CREDIT SCORING
Keywords
Performance definition
'Goods' and 'bads'
Roll rates
Indeterminates
Predictive power
Reject inference

Exploding the myths of scorecard development

_by Eva Neves

Best practice guidelines are all very well, but sometimes common sense is all that is needed in scorecard development. Eva Neves, senior consultant at South Africa-based risk management consultancy PIC Solutions, highlights some scorecard design myths and explains how these can be exploded by examining the client's culture and processes, and by understanding the problem that the scorecard is trying to solve.

One of the biggest concerns that credit risk professionals have regarding credit scoring involves the number of misconceptions that exist in model development.

Take, for instance, the amount of scoring models that are built just because it is fashionable to have one. I was once taking part in a customer delivery meeting and discussing the possible uses of a scorecard that I had just built when I realised that the client had never planned to actually use the scorecard - even though a lot of money had clearly been spent in the development and implementation of the model. In this particular organisation, the branches had most of the decision-making responsibility and they were not going to allow the model from head office decide for them.

Needless to say, I asked the client company why it was prepared to spend money on something it would never use. The company replied that it all boiled down to having something that its competitors didn't have, and that even if the scorecard was never going to be used, it still saw the scorecard as a benefit. However, I could only think of the benefits that the scorecard could have provided.

Many organisations build models without thinking through how they are going to be used. The most important step they need to take is to clearly identify a business problem for which they might need a model. Once this is done, companies can start designing their models and optimise their decision-making.

Performance definition

Once the problem has been identified, the next step is to provide a performance definition. This is the where the decision is made on what the model is going to predict. Whether it be risk, attrition, spend or response, this has to be completely tailored to the organisation and product environment. There is a lot to be said for best practice in the design of models. However, there is no best practice performance definition that can be translated from portfolio to portfolio or from country to country. In other words, there is no one-size-fits-all solution - each portfolio and model needs to be analysed on its own merits to ensure that the organisation's business objectives, its collections practices and the model's future effectiveness is optimised.

In the risk area, for example, it is important to analyse what the organisation considers to be a 'bad' customer. This is usually the starting point. The main question asked during scorecard developments is "who is the customer that the organisation would rather have not accepted had they known what they know now?" The best way to determine this is to approach the question from the profitability point of view. A good indication is usually obtained by observing the current and future collections strategies applied through the account life cycle. By analysing this, it is usually easy to pinpoint at what level of delinquency the organisation should start taking harsher actions. This should be the first stab at the performance definition of the model.

Measuring the write-off stage

It is also imperative to find an early measure of a write-off stage. After all, the main business objective when building these models is to minimise bad debt. During scorecard developments, it is not possible to go back in time and observe the entire life cycle of an account, as this might make the sample too old

Many organisations build models without thinking through how they are going to be used

Keywords
Performance definition
'Goods' and 'bads'
Roll rates
Indeterminates
Predictive power
Reject inference

to be representative of future populations. For this reason, it is very important that roll rates are analysed. Roll rates are transition matrices that look at the movement in delinquency stages. They provide an indication of how accounts roll forward into higher stages of delinquency, and how they 'cure and mill' (pay sufficient amounts to stay at the same level). If the rolling to written-off rate is included, then this is the final measure of what the future write-off rate is going to be, and it is best practice to back the initial 'bad' definition with the figures obtained from this report.

Once you have selected the delinquency level, it is best practice to define a bad customer. Some organisations argue that if the customer ever reached a certain level of delinquency but then paid up, they should not be classified as bad accounts. In reality, this raises the issue of whether the performance definition is correct in the first place.

In order to build an effective model, it is recommended that the definition is kept simple. Remember that it must be as stable as possible but also has to be easy to understand and remember. The whole strategy and use of the scorecard depend on what you are trying to predict, and there is no benefit in complicating the daily use of the model by creating a very complex definition.

Although there are fundamental differences, the concepts themselves apply to any type of model and business objective. First, identify the business issue that you are trying to address and work out what type of model, if any, suits it best. Ensure you understand the processes behind it and at what level the action is going to be taken. Finally, make sure you cater for any changes that need to be made to your systems to effectively implement and use the model. The definition of a bad account (being risk, closure or no response) will then easily be determined.

Concept of indeterminates

Even when these analyses are carried out, it is very difficult to select an exact break where a bad is clearly differentiated from a good. Furthermore, some of the accounts that are still on the books will not have had enough time to mature and show their true performance. To take this into account and ensure that the model can easily differentiate the good and bad patterns in the data, the concept of indeterminates was created. This makes it easier for the scorecard to work more effectively in those grey areas where it is needed the most.

Usually, a certain percentage of indeterminates is considered optimal. The guidelines are really common sense: if this group is too small, the scorecard will struggle to differentiate between the characteristics that make a good different from a bad. If this is taken too far, though, the model will only be able to predict the super goods and super bads, giving no benefit to the decision-making process. As an example, the rates in application scorecards should be about 5 per cent to 15 per cent of the population. In behaviour, this is about 10 per cent to 20 per cent due to the shorter performance period.

Predictive power

There are many measures of how predictive a model is. What is important to realise is that you may increase or decrease these measures by simply changing the design of your model. If you obtain an unusually high predictivity when building a scorecard, you need to investigate why this is happening. In some cases, if the design is right, it may mean that the data has been overwritten and is not what was available when you actually made the decision. It is obvious that customers that go bad usually change profiles to a worse level of risk after applying for a product. This will not have been known at the time of application, though. Any data obtained after the customer has defaulted will usually become highly predictive, so looking at unusual predictive ranges helps identify this problem.

The way in which predictive power can be artificially manipulated is by simply increasing the indeterminate rate. As the patterns are more distinguishable between super goods and super bads, the predictivity of the model simply increases. In one of the developments I was once involved in, I came across a model where we just could not get a good enough prediction. It was so low because of the data we had available, but we were so worried we actually went back and changed the performance definition to increase the indeterminate rate and observe the effect. Apart from the predictivity increase obtained, the patterns observed remained the same, only that they became stronger in each characteristic analysed (ie, they were just more prompt to identify the risk extremes rather than the grey areas). We still kept the indeterminate rate within an appropriate range, so this did not damage the model, but it certainly changed my perspective on how to design models.

To obtain a good model, the design must be right and the data used for development must be checked thoroughly. The best measure for the model is whether this is right for your organisation and objective.

This also raises a question when comparing scorecards, particularly when built by different vendors or in-house teams. In order for this comparison to be fair, the models must use the same performance definition. If not, the highest indeterminate rate might win, even though this might decrease the benefits of the model. This also applies to any other exclusions made during the development.

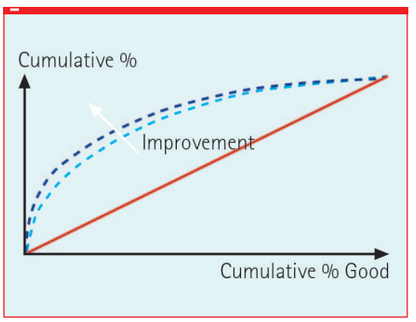

Figure 1: Trade-off curves: two-scorecard comparison

There are many different measures of predictive power. The main ones when building scorecards are Divergence and Gini. Although these both provide a very good indication on the separation of goods and bads, they are just summary statistics; they do not explain where the lift actually comes from. Personally, I always think that trade-off curves (also known as Lorent curves) provide the best view on how good a scorecard is.

Figure 1 (above) provides an illustration of this type of curve. Ascending cumulative percentages of goods and bads are plotted as score increases. The closer the curve is to the top left corner, the higher the ability to predict a bad account. If the curve lies on the red line, the ability to differentiate the goods from the bads does not exist as for every good account there will be an equal number of bad accounts.

Trade-off curves provide information on where the lift comes from. This is very important when comparing different scorecards and is widely used in segmentation analyses. In general, even when one scorecard has better divergence/gini than another, what is important is whether the lift obtained affects the area where the organisation operates. If, for example, 95 per cent of the applications coming through the door are accepted, the fact that a scorecard is better than another at the high levels of scores makes no difference whatsoever to the organisation.

Selection of characteristics

One of main steps in the scorecard build is the selection of characteristics. The types of

variables do vary across different types of models but there are general guidelines that are applied throughout.

The scorecard developer must ensure that any characteristic included can be implemented in the life system. The data required to score must be easy to obtain or retrieve from the different systems and any calculations required must be taken into consideration when building the scorecard.

Also, the use of currency value characteristics like income has been heavily criticised due to the effects of inflation, so it is usually best to use percentages like debt ratio. Although preferred, the latter can sometimes be more difficult to interpret or even be a lot less predictive, and therefore might be inappropriate. The argument of inflation also loses weight in countries where this is not really an issue, and the variable can be made a lot more stable if classed tightly. As this affects all customers in the same way, a possible solution is to simply apply inflation factors to the bandings of these characteristics on a regular basis, as is usually done within the credit limit strategies.

For the model to be robust, stable characteristics have to be selected. If the organisation heavily influences the frequency of the loan that customers are allowed to apply, it is also an issue to take into consideration.

What is most important, though, is to ensure the model is as independent as possible from external factors. For example, it is recommended that whenever bureau scores are used in decision-making, these are not included as a characteristic in the scorecard. This ensures that any bureau scorecard redevelopment or system down time does not affect the model. Best practice would be to combine scores using dual matrices, which provide the same results and greater flexibility when changing processes.

Inclusion of bureau information

Many scorecard developers do include bureau information on the actual scorecards. In most countries, however, this only includes negative information, so even though this has its benefits, it is important to understand why these characteristics usually overpower the other application information. Most of these will actually affect a very small number of customers, mainly the super bads, so although the variable separates this small group of customers very well, it is not providing benefit in the "not so bad" accounts, where the bulk of the portfolio really is.

Although using them provides certain uplift, it is best practice to ensure that they do not dominate the scorecard. They also make the scoring process dependent on the bureau, so before selecting them for the model, this has to be taken into account both from the cost and the processing point of view. If included, a strategy must be put in place in the cases where no bureau is available at the time of scoring.

One issue that is fundamental in selecting the characteristics is palatability. This affects the scorecard in two ways. First, the observed patterns must be reasonable so that no overfitting of the sample exists. If there are factors that cannot be understood, there might be an issue with the data used for development. Unless the variable is not correlated with any other characteristics, it is not wise to include in the model anything that cannot be explained. The second issue is the buy-in from the business side into the scorecard. The organisation must believe in the scorecard, so it is sometimes best practice to lose predictive power in order to gain trust from the model's end-users. This will also reduce the amount of overriding of the scorecard.

Reject inference

Reject inference is the process by which applications that were declined in the past are inferred to ensure that the future scorecard caters for all risk profiles on which the organisation makes a decision. This process is essential when the acceptance rate is very low and mostly applies to the application area.

“The scorecard developer must ensure that any characteristics included can be implemented in the life system. The data required to score must be easy to obtain or retrieve from the different systems and any calculations required must be taken into consideration when building the scorecard.”

Although this process mainly applies to declined applications, it is often forgotten that it is also required in the case of uncashed applications or applications that are not taken up. In some types of products, the uncashed rate can actually be higher than the reject rate, and if not inferred, the scorecard might not be applicable to overall population.

Eva Neves

The benefits of reject inference have been questioned over and over. In reality, it is very dangerous to infer performance when acceptance rates are very low. Because some of the profiles of the rejects have never been accepted, it will be impossible to know whether they would have been good or bad accounts.

This is why reject inference is actually an art rather than a science, and it completely relies on the experience of the scorecard developer. The ideal situation would be if all applications were accepted for a period of time and their performance observed. As this is not possible, we still have to rely on the common sense and experience of the analysts.

Regardless of the pitfalls though, it is definitely better to include this process as part of the scorecard development rather than excluding a whole portion of the population. However, the risks are such that it is important to understand the scorecard characteristics as much as possible to ensure nothing was done wrong in the inferring process.

In the main, however, the success of any model will depend on the design of the actual business issue and the understanding applied throughout the scorecard development process by the business and scorecard development teams. Without these, the scorecard will either not cater for the specific organisation's needs, or it will overfit the sample used for development, or it will simply be overwritten because it has not been sold to the business properly - in which case, why was it built in the first place?

Eva Neves is a senior consultant at risk management consultancy PIC Solutions, South Africa
Email: ENeves@PICSolutions.com