Credit scoring: a future beyond empirical models[1]

David J. Hand
Professor of Statistics, Department of Mathematics, Imperial College, and Institute for Mathematical Sciences, Imperial College

Adam R. Brentnall
Research Associate, Institute for Mathematical Sciences, Imperial College

Martin J. Crowder
Professor of Stochastic Modelling, Department of Mathematics, Imperial College

Abstract

The paper argues that 'iconic' models, models based on some sort of theoretical foundations, are likely to be more robust in uncertain retail credit scoring economic situations than purely 'empirical' models, models based on pure data analysis. The two types of models are set in a historical context, both being contrasted with judgmental models, and a variety of models constructed for retail credit scoring are described, ranging from the predominantly empirical classic segmented scorecard to the mainly iconic MIMIC (multiple-indicator-multiple-cause) models. Relationships to behavioral finance, the Lucas critique, and the structural/reduced form distinction are noted.

1 The work of Adam Brentnall on this project was supported by EPSRC grant number EP/D505380/1 and the work of David Hand was partially supported by a Royal Society Wolfson Research Merit Award.

Broadly speaking, there are two different kinds of statistical model [Box and Hunter (1965), Lehmann (1990), Cox (1990), Hand (1985, 1994, 1996, 2008)]. The first type goes under various names, including mechanistic, phenomenological, substantive, and iconic, and are based on formulations of a simplified representation of some kind of theoretical or observational construct. Thus, a theory of how falling bodies move in a gravitational field will be represented by a system of equations — an 'iconic' representation of the 'mechanism' or 'phenomenon' of the physical system. Statistical methods are then used to estimate the parameters of the model. The second type, likewise referred to by various names, including empirical, data-driven, and descriptive, merely seek to summarize the relationships between variables in a convenient way, and the model form has no theoretical base at all. Thus, for example, a linear regression model might be constructed to allow us to predict propensity to purchase a product from the observed values of a number of covariates, without any theoretical link between the covariates and the response. These variables might be chosen from a very large set of potential predictors, by a statistical variable selection procedure aimed at maximizing predictability, regardless of semantic content. Clearly there is a grey area between these two classes of models. A complex model may have both iconic and empirical aspects. Furthermore, a model may start out as merely empirical, and then become iconic as knowledge advances and theory is developed.

Observation also shows that different disciplines place different degrees of emphasis on the different model types. In the management sciences, for example, empirical models (with a linear regression structure, for example) are very popular, whereas in the natural sciences, such as physics, iconic models are perhaps preponderant (often based on mathematics such as differential equations relating the system variables). In this paper we argue that the retail credit industry has tended to favor empirical models, or, at least, models with a large empirical aspect, but that one might expect iconic models to show superior performance in certain situations.

This paper is restricted to 'input/output' models, having the form $y = f(x1,..., xp)$. The $x = (x1,..., xp)$ go under various names including predictor variables, covariates, input variables, and even independent variables. Variable y is the outcome, response, dependent variable, etc. The model form, f, is constructed on the basis of past data for which (x, y) pairs have been observed. In the retail credit industry, each of these pairs may refer to an individual, a credit line, a transaction, etc., according to exactly what is being modeled. For example, if creditworthiness of individuals is being measured, x might include characteristics such as time at present address, type of accommodation, etc., and y might be whether or not they previously defaulted on some financial product. In contrast, if an individual transaction is being scored (as, for example, in fraud detection), x might include size of transaction, time since previous transaction, time of day of transaction, etc., and y might be whether or not the transaction turned out to be fraudulent. In constructing retail credit models, the aim is to use the observations of multiple past (x, y) pairs to model the relationship f so that one can predict the likely outcome value y for a new observed vector x.

Iconic models are not to be confused with judgmental approaches. The next section gives some historical background to this, describing judgmental approaches and how these differ from and preceded the widespread adoption of empirical models in the retail financial services sector and other areas. In a different context, the two approaches are termed clinical (i.e., judgmental) and statistical or actuarial (i.e., empirical), and, as Dawes et al. (1989) put it, "In the clinical method the decision-maker combines or processes information in his or her head. In the actuarial or statistical method the human judge is eliminated and conclusions rest solely on empirically established relations between data and the condition or event of interest."

The aim of this paper is to describe a move to yet a third class of approaches, iconic model-based approaches, in which the models are based on relations between data and the condition or event of interest derived from theoretical considerations. Grove and Lloyd (2006) define judgmental approaches, "data can then be combined by a professional using trained judgment," and empirical approaches," data can then be combined…mechanically (e.g., by a formula, actuarial table, or computer program)" and then say, "There is no true hybrid of these data combination methods." Whether or not this is true, the models we term 'iconic' are not so much a hybrid as a third alternative. They do not rely on professional judgment, but do rely on well-formulated and predictive theories describing the underlying processes. They do not rely on blind analysis of data, but on analysis of data fitted into a theory.

Judgmental versus empirical approaches

The past half century has witnessed a radical change in the decision-making strategies employed in the retail financial services sector. In particular, the human judgmental approaches, which were the familiar way of making credit granting decisions throughout the bulk of history, have been replaced by objective procedures based on data. Instead of judging an applicant for a loan or financial product on the basis of personal knowledge, their standing in the community, or such subjective characteristics as their appearance, decisions are now based on elaborate statistical models, often referred to as scorecards. This permits the automation of decision-making processes, so that far more credit granting decisions can be made than would be possible if it was necessary to rely on human judgment for each one. Moreover, the advent of revolving credit operations such as credit cards have necessitated the monitoring of ongoing behavior, so as to warn of impending difficulties, increase the credit limit, and in general to manage customers in an optimal way. Once again, the sheer numbers involved would render this impossible if it relied on human judgment.

In addition to these practical aspects, statistical scorecards have a range of properties which make them particularly attractive in this role. They do not tire if applied repeatedly minute after minute, hour after hour, and day after day. A specific statistical model will always give an identical result on identical data. They are not subject to irrational prejudices. They can also be incrementally improved, by experimenting with different structures, such as additional predictor variables to find those which enhance predictive performance. Furthermore, one can objectively assess them against legal requirements, such as the need to be able to provide some sort of explanation for adverse decisions. There are also psychological phenomena which affect human judgment, but not, of course, objective statistical models, such as a poor ability to determine prior probabilities (i.e., in the relative numbers of good and bad customers), a tendency to take account of variables which may in fact not be predictive (one might call this 'superstition'), and a tendency to be overconfident on one's own judgments.

Of course, the balance of relative merits is not entirely one-sided. Scorecards are not good at detecting the anomalous customer, such as the medical doctor (generally a good risk factor) who has previously served prison time for embezzlement (generally not a good risk factor), or the new CEO of a blue-chip company (generally a good risk), who has just taken up his post (short time in present employment being a poor risk factor), having moved from overseas (and so having no credit record).

In the early days of objective credit scoring, there was considerable doubt that statistical models could perform as effectively as human perceptions and judgment. This led to experimental assessments comparing the relative performance of the two approaches, where it soon became clear that the formal models had the edge. For example, Myers and Forgy (1963) review studies dating from 1941 to 1960 and comment that "While results from these studies have differed, all have shown that a properly constructed numerical rating system can offer at least some degree of improvement over the purely subjective or judgmental approach to evaluating credit. In some cases, the amount of this improvement has been substantial."

This work was paralleled by analogous research in other areas. For example, the eminent psychologist Paul Meehl published a book [Meehl (1954)] in which he contrasted 'clinical prediction' (i.e., human judgment) with 'statistical prediction' (i.e., the data-driven models). The book reviewed 20 comparisons of the two approaches, concluding that 19 of the 20 demonstrated superior or equal performance by the objective approaches. Further subsequent studies supported this. Grove et al. (2000) carried out a meta-analysis of 136 studies in psychology and medicine, each of which compared the performance of "at least one human judge to at least one mechanical-prediction scheme." They concluded that: "On average, mechanical-prediction techniques were about 10% more accurate than clinical predictions … These data indicate that mechanical predictions of human behaviors are equal or superior to clinical prediction methods for a wide range of circumstances."

Paul Meehl (1986) produced a nice simile in favor of the empirical approach, writing, "When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "well, it looks to me as if it's about $17.00 worth; what do you think?" The clerk adds it up."

Iconic versus empirical approaches

One can think of models as being based on various components, analogous to a set of basis functions, which are combined so that the whole yields the function f. Familiar examples of such components in credit scoring contexts, which will be found in some models, are: the particular variables included in the model, a linear combination of the predictors, a transformation of the linear combination to yield the outcome measure, a segmentation of the customer base, binning of continuous variables, derogatory trees, weights of evidence, use of macroeconomic variables, use of hidden states, inferred reject classes, etc.

There are two distinct types of justification for including each component. On the one hand one might believe that a component represents some aspect of the underlying behavior of the system being modeled. These are iconic components. For example, one might have theoretical grounds for believing that different types of people will behave in different ways, and hence segment on the basis of that understanding. While we do not wish to enter into a philosophical discussion on the existence of underlying true models here ["all models are wrong, some models are useful" Box (1979)], it will be convenient for the exposition to refer to the underlying 'truth' in the way people actually behave, so we will omit the inverted quotation marks from what follows. On the other hand, one might simply observe a relationship in data being used as the basis for constructing a model and choose to include a component which manifests that relationship. These are empirical components. For example, one might have chosen which ten predictor variables to include in a model after an extensive step-wise search through all the variables in the database. There may be no behavioral theory suggesting that such a model form represents the truth.

The grey area between iconic and empirical models is perhaps illustrated by a weak behavioral theory suggesting that a linear combination is a first approximation to the truth (the first term of a Taylor expansion, for example). To construct an overall model, of whichever kind, an implicit search is carried out, over a universe of potential components, to decide which ones to include. The universe of potential iconic components is derived from theory, and the universe of potential empirical components is derived from observation of relationships in the past data (or, more generally, in

other past data – and hence also from experience of similar models in the past). Of course, models may also use a mix of component types, and we will see examples of this in the retail credit context in the next section.

If the underlying theory is sound, in the sense that it does accurately represent aspects of the underlying truth, then the iconic model will be based on components which likewise permit an accurate representation of the truth, so that a good model will result. Conversely, of course, if the underlying theory is poor, then the iconic model which results will be poor. Poor iconic components mean that one cannot accurately reflect the true relationships – in a way analogous to that in which inadequate models lead to bias in lower level statistical model building. On the other hand, the fact that the search space is limited to the set of components with a theoretical justification means that it is likely to be smaller than the empirical component search space. That means it is likely to have a lower variance.

Empirical models are likely to have a larger component search space because, by definition, there is no theory constraining the set of components which might be considered. We could use any transformation of characteristics and combinations of characteristics which came to mind, as potential components (as, for example, in genetic algorithms), in the hope that they would yield a good model for our past data, which we can then use to extrapolate to make future decisions. This large search space means that the variance of the final model is likely to be large. On the other hand, this very flexibility in fitting past data means that its bias is likely to be small, again in a way analogous to lower level statistical modeling. The larger search space may mean that predictive components not considered in the iconic search are included. For example, one may not contemplate including the color of the applicant's car in the iconic model, on the grounds that one can see no grounds for supposing a relationship between this and creditworthiness. However, it might be the case that the empirical search shows such a relationship and hence included it in an empirical model.

This has mixed benefits. It may lead to superior prediction, at least for data drawn from the same distribution as the original design data. But on the other hand, it might raise regulatory eyebrows – and indeed concern has been expressed in the past when variables with no obvious link to the outcome have been included in such models. There are also potential issues of such variables acting as proxies for excluded variables (such as gender).

The two classes of models thus have complementary characteristics. Iconic models will be biased to the extent that the theory is unsound, but are likely to show less variability. Empirical models will show less bias, but greater variability. Empirical models can be used even if one does not have any ideas about underlying behavior, but if one does (and the ideas are right) then iconic models are likely to be more accurate. Of course, these are generalizations, and things are complicated by considerations such as the sample size and dimensionality. In particular, in so-called 'large p small n problems', that is, high dimensional problems with relatively small sample size, dramatic regularization and the use of simple and less flexible model forms typically proves superior, even though this is purely empirical and non-iconic, so that advantage is gained by accepting some potential bias in exchange for significant reduction in variance.

The papers cited at the start of this article mostly discuss the two kinds of models in the statistical context. Hand (2008), however, has extended the discussion to the kinds of models typically used in data mining, arguing that these are usually empirical. It is almost intrinsic to the nature of data mining that it involves a search for serendipitous discoveries of value or interest or fits models unconstrained by theoretical notions of what model forms might be appropriate.

Models and changing circumstances

In a stationary environment, given a sufficiently large dataset, one should expect an empirical model to yield a good fit to the distributions which generated the data, and hence relatively accurate predictions. Iconic models will be more or less accurate depending on the quality of the underlying theory. In non-stationary environments, however, one might expect the performance of both model types to degrade. For empirical models, the choice of model components will have been made on the basis of the available data, and hence, by implication, the model will be matched to the distributions from which the data arose. Adapting them to match the changed or changing distributions requires repeating the elaborate search over the component space. In contrast, while iconic models will also degrade if the distributions change, their basic structure will remain sound (assuming they reflect the truth well). All that needs to be done to update an iconic model is to re-estimate its parameters. For example, if we base predictions of the strike point of a shell fired from a cannon on an iconic model using Newton's Laws, the same model also applies if we move the cannon to Mars, albeit with different parameters. For an empirical model, however, things are rather different. Such a model will have been constructed on the basis that it led to good predictions on Earth, and may require complete rebuilding to capture the different conditions of Mars. It may have no component analogous to the uniform gravitational field approximation to the least squares law leading to the parabolic trajectory of the shell. Indeed, for a more extreme astronomic example, contrast models for the apparent motion of stars and planets based on Ptolemaic epicycles with models based on an iconic Copernican sun-centered approach, and then imagine using models of each type to predict corresponding motions in another solar system. The Copernican model would lead to sound predictions using a tiny fraction of the data needed for the Ptolemaic model.

In fact, given the sound theory of a good iconic model, one can often go even further than this. In the context of a system evolving over time, one might be able to forecast likely future values for parameters. This means that, with iconic models, one can make forecasts of outcomes, in a way which would be impossible with empirical models.

A simple illustration of the relative merits of extrapolating using models of each type is given by contrasting two approaches to predicting binary outcomes: (i) an empirical linear regression model, and (ii) a logistic regression iconic model, based on the notion that one is in fact modeling probabilities which must lie between 0 and 1. If the values of the predictor variable used to construct the model lie in a narrow region, then both will give good and comparable predictions for further data from that region. However, if things start to drift, and future predictor variable values lie far from the original region, then the linear model is likely to behave poorly (indeed, it may even give negative predictions for the expected probability).

The phenomenon of a model performing perfectly well in circumstances identical to or similar to those under which it was constructed, but decaying dramatically if and when those circumstances change, has been termed the cliff-edge effect [Hand (2008)]. As illustrated in the above example, we might expect empirical models to be more susceptible to this effect than (well-chosen) iconic models. All this is of particular relevance in the retail credit sector, since a major characteristic of the sector is that it is generally not stationary [see, for example, Kelly et al. (1999)]. The distributions involved change in response to changing competition (i.e., new players entering the market), changing technology (i.e., Internet banking), changing products (i.e., customer specific risk models), and, particularly important at present, changing economic circumstances. These ongoing changes are the reason why credit scorecards are typically rebuilt every few years. Since the models in the sector are primarily empirical, such rebuilding typically involves a major search over the component space. An iconic model, however (provided it is a 'good' one, with components properly representing the truth) is likely to be less susceptible to such changes. Updating the parameter estimates, and, indeed, building forecasts of likely future values of the parameters, would continue to yield effective credit models, without global rebuilding. This clearly has implications for cost and predictive accuracy.

Credit scoring models

So far in this paper we have discussed general issues of credit scoring models, showing how the discipline advanced from judgmental approaches through empirical models, and suggesting that further potential advantages are to be gained from iconic models. In this section we examine some credit scoring models in detail, to see how they fit into the empirical/iconic classification.

The traditional scorecard structure

The basic form of a segmented scorecard or logistic regression tree is very widely used in the retail credit industry [Hoadley (2001)]. Such a model can be a mix of empirical and (weak) iconic components, but is very largely empirical. To illustrate the mix, let us first look at how the segmentation is achieved, because this can be based on either of the two philosophies.

Firstly, the overall population may be divided into segments on the basis of ideas or beliefs about what are likely to be the important characteristics in distinguishing behavior types. Thus, for example, one might have a high level division into those customers with clean credit histories and those who have shown prior mild or severe delinquency, believing that people distinguished in this way are likely to behave differently. Such a segmentation is illustrated by Fair Isaac's NextGen model [Fair Isaac (2004)]. The division is made because of the prior expectation that the three different groups will behave differently. Such a segmentation is clearly iconic — although only weakly so because it is based more on prior experience than on any firm theory.

Secondly, and alternatively, the overall population may be divided into segments based on an elaborate model search over possible partitionings (for example, using the CART software of Breiman et al. (1984)). A partitioning will be adopted if it is seen that it leads to improved predictive power, regardless of whether there is or is not any underlying rationale for such a partitioning. It is entirely possible that it could include segmenting variables for which one finds it difficult to construct a rationale underlying different behaviors in the different segments. A segmentation achieved in this way is clearly empirical.

It is important to note that often scorecards include splits based on both reasons, so that a mixed empirical/iconic model results.

Within segments, the classical segmented scorecard typically uses a logistic regression (or something equivalent to this). Most often, also, any continuous raw variables are binned — they are split into intervals, with these being treated as indicator variables, scoring 1 if an individual takes a value in the relevant bin and 0 otherwise. In fact, the binary score for an indicator variable and the regression weight associated with that indicator variable can be merged and regarded as a score for that indicator. The final score for an individual falling into a particular segment is then a sum of points values, one point value for each of the indicator variables on which they take value 1. The binning has the familiar disadvantages of enforcing a rigid partition between neighboring values on opposite sides of a bin division, but the advantage of being a nonlinear transformation of the raw variable. The resulting model is a generalized additive model [Hastie and Tibshirani (1990).]

Models of this form can be written on a sheet of paper, with rows corresponding to each raw variable, and with its bins and their point values across the page. To score a new customer one simply sums, down the page, the values of the bins into which they fall. This format is the historical reason underlying the term 'scorecard.'

The various steps in constructing models within each segment, the selection of which variables to include, and the binning are typically achieved by elaborate searches (over a population of potential variables; over possible split points for binning), so that these models are predominantly empirical.

Superscorecards and beyond

Segmentation is, of course, a way to allow for an interaction between the segmenting variable and the variables used to construct the score within each segment. A more traditional, and markedly empirical statistical approach would be to include these interactions explicitly, as additional product terms in the model. However, that could easily lead to vast models, and a search over an even vaster interaction space. Already, for example, Experian-Scorex gives a set of 447 variables from which to choose when constructing a scorecard [Experian-Scorex (2004)], so that including interactions by multiplication could very rapidly lead to a vast potential number of variables to include. Even with the large datasets typically available in credit scoring applications, overfitting could be a problem. Sometimes so-called derogatory trees are used to define new, interaction, variables. These are small tree structures which partition just a handful of variables.

One of the advantages of the simple weighted sum (or scorecard) format within segments is interpretability. In particular, it is easy to identify those variables which have most impact on a score and hence to satisfy the legal requirement to be able to explain, at least at some level, any decision not to give a financial product to a particular applicant. This interpretability requirement mitigates against the use of some of the highly flexible modern tools such as neural networks and multivariate adaptive regression splines.

Suppose that a scorecard is based on p raw variables, and that that an individual takes a value in bin r_i of the i^{th} variable, and that the weight for this bin is c_{ir_i}. Then the traditional scorecard computes this individual's score as $S = \Sigma_{i=1\rightarrow p} c_{ir_i}$. Superscorecards [Hand and Kelly (2002)] extend this form by combining two scorecards of the traditional form in a multiplicative way: $S = S_1 S_2 = \Sigma_{i=1\rightarrow p} a_{ir_i} \Sigma_{i=1\rightarrow p} b_{ir_i}$.

Now, instead of each bin having a single parameter, c_{ir_i}, each bin has two parameters, a_{ir_i} and b_{ir_i}. In practical terms (and in terms of explaining such a model to someone familiar with traditional scorecards), the resulting score is simply the product of two standard models.

The model form is chosen on the grounds that it is a straightforward extension of standard forms, but it is still clearly empirical: statistical notions rather than semantic notions were the driving force. Since it includes the standard form as a special case (simply set the $S_2 = \Sigma_{i=1\rightarrow p} b_{ir_i}$ component identically equal to 1 to regain the standard form) it is more flexible and will allow more accurate models.

Graphical models

Although the focus of this paper is on models which have a well-defined outcome variable, y, it is often the case that multiple outcomes are of interest. Indeed, the basic Basel II model makes use of probability of default, loss given default, and exposure at default. More generally, one might be interested in such things as probability of defaulting on a loan, probability of becoming delinquent on a credit card, or a debit card, of falling behind in mortgage or car finance repayments, etc. While, clearly, models can be built for each outcome separately, it is equally clear that there are relationships between these variables, so that it might be more effective to construct a single global model which fits them all together, along with all the potential predictor variables. There are various approaches to such global models, but one which has attracted some attention in the retail financial services sector is the graphical model [see, for example, Hand et al. (1997), Sewart and Whittaker (1998), Stanghellini et al. (1999)].

Models of this kind decompose the joint probability distribution of all the variables into a product of factors, each of which refers to just a few of the variables. For example, with six variables, $z_1,\ldots,z_6$, (we do not distinguish between response and predictors here), the joint distribution $g(z_1,\ldots,z_6)$ might be factorized as $g(z_1,\ldots,z_6) = g_6(z_6)g(z_1,z_2|z6)g(z3|z_1)g(z_4,z_5|z_3)$. Here none of the factors involve more than three random variables, so that a simplified description of the joint distribution results.

Constructing models of this kind requires choosing its topology (which variables are involved in each factor) and estimating the numerical values involved in each factor (the relationships between the variables). There has been a considerable amount of work on both of these problems in the last couple of decades. Determining the topology can clearly benefit from constraints imposed by iconic considerations, but empirical search procedures have also been developed. Determining the numerical values in the relationships is essentially an empirical problem.

Survival models

The basic form of credit risk model in the retail sector seeks simply to distinguish those who are likely to default (according to some well-formulated definition) from those who are unlikely to do so, within a specified time. In fact, however, defaults may occur at any time during the term of a loan. Recognizing this structure to the

system, various authors have developed approaches to modeling the time at which a default occurs, rather than merely whether one occurs [see, for example, Banasik et al. (1999), Hand and Kelly (2001)]. If the aim is still simply to determine if a default occurs (i.e., prior to the loan term expiring) then this approach is fitting a more elaborate model based on the known properties of the system. Of course, such models may make assumptions (i.e., about the form of the survivor function), so that they perform poorly if this iconic aspect is poorly chosen. Nonetheless, they do represent a step towards iconic models.

Structured class definitions

A common definition of default in the retail banking sector is that someone falls three months in arrears within a twenty-four month period. However, there is nothing rigid about this definition. Either of the two numbers involved could be varied. Indeed, much more elaborate definitions for the two (or more) classes which are the object of the predictive model could, and have been used, based on a variety of characteristics of the customers and their behavior. For example, a 'good' customer could include such things as how long they have been on the books, how actively they use their account, how many credit lines they have, the size of their purchases, and whether they have ever been one month in arrears.

When such more elaborate definitions are involved, one could either try to predict the classes directly from the predictor variables (which is the common approach), or one could try to predict the future values of the variables used in the definitions of the classes, combining these predictions to yield the predicted class membership [Hand et al. (2001), Li and Hand (2002)]. Clearly, since one is using the known system property given by the class definition, this is an iconic approach (though, of course, the individual component models may be empirical).

Behavioral trait models

More elaborate models seek to decompose the fact of defaulting into component causes. For example, one might conjecture that default depends on (i) an individual's natural propensity to break an agreement (a moral aspect, if you like), (ii) customer-specific random components (i.e., losing one's job or getting divorced), (iii) general economic random components (i.e., exogenous influences, such as the inflation rate), (iv) societal mores (i.e., a difference between different age groups in their ordering of meeting competing debt obligations), and (v) other influences. Now we are moving more clearly into the iconic realm.

One simple example of such a model is an attempt to overcome the problem of population drift, and the fact that scorecards are always constructed on past data, and hence are out of date even before they are first used. This approach is based on two components. The first is a postulated fixed 'behavior type' or trait for each individual. Statistically, one can think of this as a latent random effect for each individual, indicating their propensity to default, with the actual event mediated by other influences. The second component is a link between behavior type and outcome (default), with this link assumed constant given constant external influences. Because of this constancy assumption, this link can be estimated from past data. So, one can go back in time to model the link between type and outcome, and use very recent data on each new individual to determine type. This means that the scorecard will be built on very recent data. Of course, as always, this essentially iconic model depends on the accuracy of the theory for the quality of its predictions.

A more elaborate and more heavily iconic model is described in Hand and Crowder (2005). This combines the standard form of empirical predictive model with a latent factor model (where this can represent the 'quality' of the customer from a credit perspective), into a multiple-indicator-multiple-cause (MIMIC) model. It relies on the notion that some observed features of an individual can only influence their 'credit quality' (i.e., age) while others (i.e., months in arrears) are consequences of it.

Conclusion

The preceding section began with a discussion of the heavily empirical conventional and widespread model form used in credit scoring, and gradually moved towards models which were more iconic, making use of proposed theories describing how customers behave or how classes are defined. Each of the two classes of approach has its strengths and weaknesses. Empirical models can include, as predictors, things which have a relationship with the response, even if there is no obvious reason for the relationship. This is a strength, because it picks up patterns we would not think of. On the other hand, it is a potential weakness because the relationship might be a consequence of an unmeasured other variable and because circumstances may change. This risk of changing circumstances is the biggest weakness of empirical models. If circumstances change dramatically, then such models may prove very poor.

The move towards iconic models parallels the growing interest in behavioral finance. In this context, Montier (2003) says: "Like all economists, I have been brought up to judge models not by their assumptions, but rather on the strength of their predictive accuracy. That is, an economic model may have assumptions that are patently at odds with reality, however, so long as it predicts the way in which people act, then its assumptions are irrelevant." Clearly he is describing empirical models. A clear illustration of empirical model building in economics is given by Hoover and Perez (1999). In contrast, behavioral finance is based on a recognition that people do not behave as the idealized rational agents of classical economics, but are subject to all sorts of irrational whims and cognitive errors. With this as a basis, researchers in this area are seeking to construct economic models which better represent the way people

behave. To that extent, their aim is the same as that espoused in this paper.

The distinction we have drawn between empirical and iconic models is also related to that between structural and reduced form models, as used in credit risk modeling of corporations. Structural models require data on the operation and status of the company, as would be known by the people running it, and so correspond to our iconic models. In contrast, reduced form models are based on observation of the company's performance and so correspond to our empirical models.

Our distinction is also related to the so-called Lucas critique [Lucas (1976)], which essentially says that simplistic approaches to modeling systems, and taking actions on the basis of those models, may fail because the very action may change the nature of the system. Instead, one has to construct a deeper model, reflecting how and why the system behaves as it does, a model which is invariant to the policy one may adopt.

One can think of empirical models as focusing on what the system does, and iconic models as focusing on why the system does what it does. By concentrating on the latter, not only is deeper understanding gained (which may not matter, if the aim is simply predictive accuracy), but models based on theoretically sound relationships are likely to be more robust than merely empirical ones. To take a final example, physics tells us that taller people are likely (on average) to be heavier than shorter ones. This (statistical, on average) relationship is likely to continue to hold as a population's diet improves. But a purely empirical relationship, for example, one based on the observation that, in the past, weight was positively correlated with position in society (because of the better diet) may break down as things change. Almost all current statistical modeling in the retail financial services sector is based on heavily empirical models. Recent changes in the economic climate suggest that more pronouncedly iconic models might have much to offer.

References

- Banasik, J., L. C. Thomas, and J. N. Crook, 1999, "Not if, but when will borrowers default," Journal of the Operational Research Society, 50, 1185-1190
- Box, G. E. P., 1979, "Robustness in the strategy of scientific model building," Technical Report, Madison Mathematics Research Center, Wisconsin University
- Box, G. E. P. and W. Hunter, 1965, "The experimental study of physical mechanisms," Technometrics, 7, 57-71
- Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, 1984, Classification and Regression Trees, Wadsworth, Belmont, CA
- Chandler, G. G. and J. Y.Coffman, 1979, "A comparative analysis of empirical versus judgmental credit evaluation," Journal of Retail Banking, 1, 15-26
- Cox, D. R., 1990, "Role of models in statistical analysis," Statistical Science, 5, 169-174
- Crowder, M. J. and D. J. Hand, 2005, "On loss distributions from instalment-prepaid loans," Lifetime Data Analysis, 11, 545-564
- Crowder, M. J., D. J. Hand, and W. J. Krzanowski, 2007, "On optimal intervention for customer lifetime value," European Journal of Operational Research, 183, 1550-1559
- Dawes, R. M., D. Faust, and P. Meehl, 1989, "Clinical versus actuarial judgment," Science, 1668-1674
- Experian-Scorex, 2004, http://www.experian-da.com/Web/News/Newsletters/horizons-NAmerica_04fall.pdf
- Fair Isaac, 2004, http://www.fairisaac.com/NR/rdonlyres/73A44BBC-54DD-48C5-8544-ACAD129E97F2/0/NextGenFICO_CBRisk_PS.pdf
- Grove, W. M., and M. Lloyd, 2006, "Meehl's contribution to clinical versus statistical prediction," Journal of Abnormal Psychology, 115:2, 192–194
- Grove, W. M., D. H. Zald, B. S. Lebow, B. E. Snitz, and C. Nelson, 2000, "Clinical versus mechanical prediction: a meta-analysis," Psychological Assessment, 12, 19-30
- Hand, D. J., 1985, Artificial intelligence and psychiatry, Cambridge: Cambridge University Press
- Hand, D. J., 1994, "Deconstructing statistical questions (with discussion)," Journal of the Royal Statistical Society, Series A, 157, 317-356
- Hand, D. J., 1996, "Statistics and the theory of measurement (with discussion)," Journal of the Royal Statistical Society, Series A, 159, 445-492.
- Hand, D. J., 2008, "Mining the past to determine the future: problems and possibilities," to appear in International Journal of Forecasting
- Hand, D. J. and M. J. Crowder, 2005, "Measuring customer quality in retail banking," Statistical Modelling, 5, 145-158
- Hand D. J. and M. G. Kelly, 2001, "Lookahead scorecards for new fixed term credit products," Journal of the Operational Research Society, 52, 989-996
- Hand, D. J. and M. G. Kelly, 2002, "Superscorecards," IMA Journal of Management Mathematics, 13, 273-281
- Hand, D. J., H. G. Li, and N. M. Adams, 2001, "Supervised classification with structured class definitions," Computational Statistics and Data Analysis, 36, 209-225
- Hand, D. J., K. J. McConway, and E. Stanghellini, 1997, "Graphical models of applicants for credit," IMA Journal of Mathematics Applied in Business and Industry, 8, 143-155
- Hastie, T. J., and R. J. Tibshirani, 1990, Generalized additive models, Chapman and Hall, London
- Hoadley, B., 2001, "Statistical modelling: the two cultures: comment," Statistical Science, 16, 220-224
- Hoover, K. D. and S. J. Perez, 1999, "Data mining reconsidered: encompassing and the general-to-specific approach to specification search," Econometrics Journal, 2, 167-191
- Kelly M.G., D. J. Hand, and N. M. Adams, 1999, "The impact of changing populations on classifier performance," Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ed. Chaudhuri, S., and D. Madigan, Association for Computing Machinery, New York. 367-371
- Lehmann, E. L., 1990, "Model specification: the views of Fisher and Neyman, and later developments," Statistical Science, 5, 160-168
- Li, H. G. and D. J. Hand, 2002, "Direct versus indirect credit scoring classifications," Journal of the Operational Research Society, 53, 1-8
- Lucas, R., 1976, "Econometric policy evaluation: a critique," Carnegie-Rochester Conference Series on Public Policy, 1, 19-46
- Meehl, P. E., 1954, Clinical versus statistical prediction, University of Minnesota Press, Minneapolis, MN
- Meehl, P. E., 1986, "Causes and effects of my disturbing little book," Journal of Personality Assessment, 50, 370-375
- Montier, J., 2003, Behavioural finance: insights into irrational minds and markets," Wiley, Chichester
- Myers, J. H., and E. W. Forgy, 1963, "The development of numerical credit evaluation systems," Journal of the American Statistical Association, 58, 799-806
- Sewart, P., and J. Whittaker, 1998, "Graphical models in credit scoring," IMA Journal of Management Mathematics, 9, 241-266
- Stanghellini, E., K. J. McConway, and D. J. Hand, 1999, "A discrete variable chain graph for applicants for bank credit," Journal of the Royal Statistical Society, Series C, Applied Statistics, 48, 239-251