Most importantly, monitoring the univariate relationship between an input and default, because it maps naturally into a graph, is very helpful in noticing changes in relationships, and gives one greater intuition than if one is instead monitoring the efficacy of a third-order equation (eg, $0.219 - 7.324x + 9.43x^2 - 2.439x^3$) applied to a standardised ratio. Let us examine an example of this transformation method.

Looking at Figure 1, the data from the univariate relation between the interest coverage ratio and delisting rates suggests a highly nonlinear and relatively rare non-monotonic relationship. In this case this is because as the ratio moves below zero due to a negative numerator (EBITDA), highly negative ratios are more due to low levels of interest expense as opposed to very low cashflow. The resulting transformation function would simply be the nonlinear function that best fits this data, and is reflected by the curve in Figure 1 (estimate using a nearest-neighbour kernel approach). In fact, the author prefers to use a simple look-up table based on the curve generated, as opposed to mapping the relationship into a closed-form algebraic function. Note that a parametric representation of this would be difficult to uncover, especially without 'seeing' the univariate relationship directly.

Table 1 is an example of a transformation based on the univariate relation of one risk factor proxy, EBITDA/interest, to delisting a transformation taken directly from

1. Data using Compustat US non-financial firms, 1980–99, 200 bins and resultant equity delisting date over subsequent five years



the data that underlies Figure 1. In practice one should use a more granular table (say 50 buckets instead of 10). Also, the units of the transform do not matter much, because the multivariate estimation will adjust for the optimal weighting, but this author prefers to normalise each univariate transformation to have similar means in order to have a more stable multivariate estimation procedure, and just to keep things consistent at this stage.

Choosing inputs

One should avoid getting mired in highly specialised and philosophical distinctions about 'cashflow'. It is useful to know that such minute distinctions rarely help on average, in spite of their usefulness in any one case.

Table 1. Univariate transformation corresponding to Figure 1

EBITDA/int	transform
–37.25	0.6629
–4.03	0.8215
0.15	1.2116
1.60	0.7594
2.65	0.5895
3.84	0.4901
5.39	0.4329
7.79	0.4044
12.66	0.3742
30.40	0.3471

The variable selection process consists of the following exercise. First, find the most powerful ratios within each of the risk factors listed above (being a univariate exercise this is straightforward), and perhaps augment this base with your favourite structural model (eg, Merton). Take this base set and see which are significant in a standard parametric multivariate model, with a strong bias towards elimination. Do all this with the transformed ratios in order to avoid non-linearities obscuring an input's power or enhancing a correlated ratio's power. Sequentially add ratios and see if they retain statistical significance, with the expected sign. Usually, the more powerful risk factor ratio, such as net income/assets (NI/A), when used with a similar, correlated measure such as net sales margin, will generate coefficients where the more powerful ratio, NI/A, has a positive coefficient and the less powerful ratio has a negative coefficient. Do not use the additional ratio if it contains a 'wrong sign' or if it is statistically insignificant.

The 'wrong sign' problem is an especially useful exclusion device. For example, given

$$y = \beta_1 x_1 + \beta_2 x_2$$

the coefficient on x_1 is simply

$$\beta_1 = \frac{\rho_{1,y} - \rho_{1,2}\,\rho_{2,y}}{1 - \rho_{2,y}^2}$$

The coefficient will be negative if $\rho_{1,y} < \rho_{1,2}\rho_{2,y}$. That is, if the geometric average of the correlation between regressors is greater than the correlation between the regressor and the predicted variable, it will have the wrong sign. For example, if $\rho_{1,y} = .3$, $\rho_{2,y} = .8$ and $\rho_{1,2} = .4$, then the sign of β_1 will be 'wrong', ie, it will be negative in a multivariate context but positive in a univariate relation.[15]

Adding regressors always increases in-sample fit (R^2), and has also been documented to increase user confidence in the result, creating a large incentive to overfit a model by using too many inputs.[16] For example, instead of using just one measure of profitability, why not several? Thus, one is looking at net profit margin, gross profit margin, net income/equity and EBIT/interest, as well as the trends, levels and historical averages for these ratios. It is truly a Faustian bargain, since the benefits of appealing to user intuition (which is to look at as much information as possible), combined with an opportunity to fit the data better (rarely are out-of-sample tests truly out-of-sample), are immediate and certain. The costs, though uncertain, are rarely worth it for the following reason: the standard errors in a multivariate model are proportional to $1/(1\text{-}\mathrm{corr}(x_1,x_2))$, so the higher the correlation with existing information, the higher the standard errors for the individual coefficients, which implies a more precarious out-of-sample performance.[17] Higher standard errors in this context are not an innocuous effect that diversifies away, it degrades model performance, generating worse fit on new, or out-of-sample, data. Paying attention to everything creates confusion, in statistical or personal matters, and the ultimate in confusion is a non-informative interpretation. Further, the wrong sign problem caused by including two highly correlated inputs is not just meaningless partial derivatives, but the generally poor out-of-sample implications: it places great demands by requiring that the correlations among the predictor variables remain stable, which is asking a lot from a problem already severely constrained by a lack of data. There is always a trade-off in adding new information, one that becomes more biased in favour of exclusion as one adds more and more inputs. A bias towards parsimony is therefore rare but wise.

This is the stepwise process of variable selection, suggested by the univariate power, validated in a multivariate context.

Combining different inputs

At one point this was the main focus of the commercial credit analytic problem: highlighting the particular advantages and disadvantages of a particular binary model versus some alternative. For example, Altman's seminal paper (Altman 1968) highlights discriminant analysis, and several authors have documented the improvement from going to logit in this context (Ohlson, 1980). While a useful improvement, it is really a second-order efficiency compared to choosing inputs, transforming these input variables, and mapping to a meaningful output.

The early popularity of discriminant analysis (DA) owed more to the ease of computation – matrix algebra vs maximum likelihood estimation – than its predictive superiority. Logit and probit models are now preferred since most statistical software contain these algorithms as standard packages. DA used to be easier, but now software programs make them equivalently demanding, and logit relaxes some of the severe assumption needed for DA, including homoskedasticity and Gaussian distributions for both the 'good' and 'bad' observations.[20] The choice between logit and probit is less important, as both give very similar results (especially after the

transformations suggested, which diminishes the number of outliers). Logit gained popularity earlier primarily because in the early days of computing anything that simplified the maths was preferred. In this case, finding the maximum likelihood is easier for the logit density function $1/[1 + \exp(-x)]$ than the probit's Gaussian density function.

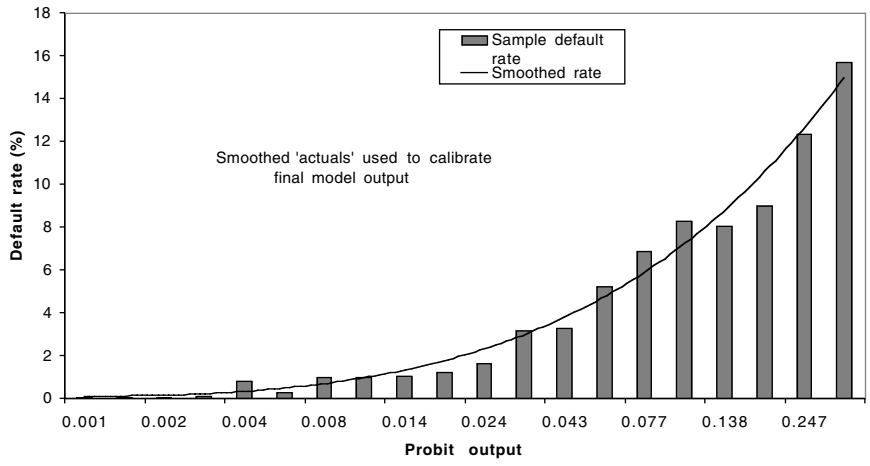
Again, computers have made this a non-issue. In any case, what used to be the signature part of a model (is it DA, Logit or Probit?) is now a subsidiary part of a broader algorithm that is not limited to a single procedure, and it does not matter that much. The author prefers probit, because many central limit theorems suggest that more phenomena have gaussian distributions (ie, normal) than logistic shapes, and the 'bad' observations generally have much greater variance than the 'goods' (a violation of one DA assumption). But even if one were to use DA, it should not be expected to affect the model significantly.[21]

Mapping to meaningful outputs

The final part of the modelling process, namely mapping, is similar to the suggestion of how one transforms the inputs. One should take the output of the probit model and find the best fitting function that maps the output into the sample default probability. This is done because invariably the output from the probit model tends to overestimate the true probability of default within sample. It is a common problem in applied probit or logit prediction, and relatively straightforward to correct. Figure 2 shows how one takes the output of the model and maps it to sample default probabilities. Note that the ordinal ranking along the x-axis implies that the output of the probit model could be in any units, it simply does not matter. One estimates the relation between model output and sample default probability using a variety of smoothing algorithms, which on a univariate problem are not meaningfully different.

Invariably one's dataset will have either too many or too few defaults for its time period, as credit cycles are quite long (5–10 years), and the probability that the sample has an average default rate equal to the population default rate is remote. To correct for this, simply scale the sample default rate up or down depending on the relationship between the sample and the population default rate. For example, if the sample default rate is 8% annually but you assume that the population (ie, long run) default rate is 2%, then adjust the initial default estimate by 1/4.

2. Map final model output nonparametrically



Neural Networks

Arguably, the best modelling decisions, like the best investment decisions, are ones that you do not make. In the author's view, a surprising number of firms sell models based on the neural network (neural net) approach, and in his candid opinion, as they exemplify many of the modelling vices mentioned above, in the context of commercial credit risk he finds them an edifying abomination.

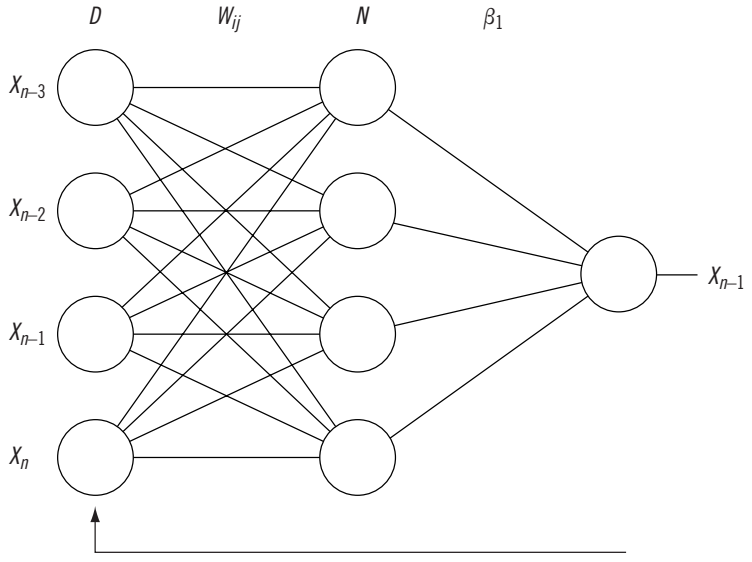
Neural network proponents are fond of expositing their algorithm with graphs, and examples that can appear as little sets of blue dots within a sea of red dots: how does one find the highly nonlinear hyperplane that separates these two populations? Another common expository device is the node graphic depicted in Figure 3.

In the author's experience, the neural net problem becomes much clearer when seeing a real example. The essence of the approach is to use a sequence of functions that, like a neuron in our brains, are either "on" or "off", 1 or 0. In practice this can be smoothed by, instead of using a discrete 0/1 function, using a sigmoidal function that ranges from 0 to 1, such as the logistic, arctangent, or hyperbolic tangent functions (sigmoidal because the function looks like an elongated S, the greek letter sigma). Then these inputs are sent to various nodes where weightings cause the inputs to cancel each other out, one to veto the others, two in combination to veto the others, etc. This is the 'middle layer' or 'hidden nodes'. Finally, all the middle nodes output are added together, in another sigmoidal or binary function.

Thus you can have a function like $g(x) = \dfrac{1}{1 + e^{-x}}$, and take an initial set of inputs and transform them as outlined above (denoted $T(x)$, which is usually a vector), and have $y = \alpha_0 g(\alpha_1 g(\beta_0 T(x)) + \alpha_2 g(\beta_1 T(x)))$. Note that the sigmoidal function g() is nested, and in this case we have two hidden nodes. Other nodes and layers generalise in a straightforward way, the number of nestings (layers) and additional functions (nodes) can extend indefinitely.

The flexibility of the structure comes from the fact that the same input vector is

3. Neural net architecture



$$X_{n+1} = \sum_{i=1}^{N} \beta_1 \tanh \sum_{j-1}^{D} W_{ij} X_{n+1-j}$$

Source: Beale and Jackson (1990).

included more than once, and therefore in one node, say $g(\beta_0 T_{(x)})$, it can diminish the output and another, it can add to it. Thus one neural node could, in effect, be turned "on" (ie, made to be at its maximum value) only when sales growth is high and profitability is low, while another node could be turned "on" only when sales growth is low and leverage is high. One could have several of these nodes, all embodying different such "if/then" statements, and this could be taken to the next level by weighting their importance based on their accuracy. Then these nodes could be used in similar "if-then" type logic (eg, if two of the three nodes are "on", the final node is "on").

That is the explanation in theory, but in practice it is difficult to assign any such interpretation because given the transformations and the multiple coefficients (there are usually at least five inputs), it is very difficult in practice to "see" which inputs are doing what; are they capturing a nonlinear interaction, or just two highly correlated inputs (at the node level) suffering from a 'wrong sign' event? Furthermore, flexibility always comes at a price, and in this case the price is degrees of freedom. Adding more parameters will explain the data better, but perhaps in the same way that adding different (but highly correlated) measures of profitability add to a 'regular' probit model's fit.

Parameter estimates are found through a 'fit and test' iteration, whereby one randomly tries some interaction weights based on their effectiveness in one sample, then tests them on a different set of data from the initial sample. Those weightings that are both most powerful and stable are deemed the best fit. The model is then finally 'validated' out-of-sample. The problem, however, is that in practice most modellers are sequentially looking for the best validation performance, and effectively make the entire 'fit, test, and validate' sample a 'fit' sample. One of the greatest advantages of explicit nonparametric approaches is that the functional form is consciously chosen and demands an explanation. For the neural net, the benefits of providing a better fit do not have offsetting costs in transparency or logic since the highly nonlinear structure is impenetrable from the start, creating a structure ripe for overfitting.

Much nonlinearity can be captured in more traditional, albeit less sexy approaches. As mentioned above, alternatives to neural nets are not simple linear combinations of raw inputs as in Altman's Z-score. The real question is whether this more complicated functional form works better, and in the author's opinion, this is just the standard Faustian statistical bargain: greater fit for today at the cost of worse fit tomorrow. Define a sample small enough and a model general enough, and everything is explained, frequently at the expense of predictive power.

In general, nonlinearity that can be captured via a function whose transformed inputs are additively separable, ie, $g(f(a_1 T_1(x_1) + a_2 T_2(x_2)))$, (where f() is a probit function, g() is the nonparametric mapping to a DP function and the T() are initial transformation functions of inputs), can be captured via the transformation/probit/nonparametric DP mapping approach outlined above. Note that if g() is convex, this implies positive interactions (two inputs aggravating each other) and if g() is concave this implies dampening interactions. If more complex interactions exist, a highly nonlinear interaction by a subset of the inputs, they would benefit from the neural net approach. Personal experience examining the data has convinced the author that such highly nonlinear interactions are neither significant nor stable. If a new sample of data shows a significant nonlinear interaction (nonlinear univariate relations would be addressed in the univariate transformation alone), it would be preferable to model it more directly in order to better understand it; its lack of precedent would make real-time monitoring interesting and important. Sometimes the argument is made that it is precisely the unknown nature of the nonlinear interaction that gives the neural net its advantage, one might be skeptical of this assertion. It could be much more probable that a spurious correlation is found through repetitive testing than a subtle but important and stable nonlinear interaction is found.

Most nonlinearity can be modelled within conventional approaches, and – to the extent that nonlinear interactions are important yet indescribable – the perceived faith in their stability going forward can be doubted. Neural nets are perhaps best left for detecting motherboard imperfections or other problems where small sample biases are not so paramount.[22]

RELATION TO TYPICAL LENDER RATINGS

Scoring is only good in relation to an alternative, accordingly it is useful to contrast scores with its most common alternative, which is not the Merton model or neural nets, but typical lender ratings. Almost all agency ratings or investment bank research are composites, qualitative opinions much like the agency ratings themselves, and in fact most ratings are explicitly mapped into the agency ratings (eg, a mapping of, say the ordinal number 4, into Ba2).

Internal ratings can be perceived to lack meaning should one consider that their intent is to capture everything. Perhaps this could be said to be a legacy of the CFA approach, which suggests that if you look at everything, you can explain anything, and most importantly, no analysis is wasted.[23] This is true in a trivial sense, but is, arguably, absurd in a practical sense. The paper, *Ten Critical Failings of EBITDA* (Stumpp, 2000), is a case in point, This was celebrated by the financial community although it essentially made the banal point that EBITDA is an insufficient metric of default risk, and that with the advantage of hindsight, ten different permutations of cashflow would have helped in various situations. True, but then how is one to know which of the ten permutations to use without the benefit of hindsight, or how best to weight these 10 permutations? Finding out that some of the innumerable distinctions in cashflow would have flagged a particular failure like Enron is of little use to those evaluating credits in real time. This is just the standard overfitting vice in a different guise.

Advocates of this approach like to highlight how models fail conspicuously for various examples, while a holistic, flexible approach could have anticipated virtually every special case. While true, it supposes that they would have acted on the right signals without the advantage of hindsight, instead of becoming mired in the numerous conflicting signals.[24] The qualitative approach emphasises explanation, narrative, and anecdotes, as opposed to the quantitative focus on prediction, models and statistics. This would all be a matter of personal preference except that statistics dominate anecdotes for the simple reason that the bottom line is a statistic – a portfolio with lower *average* credit losses, other things being equal, makes more money regardless of how many compelling anecdotes exist.

Another advantage of scores to qualitative ratings is that they are free from a curious bias where, after a certain amount of attention, the quality of the credit rating actually decreases, a bias that disproportionately affects large credits. As the US presidential election in 2000 illustrated, after a certain point a seemingly quantitative question can become purely political, and extended analysis might not shed more light, but can instead encourage vested interests to exploit the rules of the game to their advantage. It is perhaps useful to be skeptical of any credit opinion that involves extensive review by many competing parties: the final conclusion will have a nice rationalisation, but its worth is probably nil. A quantitative score is calculated just as quickly and powerfully for highly political companies as for boring ones, an unbiasedness that adds to its efficacy.

TESTING RATINGS

The best tests are those that take the least effort to understand. A model's statistical superiority should be demonstrable in a simple graph, and if a quant is unable to produce such a graph, then skepticism might be recommended. The best test involves applying common sense, and it helps if you augment your natural common sense with the following tool.

The primary testing tool for assessing statistical power – the ability to rank-order defaulters and non-defaulters – are power curves.[25] The curves graphically illustrate the ability to exclude defaulters for arbitrary cut-off points, and can be aggregated into a single statistic that allows for numerical comparisons among models. The power curve itself can be defined as follows and as illustrated in Figure 4. It maps the fraction of all companies with the worst score (x-axis) onto the fraction of defaulting companies within that group (y-axis). This creates a curve that is bowed out towards the upper left (northwest) of the chart: the greater the bow, the better.

There is in fact a mathematical relation between the power curves and frequency-of-default graphs, such that one can go from a probability graph to a power curve.

$$power(q) = \frac{\sum_{i-1}^{q} prob(i)}{\sum_{i-1}^{100} prob(i)} \tag{1}$$

$$prob(q) = 100 mean(prob)\ [power(q)\ -\ power(q-1)] \tag{2}$$

Here q is a percentile (say from the 1st percentile to the 100th percentile), and *prob(i)* is the default frequency associated with that percentile. Given the probability curve one can generate a power curve, while given a power curve and a sample mean default rate (so one can set the mean correctly), one can generate a probability curve. The curves would have the following relation: a dominant power curve is one that is always above the other power curve. The default frequency curve will cross the non-dominant power curve at a single point, and be above it at the 'bad' end of the score (where it predicts the greatest chance of default) and below at the good end. Thus, for models that strictly dominate others, one can observe such domination either through power curves (look for the more north-western line) or the default frequency curves (look for the one that generates a steeper slope).

The default probability line slopes downward in an exponential fashion, the usual case: firms at one extreme in scores fail at much higher rates than those with moderate scores. Correspondingly, the power curve rises most steeply for those scores at this same extreme. The more curvature to either curve implies greater power, and greater curvature in one implies greater curvature in the other.

The testing approach is as follows and was suggested by colleagues at the Bundesbank.[26] The main advantage of this approach is that it ensures each defaulted firm is used only once in testing (as opposed to, say, using monthly scores with future 3 year default as the dummy flag). We take each defaulting firm, and find the score 12 months prior to the default date. If a score does not exist on this date, we move backward in time until we reach 24 months prior to the default date. If no score is present we exclude the defaulting observation from this particular test. We find the score's relative rank among all scored firms (defaulters and non-defaulters) at the time of the score. This is done by using the calendar year from which the score was taken. While this explicitly corrects for the business cycle, in that aggregate fluctuations in scores are meaningful, for models that use only financial statements this consideration is of little practical importance. Each defaulter is then mapped into a percentile, and this collection of percentiles is the basis by which the power curve is created. Specifically, given a collection of percentiles of defaulting firms $(\phi_j)_{j=1}^{J}$, where J is the total number of defaulting firms, the power for each bin is simply

$$power(b) = \frac{1}{J}\ \sum_{j-1}^{J} \left\{ 1/\phi_j < \frac{b}{B} \right\} \tag{3}$$

lower than b/B, K months prior to the default date. As mentioned above, we take the score in the month closest to K up to a limit, so in effect the percentile is for the set of scores taken between K to K + ΔT back from the default date. For example, for a one-year default probability test we would take a default on May 98, and move back to May 97 to find the percentile of the score in 1997 using that month. As is most probable, the statement date is not exactly at May 97, and so we must go back in

time, to April 97, then March 97, etc, until we find the date at which we have financial statements. For the short term test, we go backward in time starting 300 days prior to default and moving backward in time to two years prior to default. For the long term test we go backward in time starting five years prior to the default date and move forward in time until we get to 36 months prior to the default date.

By going backwards at least 180 days, and ideally 365 days from the default date, we avoid the misleading results that come from model performance over irrelevant time periods. Predicting default of very short horizons, such as less than 90 days, is useless as very few statements are completed within this time, and in general the equity value is near zero and few meaningful decisions are feasible with the debt at this time. Many commercial lenders take six months to be confident that most of their middle market exposures have delivered their latest annual statements.

It is useful to know that the number of degrees of freedom is not reflected in such misleading data like the number of firm-years, which in public databases is often well above 100,000 firms. Only the number of defaults matters to the true degrees of freedom, and thus regardless of the number of firm-years, 114 defaults allow only 113 degrees of freedom – given 113 different ratios one could perfectly explain these defaults. The author's personal view is that it might not be worth building or testing a model without 500 defaults.

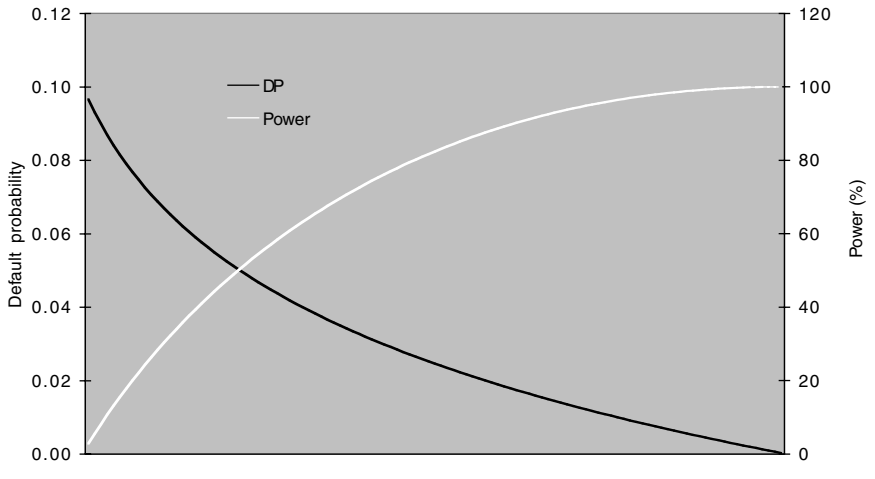
It is important to note that a testing metric, such as the area under a power curve, is always a function of a model and the particular data set. Different data sets offer wildly different power possibilities, and a common statistical 'trick' is to point out that one's model has more power than someone else's – on a different set of data. Such comparisons between models are meaningless because different data sets can produce wildly different power statistics using the same model. This is why if data are proprietary, the credibility of the presenter is just as important as the test results. If the data are nonproprietary, in which case everyone knows the 'right' answers to the test, overfitting is still a problem, though not biased in anyone's favour.

Even the most rigorous statisticians do not have their intuition reversed by a single test, and so one should not be too focused on any single statistical test, as this indirectly encourages overfitting. If one knows that a power curve on a particular data set will determine the winner of an upcoming contract or internal political battle, the clear incentive is to find the highest fit and then rationalise it.

BUILD VS BUY

Often institutions feel compelled to build models lest they admit they do not have a competitive advantage in what seems the *sine qua non* of lending.[27] Perhaps most

4. Power curve vs default probability



institutions would be better served by buying a model, which implies it has experienced much greater peer review than any internal model, and then adjust it, by say combining it with some expert rules.

There are several factors that make buying a model more attractive than building. First, most institutions do not have nearly enough data; observations are not independent, and there is usually a factor of 10 or 100 between an institution and a vendor in the number of defaults. Second, while good models are conceptually simple, they are not easy to make: sort of like cooking a good steak. Often a senior executive can feel that their intuition and experience, combined with that of a former physicist, should be able to create a model that is able to demonstrate to everyone how cutting edge their institution is. Most often, however, the intuition and experience of the lender does not help in model creation, because good modelling is often counter-intuitive: it ignores much information, and does not always target the highest correlation. Also, the arduous process of communicating the meaning and validation of a score takes time, and the more that you can rely on someone else's credibility, the better.

If one has sufficient data and wishes to make a scoring model, the following objective helps: aim to make a model that has equal power but is simpler or more transparent than its alternatives. That is, instead of focusing on increasing power, which often leads to overfitting, focus on simplifying a well-known model's structure or data inputs. This is a much more promising way to add value, playing on the fact that most models are overfit.

Conclusion

When a service is both cheaper and of better quality, it makes a compelling alternative. In the 1970s it might have been considered sacrilegious to suggest that such an important, nuanced task as economic forecasting could be trusted to anyone but a qualified analyst. Banks had huge economics departments, and lengthy reports with data and commentary were generated on regional, industry, financial and macro outlooks using proprietary, in-house expertise (Sherdan, 1999). In retrospect this appears pure folly, and while much analysis has been replaced by such sophisticatedly simple alternatives as vector-auto regressions (Sims, 1982), most of it was simply eliminated. A more direct precedent comes from consumer credit analysis which, at one time, involved the kind of exegesis currently performed on commercial credits: scoring for these smaller loans now totally dominates the underwriting process for consumer loans. The same future awaits commercial credit departments: a current welter of credit information will be focused by scoring, because it is more transparent, powerful, faster, consistent and less expensive.

1 *In general, a bond default is defined by any missed or delayed disbursement of interest or principal, bankruptcy, receivership, or distressed exchange where the issuer offered bondholders a new security or package of securities that amount to a diminished financial obligation or the exchange had the apparent purpose of helping the borrower avoid default.*

2 *See Kuritzkes and Wilkinson (2002).*

3 *EBITDA = earnings before interest, taxes, depreciation and amortisation; EBIT = earnings before interest and taxes.*

4 *This chapter will not address the following interesting and important credit topics: time series properties of defaults and implied default probabilities, ratings transition matrices, reduced form models, how credit spreads relate to scores, recovery rates, regulatory treatment of internal models, retail credit risk (credit cards), or portfolio models like CreditMetrics. These other issues tend to take model scores as inputs, or in the case of retail, involve a very different set of data.*

5 *The Moody's RiskCalc series includes an initial lengthy discussion of modelling*

issues and then applies selective commentary relevant to various countries. (See http://www.riskcalc.moodysrms.com/us/research/crm.asp)

6 *Overfitting is also called data snooping, data-mining (which ironically is not a pejorative in some circles), or a degrees of freedom problem; you want your data to have many degrees of freedom (independent observations) and your model to have few degrees of freedom (parameters).*

7 *Actually, it is the number of independent observations, or the number of observations minus the number of necessary relationships. See Good (1973). The degrees of freedom of a nonlinear model are often difficult to determine because the notion of a parameter may be blurred. For example, the kernal regression has only one parameter, the bandwidth, but each data point is used as the centre for local averaging.*

8 *Ie, the modeller should be more interested in helping the parochial interests of his business vs academic fame.*

9 *See Credit Suisse First Boston's CUSP (Credit Underlying Securities Pricing), RiskMetric's CreditGrades, CreditSight's BondScore, Fitch's Risk Rater, or Kamakura's KRM-cr models.*

10 *Some areas, however, have too few default events to score, eg, sovereign risk.*

11 *London inter bank lending rate (LIBOR).*

12 *It is interesting that CreditGrades was estimated and tested, using spread data, on a mix of investment-grade and non-investment-grade companies. It will be interesting to see how this turns out.*

13 *As an American author, leverage is used hereafter out of habit while conceding the equal vividness of the British-favoured term 'gearing.'*

14 *Famous German–American rocket technology pioneer from the 1930s to the 1970s.*

15 *This all holds in generalised multivariate models as well the OLS (Ordinary Least Squares) example presented here.*

16 *See Khaneman, Slovic and Tversky (1982).*

17 *This generalises to the multivariate discrete models as well as the bivariate case shown here.*

18 *In tests it did about as well as NI/A, and strictly worse than the naïve approach NI/A-L/A. See Falkenstein, Boral and Carty (2000).*

19 *In later papers Altman has tried mapping to DPs indirectly through agency ratings, by first looking at average scores in each rating grade. The problem is that this supposes that the power of ratings and Z-scores are the same. This has never been established. See RiskCalc (1), p. 45.*

20 *Specifically, homoskedastic errors between non-defaulters and defaulters, and inputs having Gaussian distributions.*

21 *Given the transformations of the inputs to variables that are much more normally distributed, this greatly lessens any differences created by any of these different estimation techniques,*

22 *It is not an accident that neural nets are much more popular among physical scientists than economists: degrees of freedom are much less a problem in the hard sciences.*

23 *CFA stands for Certified Financial Analyst, a US-based professional accounting group.*

24 *This dilemma is highlighted today in the US as journalists point out that there were several warnings that terrorists were going to 'do something' prior to September 11, 2001 and this information was effectively ignored, while at the time of writing there are constant issuings of warnings that something may be imminent has generated complacency. There will always be some signal that, with hindsight, should have been heeded; unfortunately it is very difficult to know which ones are right for the future without being catatonic.*

25 *Also known as ROC (receiver operating characteristic) curves, Lorentz curves, gini curves, cap plots, among other terms.*
26 *Specifically, Stefan Hohl, Stephan Blochwitz, and Thilo Liebig.*
27 *Essential requirement.*

BIBLIOGRAPHY

Altman, E. I., 1968, "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy," *Journal of Finance*, 23., pp. 589–609.

Beale, R., and T. Jackson, 1990, *Neural Computing, An Introduction*, (Philadelphia: Institute of Physics Publishing).

Boral A. K., and E. Falkenstein, 2001, "Revisiting Mr. Merton", *Risk Professional*, 3.

Cantor R. and E. Falkenstein, 2001, "Testing for Rating Consistency in Annual Default Rates", *Journal of Fixed Income*, September, pp. 36–51.

Card, A., 1998, "The Casual Effect of Education on Earnings," in: *Handbook on Labor Economics*, Volume 3,. O. Ashenfelter and D. Card (eds.), (Amsterdam: North-Holland).

Casey, C. J., and N. J. Bartczak, 1984, "Cash Flow – It's Not the Bottom Line", *Harvard Business Review*, July–August, pp. 61–6.

Chen, K., and T. Shimerda, 1981, "An Empirical Analysis of Useful Financial Ratios", *Financial Management*, Spring, pp. 51–60.

Falkenstein, E., A. K. Boral and L. V. Carty, 2000, "RiskCalc for Private Companies: Moody's Default Model", *Moody's Investors Service Special Comment*, May.

Fons, J. S., and A. E. Kimball, 1992, "Corporate Bond Defaults and Default Rates", *Journal of Fixed Income*.

Golden, R. M., 1996, *Mathematical Models for Neural Network Analysis and Design* (MIT).

Good, I. J., 1973, "What are Degrees of Freedom?", *American Statisticians*, 27, pp. 227–8.

Jarrow, R. A., and S. M. Turnbull, 1995, "Pricing Derivatives on Financial Securities subject to Credit Risk", *Journal of Finance*, 1(1), March, pp. 53–85.

Kealhofer, S., S. Kwok and W. Weng, 1998, "Uses and Abuses of Bond Default Rates", *CreditMetrics Monitor*, First Quarter, pp. 37–55.

Khaneman, D., P. Slovic and A. Tversky, 1982, *Judgement Uncertainty: Heurstics and Biases*, (Cambridge University Press).

Kuritzkes, A., and B. Wilkinson, 2002, "Can Loan Pooling Rescue CLOS?", *Risk*, February, pp. 36–7.

Merton, R. C., 1974, "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates", *Journal of Finance*, 29, pp. 58–66.

Ohlson, J. S., 1980, "Financial Ratios and the Probabilistic Prediction of Bankruptcy", *Journal of Accounting Research*, Spring, pp. 109–31.

Sherdan, W., 1999, *The Fortune Sellers: The Big Business of Buying and Selling Predictions* (New York: John Wiley and Sons).

Shumway, T. G., 2001, "Forecasting Bankruptcy More Accurately: A Simple Hazard Model", *Journal of Business*, January, pp. 101–24.

Sims, C., 1982, Policy Analysis with Econometric Models, *Brookings Papers on Econometric Activity*, pp. 107–52.

Sobehart, J., and R. Stein, 2000, "Moody's Public Firm Model", Moodys Investors Service.

Stancill, J. M., 1987, "When is there Cash in Cash Flow?", *Harvard Business Review*, March.

Stumpp, P., 2000, "Putting EBITDA in Perspective: Ten Critical Failings of EBITDA as the Principal Determinant of Cash Flow", Moody's Investors Service.

Wilcox, J. W., 1977, "A Gambler's Ruin Prediction of Business Failure Using Accounting Data", *Sloan Management Review*, September, 12, pp. 33–46.