This house believes ..... that Credit Scoring methods and applications are stuck in the 1980s.

Jon Hinder, Rhino Risk, Gillian Groom, Southampton, University, Professor David Hand, Imperial College London, Nigel Butler, Commonwealth Bank Australia

ABSTRACT

Credit scoring has been used as a technique to assess credit risk in the UK for over 20 years. But have the methods advanced in this period to meet the needs of a more sophisticated lending environment? In this session we will debate the changes in the development, implementation and ongoing management of credit scoring. Jon Hinder will lead the debate, and proposes the motion that, fundamentally, very little has changed in the tools and techniques underpinning our Industry, and worse, that the tools that are available are not being properly used. Jon believes we are still in the equivalent of the late 1980. The 1980s spawned the mullet, “Thatcherism”, the New Romantics and the industry of credit scoring. Fortunately, hairstyles, politics and music trends have moved forward. But credit scoring hasn't.

Gillian Groom does not agree with this perspective. Credit scoring in the 21st century has changed since its earliest applications in the 80’s. The quantity and quality of the data has increased, tools have been developed to facilitate model development and academic research has identified the “best” techniques for the key stages in the modelling process. The advances in the technology to implement and manage credit risk models quickly and efficiently have made it a key tool in the management of risk. Credit Risk is now higher up the agenda of Senior Management than it has ever been, and that, in part, is due to its constant evolution and the quality of its practitioners.

Both Jon and Gillian will draw upon expert witnesses to support both sides of the debate.

Ultimately, the audience will decide whether credit scoring has indeed changed, and to conclude whether it is important that it continues to do so.

Imagine for a moment that you are an analyst or senior credit risk manager in a bank, or a start up credit card lender. You are approached by a young (ish) lady, who, it turns out, is the new Chief Executive. She politely asks your name, and takes a seat next to you.

So you're the guy who manages our credit risk exposure?

Well, yes – myself and the rest of the team here.

I know little about credit risk (she came up the MBA Business School / McKinsey route. But whilst she knows little about the technical detail of credit risk management, she knows a lot about the right questions to ask).

But please tell me, these scorecards we use to manage our risk. They will work in a downturn won't they? I see the estimates of arrears and default rates in my monthly Executive Board MI pack, but I don't seem to have anything that tells me our likely exposure if the recession worsens.

We do produce downturn estimates, but they are not exactly what we could call exact. Estimating both the model performance and our overall losses in the current environment is a really difficult problem.

I understand – but we do have historical performance data to use as a guide, don't we? And what about credit bureau data? Surely those guys must be able to help?

Well sort of. Except they have only kept data for a few years. The best we can do is come up with a good mean estimate through the current models we have, based on recent history, and try to extrapolate or benchmark to get to a longer term or even a downturn estimate. We really don't know how our models will perform in the recession, and what are final losses will be. And to be honest, neither do the consultants we use, academics or the Regulator.

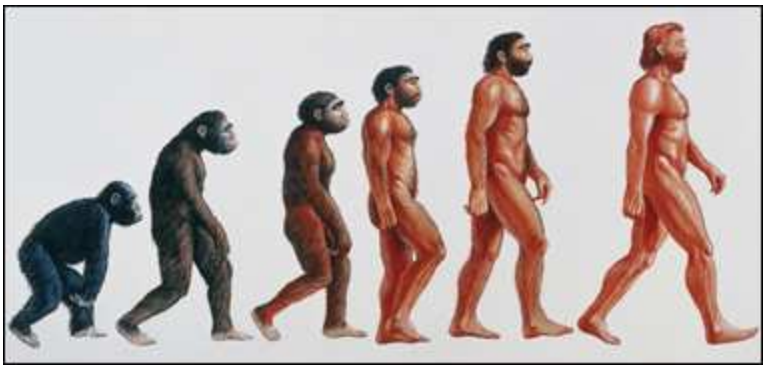
But I've been to classes in Business School where we were taught that true Risk Management is not about estimating the likely losses, but in understanding the range of outcomes, of understanding unlikely losses. Surely if a novice like me understands this is the real issue, then you guys who are the experts should have cracked the problem by now. After all - you have been doing this for a long time now - I read somewhere that credit scoring has been used in banks since at least the 1980s. You've had over 25 years to get this right.

She leans over and sees a folder with details of an upcoming Conference on Risk Management. She flicks through the abstracts. She sees workshops led by academics on heuristic modelling and quantile regression. She wonders whether genetic algorithms know they are in a fight with generalised additive neural networks for space at the top table of cutting edge methodology. She laughs to herself at the thought of a support vector machine meeting a Bayesian network.

And she sees very little on macroeconomic modelling and forecasting, and even less on understanding model error and uncertainty – what she thought was true risk management.

Is she right? Should our “Industry” be damned for not being able to crack, or worse, not even attempting to crack, the difficult problems? Do we really understand the fundamental concepts of credit risk management, and have we fully taken advantage of new methodologies and technologies to really advance both our intellectual capital and our practical knowledge?

We have had at least 25 years to get good at this. There is a famous poster that shows the evolution of man – in easy steps from chimp to homo sapient. If this portrayed the evolution of credit risk management, I suggest we would still be dragging our knuckles on the floor.



The Ascent of Credit Scoring - How far have we evolved?

Before I get too into this, I am not talking about credit risk management in its widest sense. There have been enormous changes in areas such as derivative modelling and securitisation to diversify risk, advancements in option theory, portfolio theory and so on. I am talking here about the bread of butter of consumer credit risk management – the development of credit scoring techniques and their application and use.

Most histories of credit scoring would start in the 1950s, after the “science” of was Operational Research was developed to support the logistics of moving people and supplies. Most histories would suggest that it took off as a seriously accepted tool in the 1970s and 1980s. Most large banks in the UK, for example, were using application scoring by the late 1980s, and had started to develop their behavioural scoring models and associated IT delivery systems at this point. Since then, most lenders use credit scoring in most lending environments, even if a few “old school” pockets of manual underwriting remain.

The basic assumption behind credit scoring, of course (or at least the one cited in most training manuals) is that “the future is like the past”. But that is a bit like basing the whole of economic science on the assumption that consumers always behave rationally. We do not, and it isn't.

Going back to the introduction to this paper, the new Chief Executive highlighted a particularly tough issue to tackle – namely how do we understand how models predict in a downturn? At this point, I would suggest that if we were in a Hollywood film, then there would be no problem. Someone very serious from a Government Bureau would tap me on the shoulder, and say: “Jon, don't worry. We have known about these issues for many years and unbeknown to you, we've taken our best minds from the Industry, and holed them up deep in a bunker, ready to emerge with the new models and theories, just so we could tackle these problems”. But this isn't a Sci-Fi film, and of course, they haven't.

Premise No 1 – we have focused on the wrong things

Why do I think as an Industry we are “Stuck in the 1980s?”

To help me, I have looked at the following stages that most banks will go through in the development of their credit scoring models. I suggest there are seven key stages involved:

1. Data preparation
2. Outcome definition
3. Sample selection
4. Modelling (including characteristic analysis and the model “run”)
5. Model validation
6. Model error estimation
7. Ongoing model monitoring

What is interesting is that if I had set these out about five years ago, before the advent of the Basel II Capital Accord, then step 6 would probably have not been here. But this is actually the key step (with the possible exception of the data preparation stage). What do I mean by this model error estimation? To me it can be viewed as a catch all for trying to understand and estimate all the unknowns around the model performance. This therefore brings into play three very interesting, and difficult, areas of modelling:

- What level of conservatism do I need to ensure my model estimates are robust, relying as they often do on non perfect data and samples which are always only a best estimate of the population?
- How can I estimate my model's performance through an economic cycle?
- If I have less certainty around my model (and hence loss) estimates, how can I use stress testing, simulations or other forecasting techniques to help me understand this?

In general, the techniques used to derive the outcome definition and sample selection have hardly changed over the past 50 years. Sure, there have been endless debates and academic research into various missing data classification problems (reject inference being the most well known), but, by and large, these areas are now well known. The standard modelling tools such as logistic regression, linear regression and decision tree approaches have been under attack from the aforementioned methods such as neural networks and genetic algorithms. But again, by and large, they have stood firm, and the modelling techniques used in the 1980s have remained the “standard”. Model validation has improved, but is still fundamentally based on the twin concepts of a performance measure (gini, KS and such like) and a measure of alignment (actual v expected, using statistical tests such as Homer-Leadership).

So my fundamental question is this. Given that the key techniques in most areas have remained the same as in the 1980s, why has so little emerged in the way of new methods and theories to help us with the key challenge of understanding model error estimation, particularly in the context of changes in the economy?

To support this argument, I have looked back at the past years' papers submitted to this conference, from 2003, 2005 and 2007. I have broken these down into various topics, and in doing so, was primarily interested in seeing where our Industry has focused its efforts over the past eight years. The first seven categories mirror my seven stages of model development. However, there are many others, focusing on other broader areas of credit risk management. My findings are set out in the following table:

	2003	2005	2007
Data preparation (including bureau analysis)	4	2	2
Outcome definition (including reject inference)	2	2	3
Sampling design			
Modelling method (1)	14	19	21
Model validation	2	3	4
Model error estimation	2	4	9
Ongoing model monitoring			
Credit strategies / optimisation	6	8	4
Profitability / revenue modelling	8	4	3
Others (fraud, regulation, analysis of lending markets, non retail)	9	20	31
(1) Subset - Small sample modelling / survival analysis	3	3	3
Total	47	62	77

In 2003, there were fourteen papers which focused on model methodologies. Three of these were on topics such as small sample modelling and survival analysis (I have stripped these out as being of particular interest). The rest, from what I can see, were pretty much aimed at trying to maximise the power of the models, by researching new methods. These compared to only two papers tackling what I think of as the most difficult, and actually, the most important aspects of credit modelling - namely model error estimation. I may be harsh here, but the areas of modelling which incorporate the challenges of model error estimation, particularly when applied to future scenarios, such as incorporating economic data, understanding how models work through a cycle, and having litigants in the form of stress testing and forecasting models, are exactly the areas that separate the chimps from the fully upright homo sapient of the credit risk world. And only two brave souls back in 2003 had enough foresight to focus on this. In effect, in 2003, most in the Industry (lenders, consultants and academics) were doing what I call “Chasing the Gino”.

Turn to 2005, and the pattern is similar. There is still a strong focus on optimisation and profitability modelling, and many papers were tackling issues such as fraud, or providing insight into areas such as lending policy in emerging markets. But there is also a huge number presenting papers on modelling methods. Nineteen to be exact. Out of 62. That's 31%, of the three day conference devoted to presenting, discussing, debating and ultimately rejecting a variety of modelling methodologies. I say “rejecting”, because very few of these are now used in the mainstream. To be fair, 6% was devoted to tackling the tough questions I raised earlier. That's an improvement from the 4% in 2003.

By 2007, we see some interesting things happening. Firstly, there are far more papers which I have not classified specifically, and many of these relate to small business lending, and in larger numbers, Basel II modelling. In addition, there is a large rise in the number of papers focusing on my “catch all” group labelled model error estimation. The majority of these in 2007 were looking at the methods to incorporate economic data directly into the models themselves, or addressing the concept of how models deviate through an economic cycle.

Here, I make two interesting observations. Firstly, whilst it is pleasing to see this focus in 2007 on these tough modelling problems, it is interesting to note that this is around the same period when the lenders needed to tackle this issue to meet the requirements to estimate the long run average PD and downturn LGD under the requirements of the IRB approach under Basel II. In other words, the Industry has reacted here to the regulatory bodies requiring further effort in this area. As we have previously seen, this problem only represented a small proportion of the academic effort in previous years, and I therefore cannot help wondering why it took a regulatory prod to kick start this research, when, to be truthful, credit scoring and risk modelling has required this for over 50 years. And secondly, in 2007, the whole area of stress testing, simulation and forecasting in general was still not on the Industry's radar, with only one or two exceptions.

If we look at the 2009 papers, there remains preponderance for presentations that focus on subjects such as decision trees, clustering, optimisation theories and so on. And, I my view, still not enough on the hard topics that really matter.

Premise No 2 – we have not focused on the fundamentals

My concern that we are not sufficiently forward looking as an Industry would be mitigated if I thought that we were at least doing the basics correctly. Unfortunately, we are not. I see banks, for example, that implement models without a clear validation plan.

I have seen some lenders implement an application scorecard and wait a full twelve months until they monitor its performance (one does not need a full twelve months to generate a very good picture of whether the model is performing to expectations). In fact, I would argue that with between four to six months outcome, the key measures of model performance power and alignment will emerge to a sufficient degree of tolerance. This means that allowing for the twelve months for the sample outcome period, and the standard four to six months for development, implementation and testing, these lenders are using models built on data which is at least two and a half years out of date, without performing an out-of-time validation.

One interesting angle to model validation is that lenders often see a drop in their model predictive power – say a gini drop from 75% to 70% – between the development sample and a subsequent through the door validation sample. They may ask me, why has this happened? Or is it a cause for concern? A quick check shows that the development sample actually contained a relatively small sample of defaults. A quick statistical check then shows that a confidence interval for the Gini actually estimates this to be in the range 68% to 82%. With this additional information, the “drop” from 75% to 70% has a completely different context (there may of course still be underlying reasons for the reduction in model power – but one of these could simply be that this was overstated in the first place).

Lenders have also made huge mistakes in not realising that the “actual v expected” is the name of the game, and not, as they assumed “chasing the Gini”. What I mean by this is that it is better to have a weakly predictive model, but to fully understand its limitations, potential errors and hence its ability to accurately estimate risk (loosely we can call this alignment, or the ability to predict actual losses against those expected), than it is to have a more strongly predictive model, but which breaks down when trying to accurately predict losses across its grades.

And I believe in the idea that a little knowledge is a dangerous thing. I sense within the Industry that because Senior Management can hang their hat on a statistical measure of performance, this creates a false sense of security. My Gini meets Internal Model Standards? Great! My Gini is bigger than yours? Even better! Of course, the model performance measure says nothing about their overall credit policy or strategy, or as I

have suggested, the real issues at hand – such as its possible range of error and its ability to continue working under economic stress.

And let me leave you with one final thought.

Imagine a large bank which has a mortgage portfolio of 100,000 loans, with a historical Probability of Default prediction of 1%. They lend across all regions of the country, to a wide range of customers. Imagine a second lender, a regional Building Society, again with 100,000 loans, and a historical Probability of Default of 1%. However, this lender only lends within their core region, say a 25 mile radius. In addition, most of their lending is to two or three key customer groups – young customers perhaps, or because of a particular marketing focus, they have a high proportion of teachers and other professionals. Both have the same "raw" PD. However, their real risk, what we can loosely be described as Unexpected Loss, may be quite different. We call this problem Concentration Risk. The first portfolio is pretty well diversified. The second is not.

I raise this example because it is a real world problem. In some instances concentration risk has recently been a significant factor in causing lenders to fail. And yet it is a problem for which there is no "best practice" method. And worse than there being no standard method to address this problem, there seems to be little – in fact from my current understanding, zero – research from academics, consultants and practitioners in this area (forgive me please if in fact one or two readers have indeed focused on this topic).

So how could I not conclude that we have failed in focusing on the real issues?

What do you think? Am I being unnecessarily harsh? Do you agree we are, metaphorically at least, stuck in the 1980s?

Jon Hinder

Thanks Jon, good to see you have such confidence in our people and techniques.

To be fair, you raise some interesting points. But I cannot let you get away with painting what is essentially a simplistic conclusion when we can clearly see significant innovation over many years.

Time has not stood still in the credit risk world. In fact the evidence presented in this paper shows that there have been significant improvements in many aspects. In order to focus on the key issues my argument will focus on four areas I call the 4 “T”s of change, namely Techniques, Technology, Training and Talent.

1. Techniques

First let us study the techniques of the scorecard development process. It is true that the techniques at all stages, including sampling, performance definitions, reject inference and modelling have been well established and to some extent standardised over the last twenty years. However this does not mean that extensive research and development has not been completed and continues to be undertaken to ensure that all the benefits of new data, technologies and analysis can be fully exploited.

As you quite rightly point out getting a bigger Gini does not necessarily improve the whole credit risk picture. In some companies the focus of the analytics has moved away from just the scorecard on to the bigger picture of the strategies for acquiring and managing customers in a structured manner. Organisations like Capital one have invested time and resources into developing and analytics based test and learn strategy for evaluating all aspects of the life cycle of the credit card customer.

2. Technology

The next area of interest is technology. This can be looked at from three vantage points. Firstly the technology available for the development of scoring models, secondly for the implementation and finally the management of the scoring process.

Twenty years ago, if you wanted to build your own scorecards you needed to use generalist software packages such as SAS or SPSS. The learning curve to develop the processes needed was steep. The time spent to develop scorecards in house was also compounded by the limitations on computer power. Admittedly this approach did have the advantage that analysts really had to understand the art and science of scorecard building.

Nowadays there are a range of customised credit scoring tool kits available, which provide for standardised approaches, based on “best” practice, to course classing, reject inference and model development. This combined with the computer power available on most desktop PCs, mean that scorecard analysts can focus on interpreting the results from the scorecard development thus ensuring that all the questions surrounding the development are understood and robust, reliable models are built. I will agree with you on one point, that this level of automation can in some circumstances result in models being built by the push of a button and this can be dangerous.

Technology also has had a significant impact on how scorecards are deployed. Early scorecard implementations were hard-coded. This meant that any changes that to a scorecard, however minor, resulted in major pieces of IT work. It was soon realised that to manage risk effectively it would be necessary to have the capability to change scorecard cut-offs, scorecard characteristics, attribute weights and policy rules. Nowadays new scorecard deployment software allows for changes to these features to be made by the user instead of requiring any IT intervention. The best practitioners will be using this fantastic flexibility to pro-actively manage risk.

Scorecard models are sophisticated tools and as such need to be checked on a regular basis to ensure they are still working to their best ability. The ongoing management of these tools is the key to effective and efficient risk management. You are right that the development of organised databases and tracking reporting have been slow to evolve. Partly this has been because while times were good studying history seemed unimportant and unglamorous. Basel has resulted in companies organising the data histories and the current economic situation is resulting in the analysis to understand what happened and what can be done in the future. The analysts among us need to take responsibility to continue to drive this forward. The work in this area has only just started but it is moving in the right direction and without technology this would have been impossible.

3. Training

In the late 80's and early 90's the tools available to train analysts in the concepts of the development and management of credit risk scorecards were very limited. At this time there was no Amazon, (I keyed in the words Credit Scoring and on a day in early July it resulted in a list of 92 books). Even if Amazon was around the number of books would have been far less. As recently as 1994 when I joined Fair Isaac the only reference book I was aware of was an "Introduction to Credit Scoring" by E M Lewis.

The other opportunity for learning was on training courses, again in the late 80's and early 90's the few training courses available on credit scoring were run by suppliers and inevitably these focused on the tools they had to offer to cover various aspects of scorecard management.

Now in addition to the courses offered by the suppliers of scorecard software tools there are a range of courses run by "independent" suppliers. Several consultancies lead the way in this, offering courses that trained scorecard developers on the important statistical concepts that underlie successful scorecard development. In addition universities such as Edinburgh and Southampton offer specific modules on topics such as Credit scoring and Basel.

4. Talent

My final observation about credit risk is regarding the investment in Talent. In the late 80's and 90's the pool of talent available with any knowledge of credit scoring was extremely limited. This was a critical time for companies to invest in talent to understand, promote and implement the use of credit scoring. This was a time of rapid change. It was the start of the buy now, pay later culture. Building societies were becoming PLC's and starting to dabble in unsecured lending and the personal loan became much more popular. Unfortunately the investment in marketing these products far outstripped the investment in the tools and talented people to manage the credit risk. At this time very few companies had credit risk represented as a standalone function on the main board.

Nowadays, and quite rightly so, credit risk is one of the main functions of financial services businesses, with all the power and significance of other functions such as operations, marketing and human resources. The credit risk teams have investment in the tools and people they need to ensure that the credit risk of the organisation is effectively managed.

These are four areas which have made considerable advances over the past 25 years. And to be honest Jon, you seem to have been pretty selective from your analysis of the past presentations at this conference. Let me give you two examples of where advances in technique, has considerably benefited our Industry – including the end consumer (who you seem to have missed from your earlier discussion).

Credit Bureau Data

Probably the one area of indisputable progress has been in the use of credit bureau data to aid credit risk decisions. Do you remember a time in the late 80's when only negative credit data was available, then positive data became available from the small companies and eventually positive data became available from all companies? Credit bureaus now hold some information on over 45 million consumers in the UK. The growth in the availability of credit bureau data both in terms of breadth and depth of data available has given lenders the ability to objectively measure affordability for individual consumers, to provide online real time credit decisions and many other innovative opportunities. Are you seriously telling me that this has not made a significant contribution to our underlying methodologies?

This data has resulted in the following improvements in the management of risk.

- Bureau scores and consumer profiles mean that lenders can make some evaluation of risk on all applicants without any prior knowledge.
- Without the improvements in credit bureau, it would be impossible for lenders to make instant credit decisions.
- Bureau data is the first defence against fraud for both the lender and the consumer.

Fraud

Fraud is now a significant financial risk, according to APACS in 2008, Card fraud losses totalled £609.9m, online banking fraud losses £52.5m and cheque fraud losses £41.9m. Analysts quickly identified that some kind of scorecard modelling could be useful in identifying fraudulent transactions, however they also quickly realised some key differences between predicting fraud and non-payment of credit. Specific challenges included:

- Large volumes of data and relatively small volumes (in terms of numbers, not value) of fraud.
- Short outcome windows in terms of minutes not months.
- Rapid evolution in the patterns of fraud, requiring models that could adapt to change.
- Complex relationships between predictive characteristics.

Standard scorecard models were tried, but soon analysts started using neural networks, machine learning, genetic algorithms and, model boosting to identify potential fraud and therefore trigger manual interventions to limit fraudulent transactions.

And they are adding real value. You may categorise these techniques as having been "rejected" by the Industry. But far from it. They are alive and kicking, but your perspective is too narrow to see this.

Some of what you have discussed Jon is valid, but much is based on hindsight.

Stuck in the 1980s? Dragging our knuckles? Far from it.

Gillian Groom

Thanks for the response, Gillian.

To pick up on a few points, I firstly want to discuss "Talent" within our Industry. Whilst I do see progress in this area, I am still often surprised by the basic lack of understanding of credit risk principles at senior levels, particularly in smaller organisations. Very few of the very senior people in our banks have come through the "credit risk" route. Most have backgrounds in operations or finance. Lending is at the very heart of what these organisations do, so it is disappointing that more from our Industry are not running these organisations. And from another perspective, I think it has been a failing from those in the most senior credit risk positions not to push this higher up the agenda.

So our "Talent" may have improved, but not sufficiently, and in large enough numbers to play the key influencing roles that are required. And I would like to make a further point related to "Talent". Perhaps the closest sibling to credit scoring as an Industry is Actuarial Science. Yet to be a professional actuary, one must study for post graduate professional exams, which are exceptionally difficult both in terms of the time effort involved (typically 5 years) and the technical acumen required. We have no similar professional body, and certainly no formal qualifications.

Secondly, I want to touch on Technology. I think you are correct, that there have been huge leaps in this area. My concern though is that as we move to easier and easier tools, the "art" of model development and validation is lost. I think it is no accident that some of our best practitioners started their careers at a time when everything was hard coded, and they were forced to fully understand all stages of model development and validation from first principles. They were also forced to fully understand the data, and to make extensive checks at each stage. And because they went through this process, they were better able to subsequently validate and monitor models because they knew them inside out. It's a bit like being an old fashioned mechanic. Nowadays, cars have complex electronics to indicate exactly what is wrong. The problem is what happens when the electronics itself goes wrong? Or what happens if it indicates something "out of the box" which the mechanic has not seen before?

At this point, I would like to bring in Nigel Butler

Jon Hinder

I agree with Jon that there remains a concerning degree of risk complacency and lack of understanding in the senior executives of many organisations.

The issue is not so much one of capability, it is one of application and I agree with the principle that the fundamentals around the use of credit scoring have hardly changed since it's breakthrough in the 80's.

Yes we have more models, which are based on more data. They are more accurate and in the hands of good practitioners are monitored more often. However, the primary focus remains on what is the most likely outcome of the model (as based on historical data and statistical likelihood) rather that what is the potential for the models "getting it wrong", which is surely what a good risk manager should be most concerned with?

The argument is that our focus as an industry is misplaced, with increasing focus on model accuracy and

insufficient focus on the implications of model vulnerability; whether that be expected statistical error or the potential for more extreme variation due to changes in data and/or changes in external circumstance (such as economic or sociological shifts). We know to increasing degrees of certain what is *most likely* to happen but when the question is posed "*what will happen in the eventuality that the data supporting the model assumptions becomes distorted*" we are left with, at best high degrees of uncertainty and most likely with "no idea". What Jon calls "model error" I call " model vulnerability".

In reference to Jon's seven stages of modelling, validation is considered to be the standard process for gaining confidence in the model's capabilities. Sophisticated modellers have for years used a range of validation techniques which are both within sample and also and out-of-time validation, to assess accurately how the models will perform. However, all of this is done, unconsciously or otherwise, on the premise that the data used represents the total universe of potential outcomes.

As the extremes of the recent economic events have revealed, this premise is flawed. However, a global economic crisis is not necessary to justify the argument. We are all aware, and have been for years (even if we have consistently chosen to ignore it) that the models are flawed due to the limitations of the data universe used to develop them. And yet we have not really addressed this, most often because we assume that we can "do nothing about it" but the reality is that this is not true and this is where progress has been slow.

Stress testing and/or scenario analysis is not a standard technique in the scorecard developers' and/or decision-makers toolkit, but is instead generally undertaken as an "after the fact" regulatory exercise. It is in this area, which I reiterate is what credit risk managers really need to understand, that the industry remains in its Neanderthal stages. The fact that the data is not readily available, and that techniques are not standardised to address this problem supports this view.

Nigel Butler

Jon and Nigel, I am not persuaded. Our people are better trained than ever. And the whole point of having this technology is that it frees people up to focus on the very things that, as you suggest, really matter. Such as formulating strategies, communicating with senior management and influencing the business as a whole. Would you rather have our brightest minds be tied up in spaghetti code than actually shaping the business?

Moreover, I think some of the points you previously raised are not valid. As you suggest, there is now a significant effort from the Industry in understanding some of the more challenging areas, such as model uncertainty and their performance through the economic cycle. By their nature, these are difficult issues, and to a large extent their solution relies on data that most lenders do not yet have.

I am actually encouraged by the variety and complexity of most of the research and investment in our Industry, as evidenced from your analysis of eight years of Credit Control Conferences at Edinburgh. Our knowledge is growing, and we are in a better shape for preparing for the future challenges than we have ever been.

And to support my argument, I now turn to some insights from Professor David Hand.

Gillian Groom

Thanks Gillian. Jon makes some excellent points, backed by Nigel's insight, and I'd like to congratulate them on drawing attention to a number of things which deserve consideration. In my view it is valuable for the industry to step back and take stock like this; to ask whether we are doing the best we can.

I did not interpret what Jon said as meaning he thought that there had been no innovation. This would clearly be a false assertion, which Gillian effectively demolishes via a variety of examples. I can also give many others. One only has to look at the history of innovatory companies such as Capital One and Fair Isaac to see how the level of technology has moved on from the early days. Moreover, consider the way the notion of credit scoring has widened. Originally it was merely aimed at application scoring, but now it includes behavioural scoring, fraud scoring, profitability scoring, and a host of other things, and typically not in the context of a simple single score but in elaborate segmented scorecards constructed by implicitly comparing millions or even billions of models, on vast data sets.

Instead, I interpreted Jon as asserting that the focus in the industry, the concentration of effort if you like, had been misplaced. I have a lot of sympathy with this position.

In fact, in 2006 I published a paper in the journal *Statistical Science* which discussed these sorts of issues. I focused on classification methods, rather than scorecards, but since one important use of scorecards is to assign customers to classes (e.g. accept/reject classes) this is very relevant. Like Jon's 'Premise 1', my paper argued that we had focused on just one narrow aspect of the problem, improving methods for application under highly restrictive assumptions. In particular, I suggested, the research community had put too much effort into designing classifiers to optimise narrowly defined measures of performance, such as misclassification rate, the Gini index, and so on. I noted that a large number of comparative performance studies had been conducted to compare methods, but argued that these studies failed to take account of important aspects of real problems - such as non-stationarity, selection bias, varying class definitions, and so on. Jon makes very similar points. To drive home my point, I gave the paper the title 'Classifier technology and the illusion of progress.' (Hand, 2006)

Jon also draws attention to what he terms 'model error' and Nigel 'model vulnerability'. I think this is the same as what statisticians call 'model uncertainty'. It refers to the fact that we estimate our predictive models, our scorecards, from past data, and then use those models to make predictions and decisions, but we generally do not take into account the fact that the models are just estimates.

There are two aspects to this. The first is the familiar statistical one, namely that given a particular model form (e.g. a logistic regression) then different models would be produced if we had different historical data. This revolves around well-understood issues of sampling variation, and I think everyone here will be familiar with this. This aspect means that we can produce confidence intervals for our predictions. Such error bounds are important for portfolio management and assessing likely future risk. They are less important, however, if the model is being used to yield estimates for individual people, for accept/reject decisions, for example. Here one just needs a single best estimate.

The second aspect is more interesting. It is that the model form itself - the logistic regression, for example - is not provided by nature, but is a particular form we have chosen to adopt. There is thus intrinsic uncertainty about the very family of models.

In this context, Jon then asks 'why has so little emerged in the way of new methods and theories to help us with the key challenge of understanding model error estimation, particularly in the context of changes in the economy?' I'll return to the qualification 'particularly in the context of changes in the economy' later, because I think I agree with him about that. But let me first focus on the broader question of 'why has so little emerged in the way of new methods and theories to help us with the key challenge of understanding model error estimation?'.

The answer lies in the nature of the models used in credit scoring. Broadly speaking, in statistics there are two distinct kinds of models (Hand *et al*, 2008). They go under various names, but here I shall use the terms *iconic* and *empirical*. Iconic models are based on some underlying theory: for example, one might postulate a relationship between force and acceleration and then use statistical methods to estimate the parameters of the relationship. Empirical (predictive) models are based on simply building the best model one can, where best is defined in terms of some measure of predictive performance - Gini, for example. The development of credit scoring over the past 40 years has been an extraordinary triumph of empirical modelling.

The point is that whereas iconic models reflect an underlying theory, which one might regard as an attempt to describe 'reality', empirical models do no such thing. While one can experiment with different empirical models, it is not clear that it is right to speak of 'model uncertainty' in the same way. One can certainly average over a collection of models, and highly effective statistical strategies, of both Bayesian and frequentist orientations, do this. A simple example would be the use of random forests - certainly one of the most effective of statistical prediction tools.

Returning to Jon's qualification 'particularly in the context of changes in the economy', I think there is force in this. There certainly has been some work on this in the past, and Jon's statistic from the last three Edinburgh conferences, suggesting that Basel II has stimulated an increase in the amount or work is interesting.

I am also in broad agreement with Jon's second premise. He described the (appalling) example of a bank implementing a scorecard and waiting 12 months to see how it was doing. In my view, effective assessment and evaluation is at least as important as the model itself. In fact, I think it is so important that I have written two books on model evaluation (Hand, 1997, Krzanowski and Hand, 2009).

However, where I disagree with Jon is that 'actual v expected' is the 'name of the game'. The fact is that the 'name of the game' depends on what you are using the model for. Choosing between actions (e.g. accept/reject), estimating default probabilities, estimating overall portfolio losses, and so on, are different

usages, and require different performance measures - and probably different models. A model which is good for one type of application may not be good for another. If we are making a decision based on comparing a score with a given threshold, then all that matters is whether the score is above or below the threshold, not how far it is from the threshold. Remember, again, we are not trying to model 'reality'. We are simply trying to predict what might happen, and with some objective in mind. We are using empirical models.

Incidentally I need to correct a misunderstanding by Jon's hypothetical CEO, who says 'you guys are the experts [so you] should have cracked the problem by now'. The fact is that we have. But, perhaps because (according to Jon) she came up the MBA Business School / McKinsey route instead of having a proper quantitative training, she misunderstands the nature of the problem. I may be an expert on probability, and we may have cracked the problem of what probability means and fully understand it, but that does not mean I can tell you whether a tossed coin will come up heads or tails. If the CEO expects us to do that it means she doesn't understand the very nature of probability.

Jon also says that the basic assumption behind credit scoring is that 'the future is like the past'. I have two comments on that. This first is to ask what else you are going to base your predictions on? That the future is like nothing you have ever seen before? And the second is to say it depends what you mean by 'like'. This is a bit more subtle. Suppose I have observed, in the past, that people with particular characteristics default 10% of the time. Then I could define 'like' as those with the same characteristics. But I could also define 'like' as those with the same characteristics adjusted for the currently prevailing economic conditions, or any other factors I thought relevant (Hand, 2008). 'Like' can be measured by various metrics and effective prediction hinges just as much on choosing a good metric as it does on choosing a good model.

Professor David Hand

O.K. David and Gillian, a balanced argument can be made that we have made progress. We can all perhaps broadly agree that:

- Techniques to develop more accurate models have certainly approved
- The capabilities to build such models quicker and easier have also improved
- Data quality is constantly improving, such as the example of credit bureau data
- The models are now applied to a wider range of solutions than simply applications for credit, with particular sophistication of modelling being applied to fraud identification
- The intellectual capabilities of academics and credit scoring practitioners have progressed through experimentation, experience and training

However, I am sticking with my original premise, supported by Nigel. There remains a concerning degree of risk complacency and a lack of understanding by the senior executives of many organisations. We are still content as an industry with taking the "expected outcome" of a model as the only outcome and basing our decisions on this single, assertion. This is despite the realisation that the one thing we can be absolutely certain of is that the model will not actually perform exactly in the way we expect.

In broad terms, we have yet to really tackle the hard challenges within credit scoring, and for all intense and purposes, we remain "stuck in the 1980s".

Jon Hinder

References

Hand D.J. (1997) *Construction and Assessment of Classification Rules*. Wiley

Hand D.J. (2006) Classifier technology and the illusion of progress (with discussion). *Statistical Science*, **21**, 1-34.

Hand D.J. (2008) Mining the past to determine the future: problems and possibilities. *International Journal of Forecasting*, **25**, 441-451.

Hand D.J., Brentnall A., and Crowder M.J. (2008) Beyond empirical scorecards. *Journal of Financial Transformation*, **23**, 121-128.

Krzanowski W.J. and Hand D.J. (2009) *ROC Curves for Continuous Data*. CRC Press