

Not all scorecards are the same

_by Murray Bailey

Category: Credit Scoring (development)

Keywords: Basel II
 Bayesian Theory
 Master/Slave scorecards
 Neural Networks
 Gini coefficient
 Small samples

How can you build a better scorecard? There are many techniques and yet the majority of in-house developers use linear regression. It is undoubtedly the simplest, but could a new technology deliver benefits to the business?

Whilst there is a best practice for building scorecards (see CRI April 2003)¹ there are many different approaches. The modelling technique used to find the correlations between the variables is often multilinear regression, but alternative techniques are now coming to the fore. In this article, I first review some of the approaches, before talking to a few scorecard developers about the special techniques they use. After all, scorecard building is all about discrimination and the best developers have their own preferred ways of maximising the separation of good and bad customers.

Regression

The most common technique for building scorecards today is to use multilinear regression. The regression finds the line-of-best-fit to predict risk from the variables (characteristics) entered. The result is a series of coefficients that represent the correlation in the data. If one predicted risk without taking into account the overlap of information, the scorecard would effectively double count.

Multilinear regression has gained wide acceptance because it is simple to explain and to use. And yet it relies on principles that are violated. For the modelling to be accurate, the data should be normally distributed and continuous. It rarely is. Scorecards are typically built with a handful of discrete outcomes, such as under 25, 25 to 30, 31 to 40, over 40. Even the continuous variables, like time at address are grouped to leave a series of discrete bands.

The reason developers do this is sample size. The more records there are, the more the sample will represent the whole population. Having continuous variables can lead to very low 'cell counts' (numbers of records with a given attribute). This in turn will lead to spurious results.

Logistic regression provides a nice solution to our problem. It fits the data in a probabilistic way using discrete binary (1,0) outcomes. Figure 1 illustrates logistic and a linear regression lines between risk and the variable owner. This simple diagram shows how linear regression is a poor approximation when there are four opposing points that a line is drawn between.

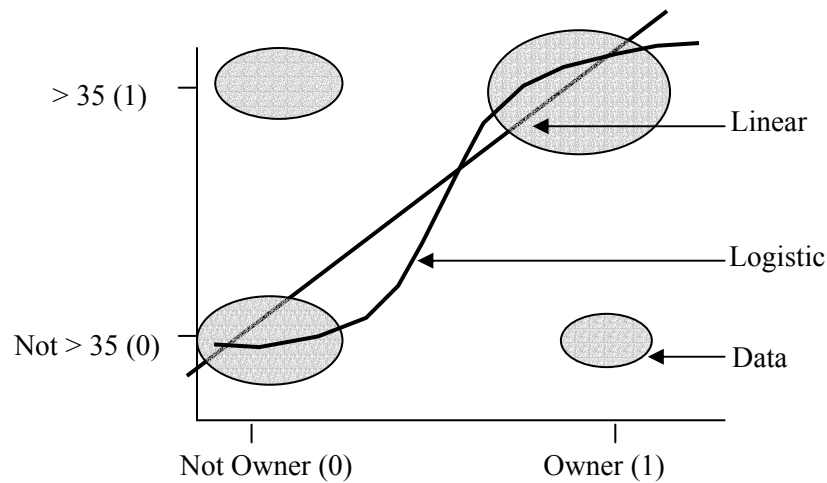


Figure 1: Linear and logistic regression comparison

So why do developers continue to use linear regression? Dave Poole Head of Marketing and Risk at store card lender IKANO Financial Services explains: “We have tested logistic regression and found that there was negligible difference. The logistic regression took longer and linear regression was easier to explain to the business.”

Developers must never forget that the business is more important than the model. There is no point in developing a statistically perfect solution if it is impractical. It must be implementable, understandable and provide business benefits. Statistics can sometimes show an improvement in the discrimination between goods and bads, but the benefit is not translated into practice. What matters to the business are acceptance rate, bad rate, processing speed, automation etc. - not the Gini coefficient.

The technique becomes an issue when available data is limited or the quality is questionable. Simon Trupp, Senior Consultant at consultancy PIC Solutions, South Africa agrees. He believes this is where new technologies like neural networks come into their own. “With limited data sets, it will be important to identify and use interactions in the data therefore neural nets may be more useful,” he says. “Typical data sets have large sets of data items and this removes the practical benefit of using interactions between variables as another variable may be correlated with the identified interaction between two variables.”

Neural networks

Neural networks are widely used for fraudulent transaction detection because of the large amount of available credit card transaction information. However, they are frequently criticised for being ‘black-box’. Historically, one major hurdle to neural networks was the difficulty of implementation. In the early 1990’s Household International built a neural network decision system for Canadian mortgages. It took days to build, but nine months to implement.

Today, vendors appreciate this issue and many provide proprietary software so that the solution is almost ‘plug and play’. They also focus on the attraction that neural networks should outperform regression-built scorecards.

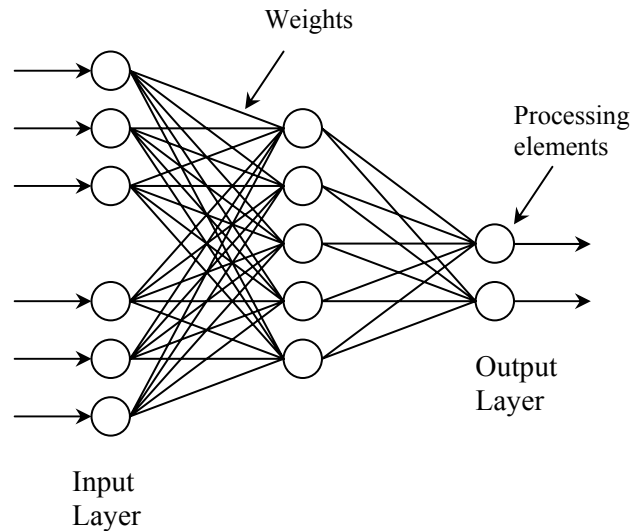


Figure 2: A multi-layered neural network

Figure 2 illustrates the most commonly applied neural network – the multi-layered perceptron (MLP). Neural net models are nested logistic regression equations. An MLP neural network has an input layer (the variables) and an output layer (risk). The difference between logistic regression and an MLP is that there is also a hidden layer. The benefit of the hidden layer is that it deals with interactions in the data – something that can be solved with regression scorecards only by building multiple scorecards.

“In principle, these should work at least as well as scorecards,” says Alan Lucas, Senior Manager in the Retail Credit Risk division of consultancy KPMG in the UK. “However, inexperienced analysts can over-fit these models. I have done tests to show that once the number of characteristics exceeds a certain number, neural nets provide no real extra power because the extra characteristics correlate with the interactions in the data.”

The United States (US) has an act that regulates the build and application of scorecards. Two of the regulations are that reasons should be provided for adverse credit decisions and that automated decision systems should be statistically based. Neural networks typically fall down in both these regards. Lois Brown VP of Marketing at software and modelling company Austin Logistics in the US explains the problem: “When using origination models, the law requires that adverse action reasons be provided to applicants who are declined. With neural networks it is very difficult to determine adverse action codes and they can be counter-intuitive which may create irate applicants.”

In France the central bank (Banque de France) requires copies of all scorecards in use. These are held by the bank in confidence. The European Union is considering adopting this practise Europe wide. Sharam Sharifi, head of the credit function for Lloyds TSB’s card services in the UK, says, “Sending a neural network to the European Central Bank (in text format) would be a nightmare. Most neural net packages do not allow it: they produce a program in C++, which you are supposed to plug straight into your system.”

Basel II also has implications for neural networks. It stipulates that all decision systems should be monitored. However, monitoring a neural network is complicated. Sharifi again: “The salesmen for neural net packages do not distribute software for monitoring the network: they just tell you that you can rebuild it every six months instead, with one click of the mouse! In fact, if neural networks are neither looked at nor monitored, they become vulnerable to internal fraud.”

Dr George Bolt, Research & Innovation Director at Neural Tehnologies, UK strongly defends neural networks against these criticisms. “When replacing a traditional scoring solution, substantial improvements in performance can often be achieved by taking into account the subtle non-linear interactions between fields,” he says.

“Any scorecard based on any technology that exploits these interactions (as neural networks do) can be complex enough to look like a black box. However, modern scorecard development software includes a range of visualisation tools that lift the lid on the black box so that the scorecard analyst can understand the relationships contained within it. For example, it is common for an analyst to be able to rank fields according to their influence on estimated scores, to see how changes to each field would affect a score, etc. Careful examination of such visualisations can provide the analyst with an understanding of how even the most sophisticated scorecards derive their estimates.”

Comparison of techniques

There have been many studies of the power of alternative models. Table 1 gives the results from a sample of unbiased studies comparing some of the more popular techniques. The figures are Gini coefficients, where 100 represents a perfect scorecard and 0 represents no discrimination.

As is reflected by developer’s experience, the Gini coefficients for linear regression are not significantly lower than for logistic regression. It is also interesting to note that no one modelling technique is universally better than any other.

However, there may be specific areas where one technique may have the edge, such as a dynamic environment (transaction fraud) or where the sample is small. It therefore appears that the modelling technique is less important than:

- Meeting the business requirements
- Providing a solution that fits the operational environment
- Obtaining quality data and understanding it
- Following best practise procedure during modelling
- Performing thorough analysis of the data
- Being cautious with reject inference and testing its validity
- Interpreting patterns in the data and applying experience and common sense

Scott Horwitz, Director of Market Management, Analytics, at Fair, Isaac, US points to the amount and completeness of the data and the relationships amongst the data elements. He says: “Most importantly, a model’s value is directly related to its ability to solve a particular business problem.”

Table 1 Comparison of model building techniques (expressed as Gini coefficients)

Authors	Linear Regression	Logistic Regression	Decision Trees	Linear Programming	Neural Networks	Genetic Algorithms
Henley ²	43.4	43.3	43.8	-	-	-
Boyle ³	77.5	-	75.0	74.7	-	-
Srinivsan ⁴	87.5	89.3	93.2	86.1	-	-
Yobas ⁵	68.4	-	62.3	-	62.0	64.5
Desai ^{6,7}	66.5	67.3	67.3	-	66.4	-
Oxley ⁸	68.6	68.3	-	-	69.5	-

Sample size

Fair Isaac, the founders of credit scoring, set the bench mark sample sizes as 1,500 goods, 1,500 bads and 1,500 rejects. The smaller the sample, the more prone to errors the model will be. The characteristics are constructed by combining attributes. The number of bad accounts tends to be what matters since bad rates tend to be small. With low numbers of bads, the developer risks combining attributes that reflects the statistical variation in the data.

There is no maximum sample size, but more than about 10,000 records and the errors start to become insignificant for regression approaches. One of the criticisms of neural networks is that they require very large samples. Bolt of Neural Technologies says this is nonsense. “Whilst the accuracy of neural scorecards tends to improve with the quantity of data that is used

to train them, high performance neural scorecards can be produced from as few as 50 to 100 records," he asserts.

"For such small sample sizes, advanced training algorithms are used that are based on Bayesian principles. These algorithms encapsulate the principle of Occam's razor – that the simplest scorecard that performs well on the sample data is most likely to perform well once deployed. Using sound theoretical principles, Bayesian neural networks avoid over-fitting by automatically adjusting their complexity to the amount of data available and identify and use only the most relevant input fields.

"The most advanced scorecard building software also allows the scorecard builder to incorporate prior knowledge into a scorecard, thereby reducing the amount that has to be learned from the sample, and improving the scorecard's robustness. For example, if it is known that the score an applicant receives should increase with their income, a neural network can be constrained so that this condition is enforced."

Mike Cutter, Portfolio Manager at GE Capital in Australia advises developers to "make sure the scorecards are very simple" when there is a small sample. He says this means having few characteristics and attributes. Sharifi, on the other hand questions whether a scorecard should be built at all. "Some industry experts recommend building the scorecard using Weights-Of-Evidence instead of dummy variables. However, research presented at the Edinburgh credit scoring conferences suggest that you are better off adopting a generic scorecard built for a similar portfolio."

Most scorecard vendors will say a small sample is where the number of bads is below 250 to 200. If the number of records is this low, the first question should be: do I really need a scorecard? What are you trying to predict?

I was once asked how a company could build a scorecard when they had next to no bad accounts. My question to the business was what do you want the scorecard to do? Is it to emulate the underwriters to increase automation; to increase acceptance rates; or sales of insurance etc.? They wanted to provide a sliding scale, reflective of the current underwriting that would let them test increased acceptance. In this instance, the solution was to build an accept/reject scorecard. This predicted probability of previously being rejected and enabled them to accept a group that were most like the accepted population.

It is tempting to relax your definition of bad to increase numbers, or to take bad accounts that are very new, whereas the good accounts are over a year old. If you do so be very careful. The definition of a bad account should be based on likelihood of loss. By relaxing it from say 3 in arrears ever, to 2 in arrears, will pick up customers who may be acceptable to the business. The resulting model will therefore predict the wrong thing.

Early bads can be reflective of poor underwriting or be fraudulent. Application scorecards assess stability rather than the quality of your underwriting policies or frauds. By comparing goods and bads from different timeframes, you also run the risk of comparing apples and pears. Check that the population is stable and marketing activity was consistent before considering early bads.

Validation against a hold-out sample is vital in scorecard development to ensure that the model has not been over fitted. Brown of Austin Logistics points out that validation is a concern with small samples. "If the number of bads is just enough to build the model, additional data for validation is not available," she says. "In this case, the solution may be to use cross-validation techniques or to compare the model against a baseline generic score. "

Special approaches

Not all developers are the same and most developers have their favoured techniques. However, when approached about their techniques, most scorecard vendors talk about their flexibility. For example Scott Horwitz, Director of Market Management, Analytics, at Fair, Isaac US says, "We approach individual problems with the techniques and methodologies that make the most sense for that problem. Our analysts study the business problem and then determine whether that problem is best served by neural net models, logistic regression models, models using a Fair, Isaac proprietary technique, or other models." I understand the 'proprietary technique' to be Iterative Search, an approach also used by Scorex and Windsor.

Horwitz goes on to say, "Often, the resulting models will be developed using a combination of techniques. The end goal is to develop the most powerful models that solve the business problems facing our clients."

Jake Betts, a consultant at Windsor, UK believes that small samples require the most attention. A technique he uses is to allow manual intervention in the scorecard assessment process. "This way, you can let an Operations Head assess the benefit of a variable or scorecard from one country applied to a portfolio in another," he says. "I also like to create variables using Bayesian logic. This can make a generic scorecard really work for a lender in a new situation."

A technique used by GE Capital's Cutter is to build master/slave scorecards where there is a small sub population. This may be mostly explained by the general (master) scorecard, but can be tuned by the slave scorecard.

He also says that, wherever possible, he implements scorecards as challengers. "This gives me the ability to measure the effect of the reject inference and other changes in the population. It enables me to prove the incremental impact of the scorecard thus avoiding debates about changes to policy rules, changes in the business mix or changes in the macro economy etc."

Simon Harben Director of Strategic Decisions at Experian, UK has a similar technique to the master/slave model approach. He calls this the "High risk niche approach". For a high-risk niche, a sub-population based on the master score is defined to identify the lower quality applicants, on which the niche model is based. He says: "Unlike a true niche, the high-risk niche is applied to the total population, resulting in a scorecard that has improved discrimination for lower quality applicants, particularly those around the cut off."

Arnold Koudijs, a consultant at KIQ in Holland, believes implementation can separate his company from the rest of the field. He says, "When implementing a scorecard we also use business rules that can be easily changed (without IT involvement). When we do this, the scorecard becomes a whole scoring module that incorporates the scorecard, accept/decline policies, product specific rates or limitations, and can generate the "reasons" for a decline." Explaining an adverse decision to a customer or prospect is required by the Fair Credit Reporting Act in the US, but Koudijs says it is also very useful for the user (and customer services) in understanding the decision.

Trupp of PIC Solutions agrees that implementation is vital and focuses on the strategy. "The effectiveness of any scorecard is only achieved by correct implementation within an optimised strategy," he says. "In my opinion, the results achieved for many businesses will be mostly impacted due to these elements rather than a particular technique used or approach offered within the scorecard build itself."

Conclusion

Scorecard building, it seems, is more about understanding the problem than it is about the specific approach taken. If the power of the scorecard seems incredibly high, then make sure the validation sample reflects this performance. Too good to be true often means too good to be true.

The new technologies such as neural networks can be culprits of over-fitting. The issue is not that the techniques are worse, in fact they should be better than regression approaches. The problem is that scorecards need to be understood. In a highly mathematical discipline, the best scorecard may, in fact, be the simplest.

Author Murray Bailey is the technical editor of Credit Risk International
Email: murray@crimagcom

References

1. Murray Bailey, "Building scorecards: The pits and the pendulum", Credit Risk International April 2003, Blue Moon Publishing

2. Hand D.J., Henley W.E., "Statistical classification methods in consumer credit", *Journal of the Royal Statistical Society, Series A*, 160, 523-541 (1997)
3. Boyle M., Crook J.N., Hamilton R., Thomas L.C., "Methods for credit scoring applied to slow payers in Credit scoring and Credit Control", Oxford University Press, Oxford, pp75-90 (1992)
4. Srinivasan V., Kim Y.H. "Credit granting: a comparative analysis of classification procedures", *Journal of Finance* 42, 665-683 (1987)
5. Yobas M.B., Crook J.N., Ross P. "Credit scoring using neural and evolutionary techniques", Working Paper 97/2, Credit Research Centre, University of Edinburgh
6. Desai V.S., Crook J.N., Overstreet G.A., "A comparison of neural networks and linear scoring models in the credit environment", *European Journal of Operational Research* 95, 24-37 (1996)
7. Desai V.S., Conway D.G., Crook J.N., Overstreet G.A., "Credit scoring models in the credit union environment using neural networks and genetic algorithms", *IMA Journal of Mathematics applied in Business and Industry* 8, 323-346 (1997)
8. Oxley J. internal research by Experian, previously unpublished

This article as published in Credit Risk International May 2003, Blue Moon Publishing.