monitor

Reprinted from the June, July/August and September 2005 Issues, Vol. 32, No. 6, 7 & 8

The 12 Secrets of Commercial Credit Scoring:
Confessions of a Closet Quant Jock

In this article, PayNet's Tom Ware divulges the 12 secrets of commercial credit scoring. A self-admitted closeted "quant jock," Tom's confessions move the topic from the realm of the arcane to the world of clarity.

By Thomas E. Ware

I was originally trained as a quant jock: mathematical economics, advanced calculus, statistics and even some graduate work in model building. But after a short time working in that field, I heard the siren's call of business school, MBAs, strategy and finance. So after that metamorphosis, and with MBA in hand, I set out on what has now been a 20-year career in equipment finance and banking, working in a variety of credit, operations and general management roles. Slowly, and quite unintentionally, however, I found myself over the years pulled more and more into quantitative credit scoring. At first, as the credit officer responsible for implementing scoring, it was just as an outsider looking in; but eventually, it was as a model developer or project lead on a development team.

As a customer and/or development partner, I have had the privilege of working with all the major commercial credit scoring companies including Fair Isaac, Experian-Scorex, D&B and, at PayNet Analytical Services, with the spin-offs of these firms, PredictiveMetrics and InfoCentricity. Although I will never "out calc" the PhD mathematicians, I have a unique perspective on scoring as someone who first and foremost knows the realities and intricacies of leasing, but who also understands most of the mathematics.

Not all scoring systems are equal. And while it's true that using a bicycle is better than walking, most people would prefer the power of a Corvette or, if it were available, an F-15. The differences between scores aren't quite that extreme, but they're still substantial, so finding a first-rate scoring solution is likely to have a significant impact on a lender's bottom line.

#1. "You Gotta Believe"

The mantra of the '69 Mets has broader applicability. While this may sound trivial, I have seen cases where predisposition against scoring effectively doomed a project. An airplane pilot barreling down a runway at hundreds of miles per hour just before taking off has to believe the plane will leave the ground; the alternative, at those speeds, is unthinkable, as is the option of hitting the brakes halfway down the runway. Note, however, that the pilot's "believing" is not a matter of emotional blind faith, but rather is based on indisputably tested and documented proof that aerodynamic lift works. This same degree of proof is available in the world of credit scoring, to those willing to examine the evidence.

However, not everyone is willing, and this problem is not limited to credit granting. In a fascinating book, *What Works on Wall Street*, James O'Shaughnessy documents the result of his comprehensive historical analysis of what stock characteristics produce the most favorable returns over the long run. The entire second chapter of the book, however, is an examination of emotional and psychological barriers and human prejudices against purely mathematical odds-making. He summarizes a wide range of studies specifically designed to assess the accuracy of human decision making that consistently find patterns of mistaken human decision making in medicine, politics, gambling, insurance, investment management and other fields.

People have a tendency to over-weigh their own personal experiences above that of broader samples (in Stalin's words "One death is a tragedy, a million, a statistic"). People also read patterns into totally random data sets, prefer complicated, creative explanations to simple solutions and above all, find statistics to be boring.

There is also the human "ego" component underlying the battle of Man vs. Machine, which goes back at least as far as the Luddites destroying machinery in the early 19th century, and through to recent history when IBM's Deep Blue computer, examining 200 million moves per second, defeated the human world chess champion in 1997. No Olympic runner would ever dream of trying to beat a car in a race. Credit professionals will profit more from thinking about how we can take advantage of technology, as opposed to trying to fight it.

No credit manager has ever looked simultaneously at 100,000 deals, remembered all their characteristics at time of application, researched which ones subsequently went bad and then ferreted out how important each specific factor was in creating a high likelihood of a deal going bad. How can anyone then be surprised when a model that does do all those things ends up being demonstrably more predictive than the credit manager? If it was possible for IBM to make Deep Blue do what it did, then logically the only way credit scoring wouldn't be more predictive is if a mediocre job was done in building the credit score.

#2. Understand Your Goals

One of the first and most important steps in the model building process is to define exactly what one is trying to predict. Is it the probability of default or the probability of loss? Is it overall transaction profitability?

Though related, these are quite different things. Default is relatively straightforward, but predicting loss (at least to Basel II standards) first requires predicting the probability of default, then predicting the probability of loss given default (which factors in things like term, downpayment, collateral resale value and the ability to recover monies through litigation and other means). Profitability also considers the economics of the transaction, the cost of collecting troublesome accounts even if they never actually default and perhaps even the incremental impact on vendor deal flow if the application is declined.

Most credit scoring models today, however, are simply looking at the probability a transaction will default, and the most common specific "bad" definitions are whether an account will ever reach 60 days or 90 days past due. Banks generally prefer the 60-day metric based on the idea that if an account reaches 60, but not 90, it was still probably an unprofitable account to have booked.

Most leasing companies, and especially the vendor-oriented ones, tend to prefer a 90-day definition because to them, while an account that reached 60 but not 90 was either break-even or just slightly unprofitable, being able to book the deal helped them make their vendor happy and ensure the flow of future business.

For lenders with significant payment misapplication problems, a modeler could even develop a bad definition tailored to ignore delinquencies that appeared to be just administrative (e.g., the account was 90, but the borrower had much bigger accounts that were all current). The best models also consider other circumstances in the definition, such as whether the account had to be extended, whether there was a bankruptcy (which might not show up as a delinquency for a long time) or whether the account was a real nuisance such as 6 x 30. The modeler also has the option of counting these as "indeterminates" — neither "bad" nor "good"

Another dimension is the time period — does the model predict the probability of default within the next year? Within the next two years? Though users generally prefer long time windows, ideally wanting to know if an account will ever go bad, practical considerations usually reduce the performance time window to a year or two. First is the problem that the further out in the future one goes, the harder it is to predict what will happen, as there may be no clues today as to what will happen four years from now. And reaching further out is likely to degrade performance in the more important early years. Second is the fact that a lender doesn't really need to know what happens in four or five years — for most equipment types the lender is either in, or close to, an equity position in their collateral after two years or, in the case of soft collateral, the term was short enough that the balance is substantially paid down after two years.

The bad definition chosen should be the one that's most useful given the lender's circumstances and economics. A lender that really wants to steer clear of deals that are likely to go 60 also has to realize and accept the fact that the model that does this will be highly sensitive to the applicant's days past due at time of application, and therefore volatile from one date to another.

For example, an applicant that is 45 days past due on some account at the time they're scored will get a much lower score than they will on the following day if that account is paid and goes down to 15 days past due. Though this kind of score volatility is generally unpalatable to users, the model is simply responding to the best information available at the time.

For most lenders, however, a 90-day bad definition is more appropriate, as models built with this definition are much less sensitive to light delinquency, and as 60-day accounts are still captured when they become 90, but are not if they cure before reaching 90.

#3. Develop an Overall Scoring Strategy

There are several choices that need to be made. Which score (or scores) should you use? Should you build custom models or buy generic pooled-data models? Should you combine them? What about the scores you have been using until now? The right answer for one transaction type or size segment may not be right for another.

The first question is probably to build or buy. Building a custom score has the advantage of really focusing on the specific type of business your institution does and, as such, it should be very predictive. But building has a number of disadvantages, too. One is cost — modeling firms charge anywhere from $50,000 to $250,000 or more depending on scope of the project, and the historical bureau data necessary to build the model can cost half that much.

While it may be possible to find someone willing to build it for less, one needs to be careful not to be pennywise and pound foolish — a lot is riding on the score. No one goes to a "discount" brain surgeon.

Another disadvantage of building a custom model is that it requires that the lender have a large portfolio with significant (and electronically available) historical data, which many lenders do not. Finally, since scores require periodic updating to maintain optimal performance, this investment will need to be made again and again every few years.

At the other extreme is buying a generic score. The disadvantage here is that the score is so generic that a lot of potential predictive lift is lost, as nuances of your institution's lending market and borrower profile are washed away in a model primarily built to predict how someone will pay their phone bill.

That said, however, such models are still predictive. In some cases, where extensive work has already been done and it doesn't relate to your institution's primary expertise — most notably consumer credit — using a standard FICO score (or one of its siblings like an Automotive FICO or their new Next Gen scores) may be the best way to go for the consumer component of your overall credit score.

A middle ground is a semi-custom score, which I believe is the best of both worlds. It's a pooled-data score built to focus specifically on one particular type of lending, and it doesn't require any individual lender to make a major investment.

At PayNet for example, we have built, in conjunction with some of the top credit modelers, a score specifically for transportation equipment lending and another one for office equipment lending. Not only is this solution more cost effective, as all the major development costs are effectively spread out over many lenders, but it is also likely to be more predictive, since these models are built using a much larger pool of data as opposed to just one lender's own limited data.

In developing an overall scoring strategy, also keep in mind that different unrelated scores can be easily combined. Indeed, the more different they are, the better. So a commercial score can be combined with a consumer score, or a generic or semi-custom score can be combined with an existing home-grown manually-calculated score.

To combine scores, simply do a retro-analysis by calculating what your borrowers' scores would have been when they applied for credit and then calculate what percent of each score range went bad. When

combining two different scores, historical score values are simply calculated for both scores, and the bad rate is a matrix showing what the odds are when one score is low and the other high, and vice versa.

The bad rates in the cells of the matrix become the new blended score that can then be used for decision making. Another benefit of going through the retro process is confirming a score's applicability to your institution's style of lending and measuring the score's overall predictive power for you.

Custom scores blending both commercial and consumer data can also be created, and there are also generic scores that have been built with the consumer component already blended in. While such scores are effective, they are probably not as effective as the do-it-yourself type blending described above.

The most important point, though, is to take advantage of all the predictive information available. If a commercial credit applicant is offering personal credit information, don't use just a consumer score or just a commercial score, but rather use both, either by combining two scores or by using a blended score. While it may be tempting to use just one for cost reasons, analyze the cost of incremental losses and foregone approvals before cutting corners. The one exception to this is prescreening, or staging of data-pulls, whereby a less expensive data source is accessed first and if the credit information found is so negative that the application will be rejected no matter what information is on the other bureaus, then there is no need to spend money pulling additional bureaus for that application.

#4. Garbage In, Garbage Out

Good data is at the very heart of credit scoring. There must be a very large quantity of accurate, consistent, relevant and above all detailed data with which to build the scoring model in the first place, because the model is really "learning" the truths of the world from this data. Indeed, I know of no person who has sat down with the credit files of say, 10,000 bad accounts and really tried to analyzed why these went bad, and how they differed from 100,000 other accounts booked around the same time that didn't go bad. The model does all this for us. Compounding the need for large quantities of data ("bads" in particular) is the nature of the most powerful mathematical modeling tool, multivariate regression.

The other time that good data is critical is when it's accessed in real time to feed into the scoring model to produce a credit score and make an actual credit decision. No matter how sophisticated, advanced, thorough or brilliant a credit scoring model may be, its output can't be any better than the data that's used to drive the model. Imagine you have a daughter who is debating between marrying two young men and she asks your advice. No matter how good you are at judging personalities — and even if one has beady eyes and a shifty look — it is doubtful if you could pick as confidently and accurately as you would if you found out one of them had been an axe murderer recently set free on a legal technicality. By its very nature, fact-based decision making requires having the facts. And as a rule, the more targeted and specific the data is to the issue at hand, the more value it will have.

#5. Like Predicts Like

If you were an auto insurance underwriter trying to predict if a driver was likely to have an accident, wouldn't you want to know how many accidents the driver had had before? Sure, it would be interesting to know if the driver had been bankrupt, as there is a correlation, but if you're trying to predict "X" happening in the future, knowing whether "X" happened in the past is almost always the single most important piece of information. While it may seem obvious, this principle is often ignored in the leasing industry. Many models meant to predict lease/loan repayment use as their primary input trade credit repayment behavior, without even considering the term credit repayment behavior. Building a very predictive model requires using very predictive data.

#6. Building a Scoring Model is Both Art & Science

The Golden Rule of scoring is that a model be empirically derived and statistically sound (EDSS). Mathematics is the heart of scoring, but it is not the soul. Think of the math as the ancient Greek Oracle of Delphi — it will correctly answer any and all questions you ask of it, but you must pick what questions to ask, and that is a real art. So while many models may be mathematically correct and EDSS, some models will be much more predictive than others, depending on the expertise and creativity of their developers.

Expert systems are similar to EDSS credit scores in that they are quantitative systems that assign points for various characteristics and a decision is made based on the point total. However, expert systems are different in that they are not empirically validated; rather, the objective of the system is to produce the same decision that an "expert" would if the expert were making the decision, right or wrong.

Expert systems are OK, but they are much more powerful — and believable — if they can be statistically validated. For example, traditional Chinese herbal medicine has cures for some ailments that Western medicine does not. Yet without having these cures scientifically validated, it's hard to tell the difference between those that are just superstition and those that really work. Similarly with credit scoring, the most powerful models are built by statistically evaluating the factors that credit experts have learned to be the most important. Under the statistical microscope, some factors will prove to be more important than previously thought and vice versa, or just in certain types of cases, but the key is that existing expertise is usually the most fruitful hunting ground for predictive model variables.

However, there are challenges. The first is data — you can't check a truck applicant for "rapid expansion" without either financial statements or bureau data that shows how many trucks they've recently financed. Another is expertise — the best statistician isn't likely to be the most experienced leasing credit person, so making the most predictive models requires a partnership where mathematician and industry expert work closely together. Often a modeler can find ways to quantify abstract concepts that the credit expert cares about but can't envision putting into a model.

For example, every credit professional is going to be more positively predisposed, all else equal, to an application from American Global Systems than one from Willie Henry Enterprises, and when tested statistically, the credit person is right. As a result, we've included that insight, to the extent it's statistically supported, into our models. As long as a potential variable "makes sense" (never ignore common sense!) the only real limit is the imagination, and that when tested, the statistics support it.

While a variable based on applicant name won't have a large impact, every bit helps. There is no fixed maximum number of variables that a model can use. Indeed, models are more robust and stable to the extent that they look at a wide variety of factors, including factors that are closely related (with the one caveat that the number of variables needs to be small in relation to the number of transactions that the model is being built on, to prevent what mathematicians call "over-

fitting"). For example, while a borrower having a high portion of its transactions being 60 days past due is highly correlated to its having ever been 90 days past due, a model is usually better off for counting both factors and spreading the weight so that no one factor has too much impact by itself on the total score. This way, in those less frequent (but still common) cases where a borrower is bad on one measure but not the other, the score for the borrower won't be impacted all one way or the other, depending on which factor was chosen for the model. Although this often won't perceptibly improve a model's statistical lift, it will make it a better model.

There is literally an infinite number of possible variables and derived variables a model can use: square root of days past due, percentage of payments that were more than 60 days past due, change in the volatility of delinquency, delinquency relative to the norm for a given SIC Code, borrower state per capital income, debt per employee ... clearly the list is endless. No computer system can dream them up, so the only way they'll get tested to see if they work is if someone comes up with the idea that maybe "X" is a predictive clue to future bad performance. So although a mathematician who doesn't know the industry being modeled can still create an "OK" model, creating a really good model requires the combination of industry expertise and modeling expertise. The best models are usually built by teams.

#7. Slice and Dice

Considering how variables will affect different segments of an applicant population is important for other reasons too. Transportation equipment borrowers are different than medical equipment borrowers, who are different than office equipment borrowers. One size can fit all, but not nearly as well as a more focused custom product will. While historically a lack of sufficient data to slice into segments and/or a lack of willingness to invest in building multiple different scorecards or models has led to a "one-size-fits-all approach," neither of those constraints exists today. And segmentation is by no means limited to equipment types. It can mean new borrowers vs. old borrowers, or small borrowers vs. big borrowers, or borrowers in certain SIC codes (e.g. for hire truckers) vs. other SIC codes (e.g. private fleets) — the list of possible segmentations is long and largely dictated by the market segment being modeled. An experienced modeler will know the most common segmentations, but this is another area where having an experienced credit professional working with the modeler will produce the best results by far.

Another way of thinking of this, from the mathematical side, is that while the mathematics, generally multivariate regression, is brilliant at simultaneously evaluating tens of thousands of transactions and coming up with optimal weights for each variable, it doesn't have an ability to do segmentation on its own — it won't give back an answer saying "variable X is very predictive for all the borrowers located in Eastern states, but not for borrowers in Western states" — unless the modeler asks that question by doing (testing) that segmentation. In this example, without the segmentation, the mathematics would simply say, "This variable is pretty predictive, but not very predictive" because the majority of applicants are located in Eastern states — and the potential predictive lift would be lost. While such a geographic distinction may at first sound silly, an experienced construction equipment lender would know to look at a North vs. South segmentation because only the Southerners can work in year-round weather conditions.

#8. Beware the "Tyranny of the Majority"

Scoring works on probabilities — "What will work the greatest percent of the time?" But the modeler must also strive to prevent the "tyranny of the majority," or situations in which a model variable works for the large majority of cases, but where it really causes trouble for a significant minority. A classic example is a model variable that is the sum of days past due now, for all of a borrower's accounts. Overall, it's a very predictive variable, but there's a problem: This variable, as constructed, is subtly biased against large borrowers. For example, a borrower with 50 accounts that are on average five days past due is going to have a total of 250 days past due, which in general (and therefore in the model) is a very bad thing. It's not hard to solve this problem, but it must be recognized in order to be solved. An easy solution here would be to redefine the variable as the sum, for each of a borrower's accounts, of the number of days now past due minus ten (or some such number) which prevents a large number of insignificant delinquencies from adding up to large delinquency number.

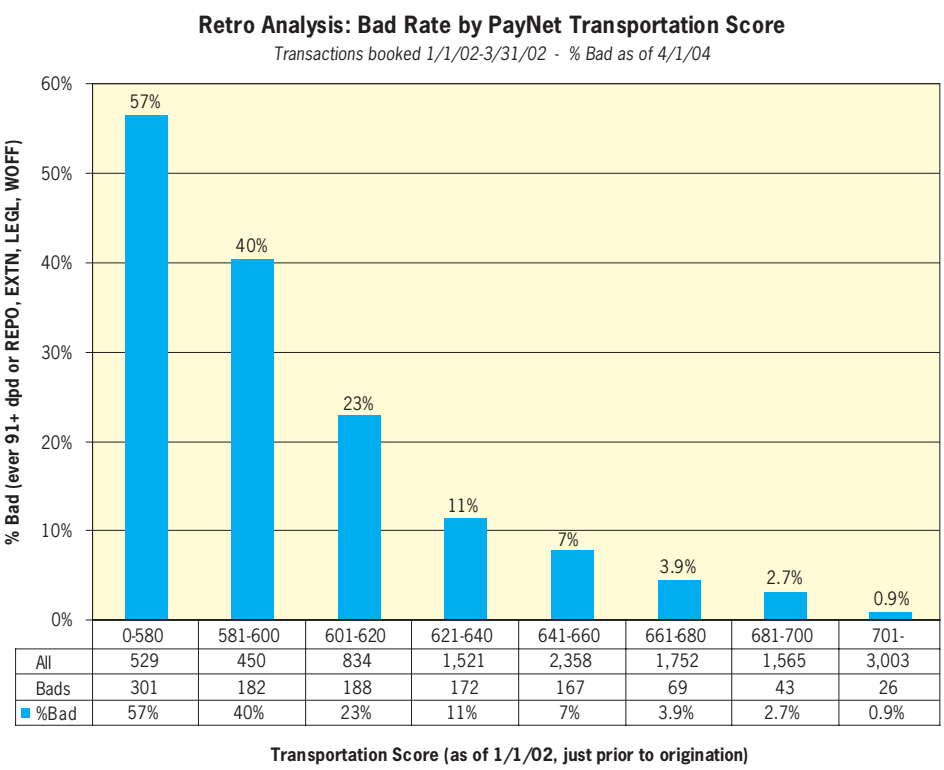
The key point though, is that while models do work overall and on average better than other decisioning methods, the modeler should not stop there. It is possible to create an even better model by making sure that the model is handling even the less common situations ("outliers") as well as possible. This can be done in two ways. The first is "top down" statistically, by performing segmentation analysis (guided by the question, "Are there some borrowers or situations where this variable could create a biased result?"). The second is "bottom up" by looking at how the model scores a wide range of individual deals and making sure that there aren't cases where the model is doing something that doesn't make sense.

#9. Test the Score & Decide How to Use It

Whether building or buying, at some stage it comes time to evaluate the score you are planning to use. Does the score really work? And up to what dollar amount? For all my business segments? Do I trust it enough to do auto approvals and/or auto declines? What should the score cut-offs be? Do I want to reduce losses or increase approvals, or both? How great are the benefits we expect to see? Are we really confident that all the necessary testing has been done and that it's time to start using the score?

While all these questions need to be answered before using the score, the fundamental analyses on which the answers are based are not difficult to perform and the results are quite accurate. Indeed, there is a strong argument to begin here — test a score that looks promising and see what it can do for you. The benefits of scoring are so great that it is really a strategic mistake not to at least do the tests.

There are two main analyses. The first is the retro analysis, calculating what the score would have been on deals your institution booked in the past and determining what the eventual bad rates were for different scores. This analysis can be done for different lines of business, for different borrower exposure amounts, for new vs. repeat customers, etc. This tells you what to expect from the score. The graph on the upper right is a typical, actual example. In this case, of the 3,003 deals that scored 701 or higher, only 26 went bad. At the other extreme, of the 529 deals that scored 580 or lower, 301 went bad. (*See Retro Analysis: Bad Rate by PayNet Transportation Score*)



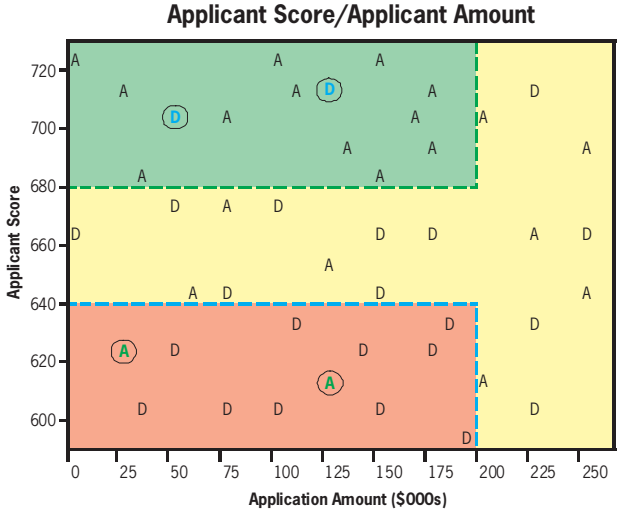
In the hypothetical example above, the lender's current practice for applications less than $200,000 is generally to decline those scoring below 640 and approve those scoring over 680. There are two exceptions in each direction to this general rule, and a circle is drawn around them in the diagram below. In cases where I've actually done analyses like these, upon careful review, the low-scoring approvals almost always turn out to be mistakes, and the person who approved them wishes they hadn't. That said, this is also an opportunity to verify that these aren't part of some special segment of business where approving lower scoring credits might be OK. (Are these deals full vendor recourse? Are these borrowers doctors with seven-figure incomes?) The high-scoring declines, on the other hand, are usually deals that have something wrong with them that's not credit-related per se, such as ineligible equipment types, environmental issues, geography or some inherent structural problem.

In the example below, the clear, general pattern of approving above 680 and declining under 640 holds true for transactions sizes up to about $200,000, at which point the pattern becomes less clear. So everything equal, with findings such as these, establishing a policy for applications under $200,000 of approving over 680 and declining under 640 would be a reasonable way to start. Over time, the gray-area, manual review band of 640 to 680 in this example can probably be narrowed, and the dollar amount up to which scores are used increased. Moreover, I've always thought it strange that an applicant who would automatically be declined for a small dollar amount might be approved for a much larger dollar amount. Except in unusual circumstances, a passing credit score should be a requirement for large transactions, in addition to whatever other requirements might be appropriate for the amount.

The second key analysis is comparing current credit decisioning practice to the scores given by the model. Given the retro analysis example above, one would hope that the credit analysts are approving all the deals that score above 680 (since their bad rate is 0.9% to 2.7%) and that they're declining all the deals that score 600 or below (since their bad rate is 40% to 57%). Without benefit of having the score, however, it is virtually certain that the analysts are approving some very low-scoring deals and declining some very high-scoring deals.

Using this differential, current credit decisioning practice versus how one would decision deals if the score were available, it is easy to calculate a "swap set" — the deals that will now be approved that would have been declined, that are swapped for deals that would have been approved that will now be declined. The swap set generally produces two major types of benefits. First is a reduction in credit losses. There is no reason why credit losses "have" to be what they are, and avoiding them has a direct impact on the bottom line. Second is an increase in approvals. Feedback on the credit granting process is usually asymmetric — one clearly sees the deals that were approved that shouldn't have been, but few people see the good deals that were declined, and their number can be large. Depending on institutional objectives, score cut-offs can be set so that all the benefit is shifted either to reducing losses or increasing approvals, but most institutions prefer to split the benefit more evenly.

When actually implementing a score for the first time, it makes sense to go slowly and examine the swap set deals carefully. Moreover, while one can and should primarily use the bad rates found in the retro analysis to set score cut-offs, it is also useful to see where the differences are by score and transaction amount. Identifying where the score's recommendations deviate from current practice helps in determining the cut-offs and the maximum transaction amount for auto-decisioning. This can be done by mapping out recent applications on a graph with score on the Y-axis and transaction amount on the X-axis, and showing each application as either an "A" for approval or "D" for decline, as shown in the chart to the right.



vendors who demand speed in exchange for deal flow should be willing to auto-approve more. A lender whose primary objective is to minimize losses may auto-decline low-scoring credits but then manually review anything before it gets approved. A reasonable way for a lender to get started is operating on a dry-run basis, using the score as more of a review rule, or simply as another factor for analysts to take into consideration.

Going from "review rule" use of a score to automated decisioning requires additional credit policy that limits the circumstances under which an automatic credit decision will be generated. Besides setting the cut-off scores for automatic approvals and/or declines, one should also put in place data sufficiency requirements. So for example, a borrower with a high credit of $3,000 or just six months of history probably shouldn't be automatically approved for $100,000 for five years. One thing that scores today don't do is measure "capacity" since it would require infinitely more data to build a multivariate regression-based score that said a borrower was safe for $X but unsafe for $Y. Instead, most lenders use the "comparable credit" concept and limit auto-approvals to some fraction or multiple of the borrower's previous high credit.

Similarly, many lenders require at least a couple years of history before granting an auto-approval unless a study has been done validating the score's predictive powers specifically for new businesses. In practice, however, this is less of an issue than it seems because new businesses generally tend to get mediocre, mid-range scores that are below most auto-approval minimums. Finally, most lenders, at least at first, prefer to put "training wheels" on their auto-decisioning by requiring manual reviews for any applicant that has, for example, ever been bankrupt, or 90 days past due, or that has any other characteristic broadly deemed as undesirable. Manual review rules like these are fine and are usually moot because deals with such negative characteristics are quite unlikely to score high enough to be auto-approved. Moreover, if it turns out that the manual decision in these cases (or an identifiable segment of the cases) is the same as the auto-decision would have been without the review rule, then the review rule can be peeled back over time.

Although auto-decisioning has many benefits, it should not be assumed that a lender that is unwilling or unable to auto-decision cannot benefit from scoring. To the contrary, such lenders can use a score's recommendation to help guide an analyst's decision, and such collaboration can even be assured by adopting credit authorities that require an analyst to get a second signature to approve a low-scoring deal or decline a high-scoring one.

It is important, however, that everyone involved realize that if there is no change in the institution's decisioning practices, no swap set, that the main benefits of scoring won't be realized. Decision speed will be improved and processing costs should be lowered, but the biggest benefits, reduced losses and increased approvals won't be realized unless some credit decisions change.

#10. Implementation: Institutional Acceptance & "The Human Factor"

There are two reasons why broad staff acceptance of scoring is important. The first is the basic need to have everyone in the organization pulling in the same direction — if senior management is going one way while the frontline staff is going the other, the results are never pretty.

Some employees are concerned that credit scoring will cost them their jobs. The good news here is that I have never heard of this actually happening. Because the adoption of scoring is a gradual process, it will mean that there will be less future hiring, even if there is substantial volume growth (which quicker credit decisioning can create). Total credit headcount may gradually decline through natural attrition or transfers to other departments, but employees really shouldn't be concerned about layoffs. Rather, scoring frees them up to focus on the more challenging borderline deals and on larger deals.

Other employees can be defensive, taking it as a matter of honor that they are "better" at adjudicating credit than the score, and go out of their way to try to prove it. This is most commonly an issue when presenting retro analyses that by their nature highlight that a large portion of the low-scoring deals went bad. The key here is for people not to take it personally — an automobile goes faster than even the fastest Olympic runner, and no one has a problem with that. It's also true that there are also many areas where human expertise beats machines, particularly working with large complex credits.

The other reason why staff acceptance is important is less obvious, and that is that there is a really opportunity for synergy between man and machine here: If they work well together, they will produce an even better result. The scoring models I've built allow users to "look inside" to see why the model is saying this applicant is good or bad, because then the credit analyst can better evaluate the model's recommendation on a particular application. Maybe the analyst will look at the key factors sited and say "Hmm, those are good points; I hadn't focused on those" — or maybe the analyst will say "Oh, that's why the score is the way it is; I happen to have information that the model doesn't have, so I know for a fact that this issue isn't a problem, and should therefore discount the score on this application." And this is actually a key point — in general when analysts look at the same information that the scoring model has, and reach a different conclusion than the model, they are usually wrong. But when the analyst has material "exogenous" information that the model doesn't have then the analyst's decision is more likely to be correct.

Finally, and most basically, credit analysts must understand the meaning of score values themselves. They need to know whether a score is predicting the probability of loss or of default (and if the later, how is it defined). Most "Empirically Derived Statistically Sound" scores predict default, while many Expert Systems type scores predict loss; and some scores combine the two. The analyst also needs to know whether the score they see is presented on an absolute basis or on a relative basis (akin to "grading on a curve"). Scores that are calibrated and presented on an absolute basis have the advantage of consistency over time — a score of X today means the same thing as a score of X a year ago — and what bad rate has historically been associated with that score. Many such scores are further calibrated to make comparisons simpler by using a standard rule such as "20 points doubles the odds" meaning that if the good-to-bad odds are 10 to 1 at a score of 650, that a score of 670 means odds of 20 to 1.

The other common way that scores are calibrated and presented is on a relative basis, and this is usually done on a percentile basis, on scale of 1 to 100. The advantage of this method is that the analyst can easily compare how a particular applicant compares to other companies. It has the disadvantage, however, that a score of X today does not mean the same thing as a score of X a year ago. In my opinion, the ideal way to present scores is both ways, showing the user both an absolute score and a relative score.

The bottom line is that widespread score education and understanding are critical to getting maximum meaning and benefit from scoring. Not only does this maximize the potential for real "man-

machine" synergy, producing a result better than either could do alone, but it also minimizes the risk of misunderstandings and sub-optimal decisions being made based on misconceptions.

#11. Score Management is an Ongoing Process, Not a One-Time Event

Time is an important dimension to credit scoring in a variety of ways. No matter how much analysis and education is done upfront, real trust still takes time to grow. Typically a lender will start using a score without any auto-decisioning, and then move to auto-decisioning just a small portion of their deals. In many ways this first step is the toughest, and the portion of deals auto-decisioned may be only 5%. But this should be thought of as a Normandy Beach, a real accomplishment that though initially small in its absolute magnitude, laid the foundation for much more widespread success thereafter. Once auto-decisioning is actually started it tends to grow rapidly, as people see additional classes of applications that clearly can be auto-decisioned (e.g., good but not excellent credits, somewhat larger deals, other market segments, etc.) Before long 15%, then 30%, then 50% will be auto-decisioned. And the percentage will continue to grow, though more slowly, as the remaining deals are tougher to auto-decision (and the toughest to gain consensus on for auto-decisioning).

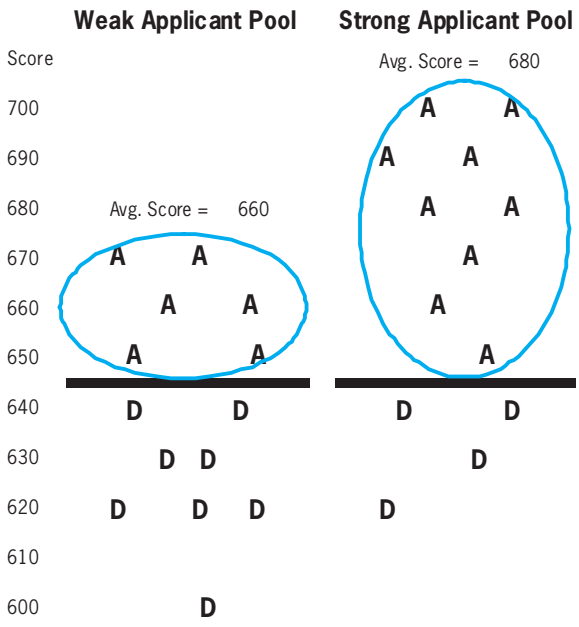
Growing the auto-decisioning percentage over time is an important activity, but just as important is monitoring score performance over time. Is the score performing as expecting? Do high-scoring deals have low bad rates and low-scoring deals have high bad rates? And are the bad rates for each score category what they were expected to be? The analysis done prior to adopting the score is important, insightful and useful, but it's never 100% the same as actually using the score.

First is the problem of "Reject Inference" — how does one really know what the performance of deals that weren't approved and booked would really have been? PayNet has done some interesting research in this area, because we have the data from essentially all the major lenders in many equipment categories, we've been able to calculate what the actual bad rates were on deals declined by one lender, but then approved by another lender — and the results have been very reassuring, as they were consistent with the bad rates predicted by the score. But even this approach isn't quite absolute "proof" in that not all declines are subsequently approved somewhere else, and there is presumably some bias to which applicants are eventually approved vs. not.

Theoretical nuances aside, the real issue to be concerned with is change in the underlying lending business. At the extreme, for example, a lender who used to do only direct business that now does only broker business cannot expect to have the same bad rates for a given score that it saw in the retro analysis. Similarly, a captive lender going into non-captive lending, or indeed any change in the positive or negative selection tendencies of the applicant population coming through the door will affect performance. A lender now charging high interest rates and advertising "E-Z Credit" will have higher bad rates for a given absolute score than one with low rates and tight credit.

Scoring also enables credit management to quantify the quality of applications being submitted, on an absolute basis over time, as well as on a relative basis. For example vendors can be evaluated by the average credit score of deals they submit, and by the average score of deals they book — and if the average score of the deals a vendor books is significantly less than the average score of deals approved for that vendor, then it is quite likely that the vendor is negatively selecting, i.e., only booking with you the deals that no one else approved. And knowing the average credit score, a true profitability measure for each vendor (or salesperson) can now be calculated, and the unprofitable vendors managed, given higher rates, ultimatums or cut-off. Similarly, management can now confidently evaluate securitizations, syndications and portfolio acquisitions, to determine whether there is any bias toward higher or lower credit quality selection, and react accordingly.

Overall portfolio monitoring by score is important because even if an appropriate minimum credit score is set, and the scoring model is working exactly as it is supposed to, the quality of the applicant population will affect the lender's overall average portfolio quality. In the simplistic example below, both lenders have set 650 as their cut-off, approving (and booking) everything over and declining everything under. Yet the lender with the stronger applicant pool has a much better portfolio, with an average score of 680 compared to an average of just 660 for the lender with the weak applicant pool.



Other macro analyses are possible and worthwhile as well. One bank, for example, looked at the distribution of credit scores that were coming from applicants who worked with loan officers in branches. What they found was that there was an abnormally high number of applicants that scored just high enough to be approved, and an abnormally low number of applicants that scored just under the cut-off. Upon investigation they found that the loan officers were "gaming" the system for marginal applicants, doing things like opening a checking account on the spot so the applicant could qualify as an "existing" bank customer. Armed with this information from monitoring, their credit policy was changed to prohibit rescoring (i.e., if the applicant did qualify with the data initially entered, then they couldn't qualify for an automated approval by changing the data).

Another area to monitor is decision overrides, and the reason for the overrides. Even if policy says that deals scoring below 600 should be declined, there will probably be some that get through on appeal. It is therefore very useful to develop a set of override codes for each type of override to distinguish between those that are essentially for sales considerations, vs. those that are based on important information not within the

scope of the scoring model (e.g., a start-up that just got $100 million of venture capital funding), vs. those where the risk of default is high, but for collateral reasons the risk of actual loss is very low. Using these codes does two things. First, it makes it clear how much business is being done on an exception basis. Second, it makes it possible in the future, to calculate what the actual bad (and loss) rates are on these deals. And in every case I know of, the score overrides done for sales considerations had very high bad rates, which credit management could then point to in their efforts to reduce the number of such approvals in the future.

Finally, and most importantly, all good things must someday come to an end, and scoring models are no exception. Models generally last two to five years, and knowing when it is time to retire a model — when it just isn't as predictive as it used to be — is one of the main purposes of monitoring. The good news is that the existing model doesn't need to be completely discarded. Rather, it becomes the base from which the next generation model is developed. If the business hasn't changed significantly, and if there are no new data sources offering the potential for increased lift, then creating the "new" model is really just a matter of updating and re-optimizing the old model based on newer data, and possibly looking for a few new variables to add additional lift.

#12. Never Lose Sight of the "Big Picture"

Building a high quality credit scoring model is a lot of work, and even if the decision has been made to buy rather than build, there is still significant work that needs to be done gaining internal acceptance, setting credit policies and parameters for scoring, and monitoring the results. It is therefore important not to lose sight of the big picture, why you're scoring in the first place. The list of benefits is so long it would be unbelievable, if each one of them weren't so clearly verifiable:

- More approvals
- Fewer losses
- Reduced overhead
- Much faster credit decisions (which improves the closing rate, and increases bookings)
- Greater customer satisfaction (at least of the customers you want)
- Increased management control
- Greater flexibility (e.g., "tightening" or "loosening" overall credit standards overnight)
- Improved consistency (so one analyst isn't declining a deal that another would approve)
- Better quality (i.e., fewer mistakes)
- Infinite scalability (if application volume doubles, you can't double the staff instantly)
- Improved operating information (e.g., quality of each vendor's applications vs. bookings)
- Improved overall transparency (of total portfolio credit quality)
- Improved predictability (of future portfolio performance, based on score vintages)
- Improved funding costs and access to funding (based on the above)

Indeed, the only "bad news" is that the benefits of scoring are so great that scoring is becoming non-optional. The innovators within a particular market will see a real benefit by adopting scoring, but the advantage doesn't last forever as more competitors adopt scoring. Over time the market prices and expectations change to the point where those that use scoring earn just a "normal" return, while those that still don't use scoring are operating at a costly disadvantage.

In general, the smaller the transaction and the larger and more homogeneous the applicant population, the easier it is to develop scores, and the sooner that lending market will adopt scoring. Consumers are easier to score than businesses, so consumer lending is about 25 years ahead of commercial lending. Though most consumers today have dozens of unsolicited credit card offers in the mail, 25 years ago getting a credit card required meeting with a loan officer at a bank, possibly discussing career prospects and the loyalty and responsibility demonstrated by having a savings account at the bank.

Today, such a process is unthinkable in consumer lending, and that is where commercial lending is heading. Virtually all deals under $25,000 are now scored, and increasingly deals in the $25,000 to $250,000 range are being scored as well. Within five years it is likely that the vast majority of transactions under $250,000 will be scored, just as home mortgages are. Several years ago the CEO of First Union said publicly that their ultimate objective was to score transactions up to $5 million. It will take some time to get there, but I'm sure that eventually we will. **m**

Thomas E. Ware *is PayNet's senior vice president of product development & marketing, and managing director of PayNet Analytical Services, which provides credit scoring, portfolio benchmarking, forecasting and other analytical services. In the 1980s he founded what became Golden Eagle Leasing. More recently, he has served as chief credit officer of American Express Equipment Finance and as general manager of a billion-dollar financial services business of Case/CNH Capital. He is a member of the ELA Credit & Collection Conference Planning Committee and ELA's Small Ticket Business Council. He graduated with distinction in Mathematical Economics from Dartmouth College and has an MBA from Harvard.*

Reprinted with permission from the *Monitor*, June, July/August & September 2005 by The Reprint Department, 800-259-0470, Part#9970-1005