Credit assessment as formalized vaticination

Akos Rona-Tas
University of California, San Diego

Introduction

Former credit bureau score sales person: The bigger picture on FICO's [Fair Isaac Co.] development as a company is that they popularized this notion of empirically deciding about making decisions. […], helping Avon figure out which people would be good Avon ladies, helping the IRS to figure out who to audit, helping
International analytics market developer:… universities to pick students…
Former credit bureau score sales person: Yes, but that failed, because they could not come up with a performance definition. All kinds of crazy projects, they are all really focused on filtering, so there were these great aspirations that people would use [it], to move beyond just banking, into insurance, into medicine, into other areas. But it just never happened. The idea is still out there…. When people ask me how it all got going, just these two guys had an idea about approaching data, when most people use computers to figure out the mechanics of an accounting system, they were off predicting people's behavior at about 1956, 58, 60. […] Then it turned out that once you have a device that can predict human behavior there are lots of people who have a need for them, bankers one thing, insurance companies, life insurance companies, utilities, oil companies, so the business developed as a way of telling industries interested in human behavior what people were likely to do in the future. So there are loads of opportunities once you define it that way. (1:36:37) [1]

The last fifty years have witnessed a revolutionary change in consumer lending thanks to a new technology called credit scoring that shifted credit assessment from expert judgment by loan officers to mechanized statistical prediction. This technology, which seeks to predict future behavior of loan applicants, was developed by a group of operations research experts associated with decision systems industry leader, Fair Isaac Co. By now it has been adopted by virtually all commercial banks in the

[1] Interview conducted by Martha Poon.

United States and it has been spreading rapidly all over the world. Credit scoring, however, is only a particular -- and particularly successful -- application of a more general technology of behavior prediction. Indeed, the core analytical problem faced by credit officers are not very different from the ones faced by airport screeners intent on identifying likely terrorists, psychiatrists trying to predict which of their patients will attempt suicide or college admissions officers hoping to select from a large pool of applicants students who will be most able to take advantage of the educational opportunity provided by the school. In each case, the prediction is not about aggregates but individuals. Airport screeners, psychiatrists and admissions officers, just as banks, must foresee behavior of particular persons and it is not enough to guess how many terrorist acts will happen in a year in some country, what the suicide rate of psychiatric patients is going to be or how admitted students will do overall . These individual predictions then are followed up by some decision about the particular person (and not just aviation security, suicide prevention or education policy in general).

Since Hume we know, prediction is hard, especially if it is about the future. It is harder still to predict what humans may or may not do as the vast philosophical literature on free will suggests.

Three aspects of technology

One way of thinking about technology is that it is a package of decisions taken for the users by its designers. Technology specifies a set of interrelated steps one needs to take to achieve some desired end. In a sense, technology can be thought of as a package of decisions taken out of the hands of the customer. What technologies promise us is that after the initial choice of the specific technology, to get optimal results, it is enough to follow the technological prescriptions. *After* the technology was deployed, its results may present users with hitherto unavailable options, but technology itself greatly reduces the role of human agency *during* the process of achieving those results.

Take one of the classic examples in the sociology of technology literature: the bicycle (Pinch and Bijker 1987). The bicycle can be seen as a bundle of choices. It could have steering wheels or handle bars or some other steering mechanism, but designers chose the handle bar for us, the users. Designers, of course, often do not know all the

possible alternatives and their choices are not necessarily optimal or well informed, but most argue for their contraptions as if they were. These choices are evident in the early days of the technology when they are contested, but once the artifact gets stabilized the human decisions become much less visible.

The package of decisions is justified in three ways: by its functionality, architectural coherence and autonomy. Functionality is the first and most important consideration. Here the claim is that the technology achieves its goals in some optimal fashion. The bicycle, for instance, delivers at a low cost a faster and more comfortable form of locomotion than walking. The engineering decisions embodied in the bicycle bring the best results. Architectural coherence, on the other hand, points to how each decision depends on others. For instance, the handlebar can be explained with reference to the construction of the frame that results in a certain body posture that makes a round steering mechanism collide with the knees. Here the claim is that even though each element may not directly promote functional optimality (it is easier to grab the wheel than the bar) in the context of the architecture of the technology, the other choices designers made, it is the best technical solution, i.e. replacing it with another part would harm functionality. Finally, a technology has to be easy to apply, otherwise it needs other technologies. There are relatively few things one needs to know to get on a bicycle and go, provided one has the requisite motor skills. But if one wants to take his two-year old child to preschool to the other end of town, one must deploy other technologies some of which has to do with child safety. If a technology is not autonomous and requires other technologies to work, it is incomplete and must expand to incorporate its complementary technological requisites. The strong claim for a technology is that it is *functionally optimal* (it achieves best results compared to available alternatives), *architecturally coherent* (its component decisions are a seamless whole), and it is *autonomous* (it functions in most contexts).

Even though it is embodied in various objects (software, computer hardware, instruction manuals etc.), credit scoring itself is not an object but a process. It is sold as "best practice," a technology that is optimal, coherent and autonomous. It is claimed to provide the best prediction of the applicant behavior and everything in the process to be

chosen to promote that. It is also alleged to be autonomous, to be applicable anywhere, anytime by just about anyone.

History of credit scoring

Credit scoring, as scientific vaticination, emerged from operations research, a field dedicated to turn science into social technologies (Mirowski 1999; Fortun and Schweber 1993). The first attempt at credit scoring dates back to 1941, when David Durand at the National Bureau of Economic Research devised a statistical method to investigate how good and bad loans differ on various characteristics using a chi-square test (Durand 1941). These statistical calculations, now taught at lower division college courses required a tremendous amount of time and effort in the pre-computer age. Understandably, banks showed very little interest in replacing the trained judgment of their loan officers with tedious calculations based on mathematical procedures most bankers at the time found hard to comprehend. When engineer Bill Fair and mathematician Earl Isaac, soon to be joined by other operations research people, began their company in 1956, they found it hard to convince lenders of the advantages of credit scoring. In 1958, when Fair and Isaac sent a proposal to explain the new technology to the 50 biggest lenders only one bothered to respond (Fair Isaac Company web site). Until the 1970s credit scoring was not an important revenue source of the company which is now the industry leader in predictive analytics.

In the 1960s, not banks but large retail chains began to adopt credit scoring. After seeing the success of the computerization of their inventory and billing system that they implemented with help from operations research experts, Ward, R. H. Macy, Bloomingdale's and others introduced credit scoring as one way of exploiting their growing centralized electronic databases of customers. Banks were more reluctant to take up the new technology. In 1974, during the congressional hearings leading to the Equal Credit Opportunity Act (ECOA 1974), an industry representative explained that formalization would

> "….freeze credit granting criteria into established molds, to the detriment not only of the creditor but of the consumer as well. This would have the effect of

introducing one rigid structure in the credit granting process, i.e., immobilizing criteria so that the creditor's option of employing its own funds to extend credit to an applicant could almost be viewed as mandated rather than voluntary on the creditor's part." (Cited in Taylor 1979 p.29.)

In other words, the technology would strip banks of the discretion to lend freely. While foregoing choice -- and thus being released from accountability externally and gaining transparency internally-- ultimately turned out to be a net asset, as we will see later, the technology has been sold primarily not on its strength that it immunizes banks against discrimination law suits and the incompetence or corrupt behavior of loan clerks. It has been promoted not even as a quicker and cheaper way of assessing whether applicants are likely to honor their obligations in the future. The principal claim for the technology was that it predicted borrowers' behavior with greater accuracy.

Credit scoring began to spread quickly after 1974 and became the technology of choice in credit card lending. Credit cards, in existence from 1958, have had only a modest growth in the 1960s and early 1970s partly because the small loans cards extend are not profitable if the screening of card applicants is a long and expensive process. But credit scoring cut costs and made credit cards cheaper to issue which in turn made cards more affordable and thus attractive. The technology got a second boost in 1995, when Freddie Mac, the giant, federally chartered mortgage lender, informed its partner institutions about the advantages of credit scoring in mortgage lending. Since then, mortgages also require credit scoring. [2]

Credit scoring has been spreading fast all over the world. Fair Isaac Co. is now a large multinational corporation, present in over sixty countries. Credit card giants Visa, MasterCard, American Express expect their partner lenders to use the technology and the soon to be introduced regulations by the Bank of International Settlement on risk management for banks (Basel 2) will make it hard for banks all over the world to avoid

[2] While no collateral, general purpose credit card lending does require careful screening of applicants, it is not entirely clear why mortgage lending is improved by this technology. Because banks own the property until it is paid in full and can sell it if the borrower defaults, their main risk comes not from non-payment but from adverse changes in the real estate market.

credit scoring (Allen et al. 2004). Today this technology is used to prognosticate not just about the financial behavior of individual customers, but about the creditworthiness of companies and even entire countries.

The technology of credit scoring

Scoring is designed to separate "good" applicants from "bad" ones. This is achieved by a statistical (link) function that connects the values of the outcome variable with a set of predictors (Figure 1.) The link function is the gears and chain that transforms the input of the pedal, our current knowledge about the applicant, into the output of wheel rotation, our forecast of his future. The outcome variable is usually categorical; it takes a few discrete values (e.g., default/no default). The link function turns the discrete outcomes into a continuous probability distribution and calculates the best weight for each predictor by using some optimization method. These weights are such that once one adds all the predictors using these weights, the final sum or score, which estimates the person's place in the probability distribution of the outcome, is the closest to the observed outcome.

The link function thus connects information from the more distant past (predictors) with information from the more recent past (outcome) and it is calculated with data from earlier applicants. When a new applicant appears, the lender gathers the information about the predictors. Then uses the values of the predictors with the weights and calculates their weighted sum. This score predicts the probability of the future outcome for the new applicants. This prediction can be given as a probability value (a number between 0 and 1) or as a percentage or using some arbitrary scale. The Fair Isaac Co. (FICO) score, for instance, has a range of 300 to 850 with a median of 723. To make the decision, the lender must transform the continuous distribution of scores back into discrete categories (accept/reject). This is done by establishing a cutoff, a minimum score for acceptance.

The main component of the functionality of the technology is the accuracy of its predictions. In principle, more accurate predictions about human behavior should result

in better pricing of loans[3] and a better selection of clients. Both should result in increased profitability. [4]

A naïve way of evaluating the accuracy of credit scoring models would be to see how well credit scores correlate with the behavior of borrowers after they were granted the loan. Analyses show that scoring models typically sort correctly only between 60 to 80 percent of the cases but how well it sorts depends on the ratio of bad to good loans. This accuracy is far from stellar, but to analyze the behavior of loan recipients is not the proper test of the predictive ability of the technology because it answers the wrong question. These models address the question: what is the probability that someone is a good/bad client, given that he was *granted* credit? Credit scoring, on the other hand, is concerned with a different question: what is the likelihood that someone will be a good/bad customer given that he *applied* for credit? The reason why banks are unable to answer the second question is *selection bias*: we cannot tell how people the bank refused would have behaved had they been given the loan. To the extent to which those who got the loan are not a random sample of the applicants, -- and the point of screening is precisely to "bias" our sample in the direction of good clients, -- the inference we can draw from client to applicant will be flawed.

There have been various attempts to correct for selection bias (e.g., Greene 1998; Thomas et al. 2002:107-120) but there has been no real solution to this problem (see Hand and Henley 1993). One might think that if banks were willing to drop screening altogether for a while and grant loans to every applicant or a random sample of those, the selection bias could be eliminated. Apart from being expensive, this solution runs aground on the endogeneity of the quality of the applicant pool. The kind of people that apply will, to some extent, depend on what applicants know about the way loans are given out. The group of people who apply for loans without screening will be different from those who request credit knowing they will be scrutinized thus the lessons learnt

---

[3] Until recently, banks rarely adjusted loan prices to scores, but instead they had a loan product with a price and just made a binary decision to accept or reject a customer for that product.

[4] More precise predictions also give better protection against charges of discrimination.

from unscreened clients will not be applicable once the lender begins to filter loan requests.

If the accuracy of credit scoring cannot be established, it can still be promoted on its merits relative to other alternatives. If the accuracy of alternative technologies suffers from the same measurement problems, we can still ask, are users better off using one or the other technology. The relevant alternative here is human judgment, the technology used before the advent of credit scoring.

Human judgment vs. statistical decision making

The claim of functional superiority of statistical models over human judgment in lending was derived both from a more general literature in cognitive psychology and a few experiments comparing judgment to statistical prediction in the lending. In an article in Science, Dawes et al. summarizing 35 years of research in cognitive psychology concluded that

> "Research reviewed in this article indicates that a properly developed and applied actuarial [statistical] method is likely to help in diagnosing and predicting human behavior as well or better than the clinical [judgmental] method, even when the clinical judge has access to equal or greater amounts of information." (Dawes et al 1989:1673.)

Eleven years later a meta-analysis of 136 studies comparing clinical judgment with mechanical prediction found that

> "Superiority for mechanical-prediction techniques was consistent, regardless of the judgment task, type of judges, judges' experience, or the types of data being combined." (Grove et al. 2000:19.)[5]

[5] Of the 136 studies, up to one half showed models to outperform humans and up to 1 in 7 found humans doing better, depending on what difference one considers large enough as evidence. The rest showed the two roughly equal. Only 5 studies included in the analysis were economic related. In four, the model did better, but only in two was there more than a marginal difference.

Empirical research on the accuracy of credit scoring compared to human judgment is not as thorough as one might surmise from literature reviews (Johnson 1992; Liu 2001; Hand and Henley 1997; Rosenberg and Gleit 1994). Chandler and Coffman, whose article is often cited as evidence for the superior precision of scoring, admit:

> No studies comparing the ability of the two evaluation methods to predict creditworthiness have been reported. Such a study would require the use of two parallel systems under well-controlled experimental conditions." (Chandler and Coffman 1979:22.)

Most evidence is from research done by Fair Isaac Co., but there are a few empirical studies looking at instances of loan officers overriding decisions brought by the scoring model (Chandler and Coffman 1979; Main 1977; Nelson 1983) showing that when humans overrule models and accept applicants the model rejected, their decision tends to be a mistake.[6] Advocates of the technology argue that scoring is more accurate because human judgment is prone to various errors and it is severely limited in its capacity to process information. They also contend that human judgment is overly pessimistic because loan officers focus on the negatives too much as their bosses would scrutinize them on bad decisions but not on good ones. Equally important, they state, that scoring is objective and can be better monitored for the exclusion of certain criteria thought to be discriminating. Scoring is consistent across individual officers making their decision not just more "fair," and more defensible, but also allowing for the accumulation of experience across officers and the correction of mistakes. Moreover, scoring makes do with less data and therefore it is less intrusive. But the trump card of advocates is that scoring does exactly what humans do, except it does it better. If a human judge can articulate the reason for a decision, in principle, this reason can be included in the model.

Detractors of the technology who advocate the superiority of human judgment point out that humans judge individuals while models always judge categories. They also

---

[6] This is a less than convincing test. First, it ignores negative overrides, when the model is more lenient than humans. Second, the relevant comparison is not between the performance of those who were let in under the cut off and the average, but those who got loans with positive override and those who the model let in at the very bottom, just over the cut-off. Moreover, the cut-off score is an arbitrary score and it is set by the bank and not produced by the model itself.

allege that humans are more flexible than statistical techniques. Models use a rigid algorithm to extrapolate from past to future and thus they are ill equipped to take into account changing circumstances. Moreover, humans are better at judging unusual cases. Finally, human judgment results in decisions that are more comprehensible to clients. Human decisions can be explained in causal terms helping rejected applicants to mend their ways. Models work with correlations and complicated statistical manipulations some of which banks treat as proprietary – i.e. secret – information. But even if banks were to disclose everything about their technology, credit scoring often does not provide the causal narrative people need to understand what they need to do differently to get their application approved.[7]

Scoring advocates believe that critics are overestimating the powers of human cognition. They retort that statistical models can handle much finer information and a much larger combination of traits than humans can (Lewis 1992:12). They also point to the prejudice ridden judgments of humans and their inability to change their minds in the face of evidence contradicting their preconceptions.

The debate on relative accuracy, now mostly settled with surprisingly little direct evidence, was contentious because it impinged on two sets of interests. The first was the interest of the credit professionals themselves, whose knowledge was now devalued and replaced by software packages. The second was the interest of the various social groups, who felt they were unfairly treated by lenders most of whom have been using the new technology. These groups wanted to know if the technology is indeed a good predictor of loan default. The answer to this second question was that credit scoring did not decrease and in certain cases did increase the proportion of minorities among the loan recipients. This avoided the question posed while responding to the concern underlying the question.

Accuracy of what?

At this point, we have to ask the question, what functionality exactly is that credit scoring is supposed to deliver. The answer seems simple: credit scoring is supposed to make banks more profitable. Then the "good" customer should be the profitable one and

[7] The Equal Credit Opportunity Act requires banks to explain their decision in a manner that allows people to mend their ways.

the “bad” the one who is not. It would follow that credit scoring should define “good” and “bad” accordingly and predict the future profitability of the applicant. But this is not the case for the vast majority of the lenders.

The prediction aims at not how lucrative the applicant will be but whether or not the applicant meets his/her payment obligations even though paying promptly and being profitable are two different things altogether. In the US, for instance, credit card holders who pay off their debts at the end of each grace period – about two in five – are, in fact, free riders. They use their cards for free and are subsidized by the “revolvers,” those who keep a large balance. People who miss even their minimum payment can be charged all sorts of extra fees. But the bank makes money not just if one pays only the minimum payment. Even if the borrower failed to pay, the bank could still have earned a handsome profit on the loan if the default happened at the end of the loan period when all the interest and most of the principal was collected.

There are three reasons why banks use non-payment instead of profitability as their target to predict. The first reason is not economic but moral. Banks are acting on a moral conviction. Most bankers believe that not meeting one’s contractual obligations is simply wrong and deserves punishment. But it is not just the bankers’ moral conviction that matters here. Most of the public agrees that loans should be given on the basis of how conscientious borrowers are. If banks started to deny loans to people who pay their debt promptly on the grounds that they are not profitable, the public would be outraged and banks would face an intractable public relations disaster. Surprisingly to economists, by punishing moral delinquency banks are providing a public good (a set of norms about credit behavior that benefit all their competitors as well) and forgo their own private interest. Secondly, most banks are not equipped to evaluate each loan account for its profitability, so even if they wanted to bankers could not model profitability. And finally, a bank’s reputation is strongly influenced by the default rate of its loans. Even if the bank makes money, non-paying customers and a high default rate shed a bad light on the institution.

Non-prediction related benefits of credit scoring

If banks are predicting the wrong thing without being able to assess the value of what they are doing, why are they using the technology? The accuracy of prediction is only one aspect of the functionality of the new technology. There are other selling points: lower cost, faster speed, higher political and professional legitimacy and increased managerial control over personnel.

Being cheap was the second most important consideration after predictive accuracy. No one could argue with the claim that credit scoring was a cheaper way of peeking into the applicant's future than a thorough case-by-case investigation by loan clerks. Technology advocates then argued that the lower expenses of scoring passed on to the consumer make loans affordable for less privileged social groups, therefore, the technology by broadening the market, made lending less discriminatory. Lower cost was not an obvious argument, however, before the recent steep drop in information technology expenses. Moreover, even with affordable software, computers and communication systems, credit scoring requires a sizeable upfront investment. Nevertheless, once the investment is made, the marginal cost of processing each additional application is low.

Credit scoring is also faster. Cutting time is especially important for purchase loans, where the customer often will not wait for days or even hours to learn if his loan application is accepted which is why retail chains were quick to embrace scoring in the early days.

The "objective," "scientific" technology also bestows political-legal and professional legitimacy. Political and legal legitimacy pivots on ECOA and Regulation B of the Federal Reserve (Federal Reserve 1985) that made banks litigation-proof against discrimination if they use credit scoring in an "empirically derived, demonstrably and statistically sound" manner.[8] This is quite curious given that credit scoring is built on

---

[8] Empirically derived, demonstrably and statistically sound methods are defined as follows:
"A system that evaluates an applicant's creditworthiness mechanically, based on key attributes of the applicant and aspects of the transaction, and that determines, alone or in conjunction with an evaluation of additional information about the applicant, whether an applicant is deemed creditworthy. To qualify as an empirically derived, demonstrably and statistically sound, credit scoring system, the system must be:

discrimination. It is not just that, as technology advocates claim, it discriminates between good and bad applicants. It discriminates against good applicants who belong to groups where others are bad.[9] In fact, if the essence of discrimination is to reward or punish people not for their own merits but for belonging to some group or category, credit scoring can easily be seen as nothing less but a formalized, mathematical system of discrimination. The reason why credit scoring is perceived as non-discriminatory is not because somehow computers and statistical techniques are accepted to be impartial and just, but because Congress prohibited the use of certain predictors in the link function. These include race, color, religion, sex, marital status, national origin, receipt of public assistance, and the good faith exercise of consumer rights.[10] Age can be used, but to avoid discrimination against the elderly (and to please the American Association of Retired Persons) after turning 63 years old one ceases to age in scoring models. It is clear that the exemptions reflect the political lobbying power of various social groups. Groups

---

- Based on data that are derived from an empirical comparison of sample groups or the population of creditworthy and noncreditworthy applicants who applied for credit within a reasonable preceding period of time
- Developed for the purpose of evaluating the creditworthiness of applicants with respect to the legitimate business interests of the creditor utilizing the system (including, but not limited to, minimizing bad debt losses and operating expenses in accordance with the creditor's business judgment)
- Developed and validated using accepted statistical principles and methodology
- Revalidated periodically by the use of appropriate statistical principles and methodology and adjusted as necessary to maintain predictive ability.

A creditor may use an empirically derived, demonstrably and statistically sound, credit scoring system obtained from another person or may obtain credit experience from which to develop such a system. Any such system must satisfy the criteria set forth above. If the creditor is unable during the development process to validate the system based on its own credit experience, the system must be validated when sufficient credit experience becomes available." (Federal Reserve 1985 202.2.(p))

[9] In 1979, at later congressional hearings on redlining, Bill Fair passionately argued that one should be able to use any variable as predictor in the link function, including race, ethnicity, gender or religion. Fair, taking a clean, positivistic stand, contended that anything that improves prediction in the model will result in greater accuracy which, in turn, will make loans more affordable. (Capon 1982:85).

[10] This last category was designed to protect people who file complaints against banks from retaliation.

that did not have political muscle at the crucial moment when the laws were debated, such as renters or those who live in areas with "bad" zip codes were not spared.[11]

Apart from legal and political legitimacy, credit scoring also transfers an air of professionalism on the lender. Using statistical models of prediction makes lenders look like they have some esoteric, specialized knowledge even if most are using the technology as a black box, the way we drive cars not knowing much about internal combustion engines or brake hydraulics. The technology also gives them cover when loans go bad. They can claim that they are using standard operating procedures which must be the best one available; otherwise they would have been replaced by better ones in the Darwinian-Panglossian world of technology.

All these benefits follow from the way this technology reduces choice. Its lower cost is partly the consequence of taking choices away from loan experts. Having made their task more routinized, skilled officers can be replaced with cheaper, unskilled data processors. No choices – no need for the skills to make good choices. The reduction of loan officer discretion also gave bank managers tighter control over lending. By cutting time for deliberation, managers also cut time to process applications.

Increased legitimacy also rests on foregoing choice. Credit scoring exempted banks from discrimination law suits, as long as they obeyed regulations by the Federal Reserve Board. By eliminating discretion, scoring eliminated culpability for rejecting deserving credit applicants. The same goes for professional legitimacy. Professional responsibility for bad decisions could be deflected by pointing finger at the generally accepted technology.

Coherent architecture

The technology of credit scoring is presented as a seamless set of technical decisions all working in concert to promote better, more accurate predictions. The claim is that even though the technical decision about specific parts may not seem to directly relate to the ultimate functionality of the technology, changing any of the parts would

[11] Zip code discrimination is now outlawed but it took two decades for that to happen.

compromise its ultimate goals. To allow that there are various alternative ways of going about credit scoring that they are consequential in an unpredictable manner or that some parts are chosen for the convenience of the designer and those have real unintended consequences for the functionality of the technology opens up the technology to criticism and raises questions about the designers' choices.

Choice of link functions

The heart of the technology's architecture is the link function. There are various link functions that are in use. The most common ones are discriminant analysis, logistic and linear regression, but financial institutions also use probit regression, neural networks modeling, genetic algorithm, as well as linear programming, recursive partitioning algorithm and nearest neighbor analysis. Table 1. compares the performance of six of these functions in five different studies measured as the percent of the cases properly predicted (Thomas 2000). The results should be compared within rows. With the caveat of selection bias and endogenous applicant pool quality, we can see that even in predicting loan recipient behavior there is no champion model.

My simulation on British data compares the three most often used models: linear regression, logistic regression and discriminant analysis (Tables 2a and 2b). What we find is discouraging. In the tournament of functions, on this particular dataset linear and logistic regressions do equally well, while discriminant analysis finishes a distant third.[12] Are the two regressions superior in this sample? Not necessarily. It depends on what you care more about. If you are more worried about undeserving people getting loans, discriminant analysis comes out winning. If you are bothered more by making the opposite mistake – refusing deserving people – the two regressions are the champions. [13]

---

[12] From a purely statistical point of view, the use of linear regression in predicting binary outcome is inappropriate as it violates the assumptions of the regression models.

[13] Throughout the analysis we sorted the predictions according to whether they were more likely to indicate good or bad behavior. This means that cases with probability greater than .5 of being bad would be sorted as such. If we use a different, more cautious threshold, say .1, which would land one in the bad category at much lower probabilities, (or to put it differently, one would need a .9 probability or higher to end up in the good group), then the vast majority of the cases would be predicted to be bad. While drastically reducing the error of deeming bad cases good, in this data set, where the majority of

Moreover, while in this sample from an aggregate perspective linear and logistic regression seem identical, from the applicant's point of view the two are different because there will be people who will be preferred by one model but not the other and vice versa. The correct predictions of these two models refer to different people even when the frequency of correct prediction is the same. If we were to use all three methods, only half of the cases would be classified correctly by all three functions. The other half would be misclassified by at least one method. For 38 percent of the cases whether or not they are misclassified will depend on the model used, the rest will be misclassified by all three models. And we cannot even trust those cases where all three functions agree. The three methods agree on only 63%, but a fifth of those are incorrect.

To sum up, 1. Models fit poorly and model fit is mainly driven -- in ways we don't understand -- by the variation in the sample not the link function used. 2. No link function is consistently superior to any other. 3. Some function will do better avoiding false negatives, others avoiding false positives in a particular data, but no model is overall better at either. 4. Agreement of multiple functions is no guarantee of correct prediction. The choice of models clearly matters, but we have no idea how and why. So any particular credit scoring technology starts with an arbitrary choice made for the user by the designer.

Model assumptions

Each statistical method makes various assumptions. These seem technical but many of them have substantive consequences. For instance, because we cannot directly observe probabilities and we can only infer them from discrete observations, most models must make assumptions about the shape of the probability distribution of the outcome. Differences between linear and logistic regression partly ride on such a difference. Or take that most functions assume additivity. Additivity means that we can simply add up the weighted predictors and the weight attached to a predictor will be the same for everyone. For instance, the weight of owning one's residence (as opposed to renting it) will be the same regardless of whether one lives in a village or in New York City. By the

---

clients are actually good, this makes the models perform even worse piling up erroneous predictions of the other kind.

same token, an applicant's income will carry the same weight when she is twenty and when she is fifty five. Finally, consider that most models assume that the cases in the analysis are independent of one another: one client going bad will not influence any other in doing the same. But an employer's financial trouble will influence his employees' ability to pay their loans; employer and employees are not independent. In recessions as well as during rapid economic expansion, people's ability to service their debts becomes even more closely linked. The closing a factory in a small town will drive the local shops out of business linking the financial trouble of the factory worker with that of the sales clerk. Moreover, people pay attention to how others behave and take their cues from others. If suddenly they see that many people around them for whatever reasons decided to go delinquent on their loans, they will feel less compelled to meet their own obligations even if they could pay. They will also soon figure out that above a certain threshold, lenders lose their ability to persecute offenders which diminishes the deterrence of possible sanctions. This is why loan default rates over a certain percentage soon doubles or triples.

Each of these three assumptions seems purely technical but each has real consequences. The shape of the probability distribution is an arbitrary choice guided by custom and computational convenience. Additivity is deployed for simplicity and often for the lack of enough cases to test all plausible violations. The independence of cases is assumed because there is no reliable information about the connections between the fortunes or behavior of clients and also because modeling this interdependence would make the task immensely complex.[14] The choice of the probability distribution, additivity and independence are all decisions made by the designer for the user who sees them as technicalities chosen from a wide inventory of substantively seemingly innocuous tools by experts to lead to the best prediction, when in reality they are driven by the designer's convenience and may have serious negative effects on the final performance of the model.

[14] In corporate lending this interdependence simply cannot be ignored. If a bank lends money to a bicycle manufacturer and a bicycle store chain their likelihood of repayment cannot plausibly be seen as independent. This makes modeling corporate loan portfolios very difficult.

Autonomy

Advocates of the technology claim that it is autonomous; it does not depend on other technologies. Its use requires only common sense and trivial procedures that users can easily figure out themselves. The application of credit scoring technology, however, is not that simple. First, and foremost, scoring needs standardized data. Without good data the operation is garbage-in-garbage-out. Users must decide what predictors to use and how to sort people properly on those variables. For choosing predictors, designers suggest to use anything that works. This is the data mining approach. One throws in everything into the model and let the model decide which variables help in the prediction and which ones are useless. Those with insignificant weights can be thrown out. To get you started they offer a set of often used variables with the hint that these will be by and large sufficient. But users are always encouraged to add other predictors if they think they might work. Yet even if one finds the right predictors, users must measure applicants along those characteristics. Before the technology can sort people into future goods and bads, the user must sort applicants into various categories of the predictor variables. While the technology addresses how to sort people by their future behavior, it is silent about how to sort people by their present or past condition. [15]

Measurement

Deciding what variables to use and how to categorize people are not that simple and it requires its own technology. What makes the technology of measurement complicated is that it must negotiate the technological choices of others which impose

[15] There has been a lot written about the power of classification (see Foucault 1973, 1979; Desrosieres 1998; Leyshon and Thrift 1999; Browker and Star 1999; Gandy 1993).

limitations on what can be measured and how. This is to say, measurement is social. It is not simply a cognitive process. Measurement must navigate around at least three limitations: *network externalities*, *complementary verification*, and *legal enforcement*.

Network externalities (Katz and Shapiro 1986) emerge from compatibility. Measuring variables others use and doing it in a common way makes measurement possible because it allows users to take advantage of information gathered by others. Using standard variables and classification systems also makes inquiries for respondents easier to interpret. Standardization facilitates the movement of information across actors. Using variables and coding schemes found in government statistics have the advantage not just that one gets access to useful aggregate statistical information but also it increases the likelihood that the inquiry on the application form will be intelligible for the applicant or her employer.

Complementary verification is necessary because measurement is useless unless the veracity of the information conveyed can be believed. How to verify information that enters the prediction model again calls for its own technology which relies on existing institutions and social networks. What is actually verified by the lenders varies enormously. Most lenders try to verify income with the help of employers and the tax authorities, many check on employment information, some ask for proof of assets. Some follow up addresses and phone numbers. Almost none requires proof of educational credentials, although it is one of the main predictors in many scoring models. Health related variables are almost never used as predictors, even in countries where this would be allowed simply because verification of health claims would be too costly or impossible.

Finally, certain characteristics (e.g., home ownership, criminal record, citizenship) are legally defined and enforced. Using these characteristics and sorting schemes have the advantage of having the legal system behind the measurement process.

What makes measurement difficult is that neither standardization, nor the ability to verify, nor the presence of legal enforcement developed with any consideration to the specific needs of credit scoring. These must be adapted to lending which calls for an elaborate technology of measurement.

Record keeping

The data gained through the measurement process then must be accumulated and stored. Banks first must solve the problems of their own internal record keeping. As banks are often organizations of enormous size with many branches scattered at great geographic distances, this can be a formidable task. What records to keep, for how long, what information should be easily accessible and which one should be stored on disks and tapes in vaults in some basement will all have consequences for the operation of credit scoring technology.

A special set of problems emerge, when banks together need to pool and store information. As past borrowing behavior is valuable information in predicting how people will deal with loan obligations in the future, and applicants' credit history is scattered across various banks, it is important that banks are able to create some common database. Because banks are competitors, information sharing about customers is not a simple matter. Information about clients is a valuable commodity that banks often acquire at a cost. Giving out information about bad customers have the advantage of punishing and deterring bad behavior as others will deny credit from that client. At the same time, lenders who paid for this knowledge are saving their competitors from similar expenses. Providing information about good borrowers, on the other hand, runs the risk that other banks will try to lure one's best customers away. What kind of information to share about clients and whether to share information at all is a difficult problem. In some countries there is no formal information sharing, in others, bank keep a black list of bad borrowers, in yet others, like the US, there is a full information credit registry that includes both good and bad credit information about borrowers (Major and Rona-Tas n.a.; Pagano and Jappelli 1993). If banks don't share any information at all, the possibilities of credit scoring are quite limited not just because crucial information will be missing from the models. The presence of a credit registry creates the threat of not being able to borrow from any other lender if one fails to meet one's financial obligations. This makes borrower's behavior more predictable. Black lists simply deter people from defaulting while full information registries give the extra incentive to build a good credit record, i.e., to borrow more and pay the loans as agreed.

Conclusion

The technology of predicting credit behavior has been a great success despite the fact that its functionality is hard to gauge, its internal architecture often follows computational convenience at the expense of functionality and its application depends on other, complex technologies that can have crucial effect on its performance. So why is it spreading so fast in lending?

The proliferation of credit scoring does not depend on its superior ability of vaticination. It is driven by its other advantages: that it is cheaper, faster than its alternative, expert judgment. It also gives more control for top managers over their subordinates and the lending process, in general, and provides legitimacy both legal and professional.

Why didn't behavior prediction succeed elsewhere to the same extent? A short comparison of selecting among those seeking credit from banks and academic credit from prestigious colleges in the U.S. is instructive. Admissions to elite schools did go through a similar formalization in the U.S. after World War II with the introduction of standardized tests (most importantly, the SAT and the ACT) and the replacement of personal interviews with the evaluation of application files. Just as in banking, pressures for formalization came from market expansion and pressures to avoid charges of discrimination.

With the growth of their markets, both schools and banks saw a surge in their applications. For banks, the post-World War II economic boom widened the market for consumer finance. For colleges, the GI Bill signaled the beginning of the tide of massification. With the large increase in the volume of applications, colleges just as banks had to find a cheaper and faster way of evaluating applicants. But the market grew not just in size but also in geographic reach. The elite universities on the East Coast started to abandon their local character and began to take applicants from a wider area well before the end of World War II. As leading universities became national institutions, and applicants were scattered all over the country, the personalized admissions process became unfeasible even before the number of applicants began to rise rapidly. Banks until the mid-1990s were by legislation restricted to operate in their states.

Private universities, and especially the elite ones, encountered no similar limitations. Banks were able to establish system of local branches that allowed them to avoid formalization for a while, although the branch network soon created the problem of coordinating and controlling what officers did far away. Thus propinquity to clients came at the expense of spatial proximity within the organization creating its own pressure on banks for formalization. Universities, on the other hand, were unable to establish and operate their own countrywide network of offices. It was left to the College Board and other administrators of standardized admissions tests to build a national system of college access.

The other pressure for formalization came from complaints of discrimination. The ECOA originally was the legislative response to women's groups who complained of discrimination in lending with other minority groups joining in with similar grievances. The formalization of university admissions was a response to discrimination against Jews at Ivy League universities. Elite colleges became concerned about the "Jewification" of what used to be the preserve of the WASP East Coast elite around in the 1910s and kept to a highly judgmental, holistic assessment of applicants and their character in their admissions process (Karabel 1984, 2005). The introduction of the SAT in the 1930s, promoted by Harvard's president James B. Conant, was partly a means to open up these august institutions to Jews and other "meritorious" minorities.

Then despite these similarities, why did colleges not go further in using formalized models in admissions? The initial quote suggests that the main reason was that its functionality was harder to assess, as colleges could not figure out what it is exactly that they want. It is true that undergraduate admissions officers in the US select students to higher learning in general and not to a specific discipline,[16] hence they work with a more complex set of goals than banks do. This would suggest that standardized scores play a greater role in admissions to graduate education than to undergraduate colleges as graduate education is more specialized. Indeed, for professional schools, we find that reliance on formal scores is more prevalent. But admissions to doctoral

[16] This contrasts with the practice of many other countries. In most European countries, for instance, one is admitted to a major or specialization.

programs, even though they are even more specialized than professional schools, tend to be less formulaic.

The key, I believe, is in the difference in the nature of the relationship between student and college vs. client and bank. The former is much more complex and embedded even in professional schools than the latter. Doctoral tutoring is the most personal of all. Professors develop a highly personal relationship to their students, while banks often never see their borrowers. While banks' adoption of credit scoring cannot be explained with the technology's success in predicting credit behavior, it can be explained with their *belief* that it can work. This belief is much weaker in colleges because of the more complex, more deeply embedded relationship between students and teachers.

References

Allen, L. G. DeLong, and A. Saunders. 2004. “Issues in the Credit Risk Modeling of Retail Markets.” *Journal of Banking and Finance*, 28:727-752

Bowker, G. C. and S. L. Star. 1999. *Sorting Things Out. Classification and Its Consequences.* The MIT Press

Boyle, M., J. N. Crook, R. Hamilton and L. C. Thomas. 1992. “Methods for Credit Scoring Applied to Slow Payers.” In *Credit Scoring and Credit Control,* ed. L. C. Thomas, J. N. Crook and D. B. Edelman. Oxford University Press

Capon, N. 1982. “Credit Scoring Systems: A Critical Analysis.” Journal of Marketing, 46/2:82-91.

Chandler, G. G. and J. Y. Coffman. 1979. “A Comparative Analysis of Empirical Vs. Judgmental Credit Evaluation.” *Journal of Retail Banking*, 1 no.2: 15-26.

Dawes, R. M., D. Faust and P. E. Mehl. 1989. “Clinical Versus Actuarial Judgment.” *Science* 243 no. 4899:1668-1674

Desai, V. S., D. G. Convay, J. N. Crook, and G. A. Overstreet. 1997. “Credit Scoring Models in the Credit Union Environment Using Neural Networks and Genetic Algorithms.” *IMA Journal of Mathematics Applied in Business and Industry*, no. 8:323-346.

Desrosières, A. 1998. *The Politics of Large Numbers. A History of Statistical Reasoning.* Translated by Camille Naish. Harvard University Press

Durand, D. 1941. “Risk Elements in Consumer Installment Financing.” National Bureau of Economic Research: New York.

Federal Reserve. 1985. Equal Credit Opportunity (Regulation B).

Fortun, M. and S. S. Schweber. 1993. “Scientists and the Legacy of World War II: The Case of Operations Research (OR).” *Social Studies of Science* 23, no.4:595-642.

Foucault, M. 1973. *The Order of Things. An Archeology of the Human Sciences.* Vintage.

Foucault, M. 1979. *Discipline and Punish. The Birth of the Prison*. Vintage.

Gandy, Oscar H. Jr. 1993. *The Panoptic Sort. A Political Economy of Personal Information.* Westview Press.

Greene, William. 1998. "Sample Selection in Credit-Scoring Models." *Japan and the World Economy* , no.10:299-316.

Grove, W. H., D. H. Zald, B. S. Lebow, B. E. Snitz and C. Nelson. 2000. "Clinical Versus Mechanical Prediction: A Meta-Analysis." *Psychological Assessment* 12, no.1:19-30.

Hand, D. J. and W. E. Henley. 1993. "Can Reject Inference Ever Work?" *IMA Journal of Mathematics Applied in Business and Industry*, no.5:45-55.

Hand, D. J. and W. E. Henley. 1997. "Statistical Classification Methods in Consumer Credit Scoring: A Review." *Journal of the Royal Statistical Society* 160, no.3:523-541.

Henley, W. E. 1995. "Statistical Aspects of Credit Scoring." Ph. D. Dissertation

Johnson, R.W. 1992. "Legal, Social, and Economic Issues in Implementing Scoring in the United States." In *Credit Scoring and Credit Control*, ed. L. C. Thomas, J. N. Crook and D. B. Edelman. Oxford University Press.

Karabel, J. 1984. "Status-Group Struggle, Organizational Interests, and the Limits of Institutional Autonomy: The Transformation of Harvard, Yale, and Princeton, 1918-1940." *Theory and Society* 13, no.1:1-40.

Karabel, J. 2005. *The Chosen. The Hidden History of Admission and Exclusion at Harvard, Yale, and Princeton.* Houghton Mifflin.

Katz, M. L. and C. Shapiro. 1986. "Technology Adoption in the Presence of Network Externalities." *The Journal of Political Economy* 94, no.4:822-41.

Lemann, N. 1999. *The Big Test: The Secret History of the American Meritocracy.* Farrar, Straus & Giroux

Lewis, E. M. 1992. *An Introduction to Credit Scoring.* Athena Press, Fair Isaac.

Leyshon, A. and N. Thrift. 1999. “Lists Come Alive: Electronic Systems of Knowledge and the Rise of Credit-Scoring in Retail Banking.” *Economy and Society*, no.28:434-466.

Liu, Y. 2001. “New Issues in Credit Scoring Application.” Institut fuer Wirtschaftsinformatik, Goettingen University, Working Paper #16.

Main, J 1977. “A New Way to Score with Lenders.” *Money*, February, 73-74

Major, I. and A. Rona-Tas. N.d.. “Why do banks share information about customers? A comparison of theoretical models for mature private credit markets and for markets in transition.” Manuscript.

Mirowski, P. 1999. “Cyborg Agonistes: Economics Meets Operations Research in Mid-Century.” *Social Studies of Science* 29, no.5:685-718.

Nelson, O.D. 1983. “Credit Scoring: Do AFSA Members Have Faith in Its Answers?” *Credit* 9, 14-16.

Pagano, M. and T. Jappelli. 1993. “Information Sharing in Credit Markets.” *Journal of Finance* 48, no.5:1693-1718.

Pinch, T. J.and W. E. Bijker. 1987 . “The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other.” In *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technolog*, ed. W. E. Bijker, T. P Hughes, and T. J. Pinch. The MIT Press.

Rosenberg, E. and A. Gleit. 1994. “Quantitative Methods in Credit Management: A Survey.” *Operations Research* 42, no.4:589-613.

Somerville, R.A. and R.J. Taffler. 1995. “Banker Judgement Versus Formal Forecasting Models: The Case of Country Risk Assessment.” Journal of Banking and Finance 19:281-297.

Srinivasan V. and Y. H. Kim. 1987a. “The Bierman-Hausman Credit Granting Model: A Note.” *Management Science*, no. 33:1361-1362.

Srinivasan V. and Y. H. Kim. 1987b. “Credit Granting: A Comparative Analysis of Classification Procedures.” Journal of Finance 42:665-683

Taylor, W. F. 1979. *Meeting the Equal Credit Opportunity Act’s Specificity Requiremens: Judgmental and Statistical Scoring*. Master of Law Thesis, University of Wisconsin, Madison

Thomas, L. C., J. N. Crook and D. B. Edelman. 2002. *Credit Scoring & Its Applications* (Siam Monographs on Mathematical Modeling and Computation).

Thomas, L. C. 2000. “A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers.” *International Journal of Forecasting*, no.16:149-172.

Yobas, M. B., J. N. Crook and P. Ross. 1997. “Credit Scoring Using Neural and Evolutionary Techniques.” Credit Research Centre, University of Edinburgh, Working Paper 97/2.

Figure 1.

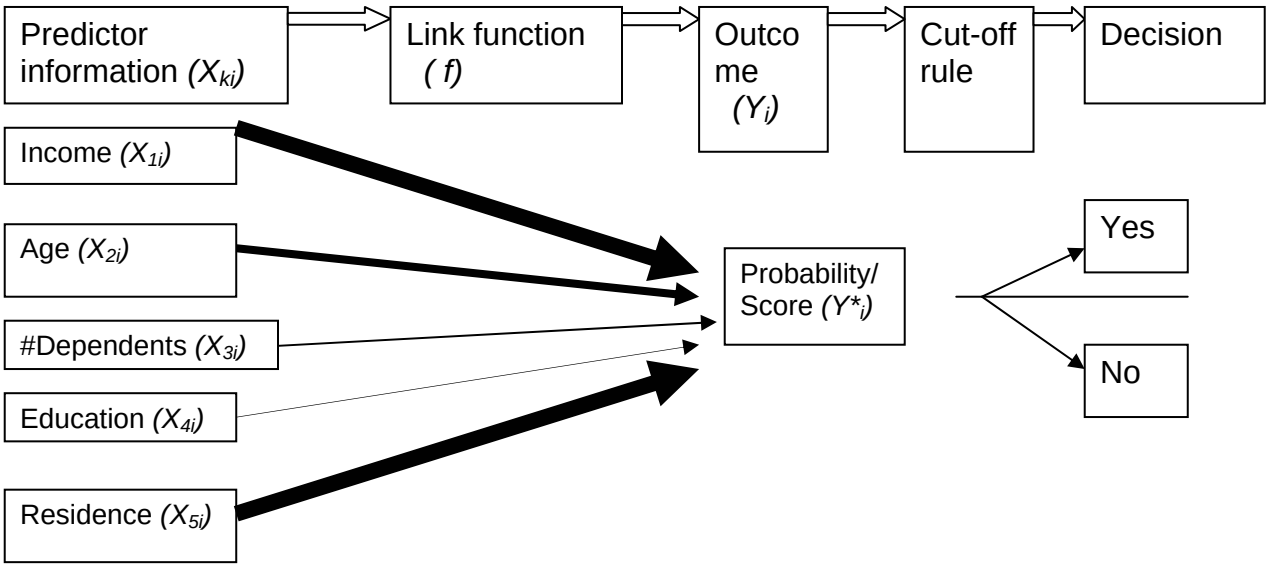
Predictor information (Xki)
Link function (f)
Outcome (Yi)
Cut-off rule
Decision
Income (X1i)
Age (X2i)
#Dependents (X3i)
Education (X4i)
Residence (X5i)
Probability/ Score (Y*i)
Yes
No

Table 1

Comparison of classification accuracy for different link functions

(Thomas 2000)

Authors	Linear regression	Logistic regression	Recursive Partitioning Algorithm	Linear programming	Neural networks	Genetic algorithm
Henley (1995)	43.4	43.3	43.8	-	-	-
Boyle et al. (1992)	77.5	-	75	74.7	-	-
Srinivasan and Kim (1987a,b)	87.5	89.3	93.2	86.1	-	-
Yobas et al. (1997)	68.4	-	62.3	-	62.0	64.5
Desai et al. (1997)	66.5	67.3	67.3	-	64.0	-

Table 2a

Comparing link functions

Linear Regression, Logistic Regression, Discriminant Analysis[17]

Link function	% correct overall	Predicted good when bad	Predicted bad when good
Linear Regression	74.4	299	14
Logistic Regression	74.5	289	23
Discriminant Analysis	63.3	135	314

[17] Using 11 predictors including income, several debt measures, age, family measures etc. Data is from the United Kingdom, N=1225 (Good=902, Bad=323).

Table 2b.

Comparing link functions

Linear Regression, Logistic Regression, Discriminant Analysis (continued)

Correct by all three methods:	
– Correct good:	588 (48%)
– Correct bad:	24 (2%)
– Error by at least one method:	613 (50%)
Error by one or two (but not all three) methods:	464 (38%)
Agreement by all three methods:	761 (62%)
of those correct	612 (79%)
of those incorrect	159 (21%)

Akos Rona-Tas is Associate Professor of Sociology at the University of California, San Diego. He is the author of the *Great Surprise of the Small Transformation: Demise of Communism and Rise of the Private Sector in Hungary* (The University of Michigan Press, 1997) and co-author with Alya Guseva of "Uncertainty, Risk and Trust: Russian and American Credit Card Markets Compared." (*American Sociological Review*, 2001)