Building a Scorecard in Practice

K. Leung* and F. Cheong and C. Cheong

School of Business Information Technology, RMIT University,
Melbourne, Victoria, Australia
**E-mail: kevin.leung@rmit.edu.au*
www.rmit.edu.au

S. O'Farrell and R. Tissington

Credit Analytics Department, ANZ Bank
Melbourne, Victoria, Australia

Credit scoring is one of the most successful applications in banking and finance. However, most studies do not explain the whole process of scorecard development, probably due to the difficulty in obtaining credit scoring data. This study addresses some of the gaps that are present in the existing literature in that it explains in detail the processes, as performed in practice, of scorecard development.

Keywords: Credit scoring; Scorecard; Practical

1. Introduction

Credit scoring has now become a very important task in the credit industry and its use has increased at a phenomenal speed through the mass issue of credit cards since the 1960s [1]. It is used to produce a score, which represents a measure of confidence that classifies applicants into either 'good' (those who are likely to repay their financial obligations) or 'bad' (those who are likely to have their applications denied as a result of their high probability of default).

While the literature on credit scoring is vast, most of these studies, to the best of the authors' knowledge, have been dealing with either benchmark datasets (e.g. Australian and German dataset which are publicly available from UCI [2]) or real datasets which are obtained from financial institutions, but which usually contain only the relevant variables [3,4]. We were not able to find any studies which explain the whole process of scorecard development and validation as performed on a real raw dataset.

The interesting and challenging part of this study is that it makes use of a very large dataset, obtained from a major Australian bank and which consists of 138 variables with 38,766 records. Considering the large number of variables, data cleaning and variable selection become a challenging task. Therefore, this study will replicate most of the scorecard development processes, which are similar to what are usually done in practice, based on a large set of raw data.

2. A review on Scorecard Development

Many studies have been done on the use of different techniques for developing scorecards. Historically, statistical techniques have been widely used. These include, but are not limited to, discriminant analysis and logistic regression. They are still being used by financial institutions because they are simple and yet robust.

Advances in technology allow other methods to be used. More recently, there has been a lot of research on the use of intelligent system (IS) techniques, such as artificial neural networks, genetic algorithms, and artificial immune systems for generating scorecards. Some studies [5,6] found that statistical techniques perform better than the IS systems, while others [7,8] concluded just the opposite.

While the aforementioned studies focus mainly on the techniques for scorecard generation, none of them explain the whole process of model development from the time the data is obtained to the point where the model is ready. This is probably because the datasets already contained the most important variables and there was no need to explain the processes prior to developing the scorecard using a particular technique.

This study will therefore address some of the gaps that are present in the existing credit scoring literature in that it will explain in detail the processes (from start to finish) of scorecard generation.

3. Scorecard Generation

A large set of raw data was obtained from a leading Australian bank. It contains many different kinds of variables including personal details, job and credit history of applicants as well as their final decision outcomes, which can either be 'good', 'bad', or 'unknown'. Most the information is obtained from the customers' application form.

Developing a scorecard in practice is a lengthy process and involves many steps: 1) data cleaning, 2) data discretization, 3) variable selection,

4) samples generation, and 5) model development and validation. These will be explained in more detail in the following sections.

3.1. *Data Cleaning*

One of the most important factors which affect the success of scorecards is the quality of data. If information is irrelevant or redundant, or the data is noisy and unreliable, then knowledge discovery during the model development phase becomes more difficult [9]. Thus, data cleaning becomes a fundamental process in scorecard development.

Data cleaning was first done at a record-level. Applications which were excluded from the dataset were classified into two categories, namely application exclusions and decision exclusions. As its name implies, application exclusions are those applications which are excluded at the point of application. Some examples of application exclusions could be that the applicant is a staff of the financial institution or that the application was already pre-approved. Decision exclusions, on the other hand, are those applications which are disregarded because their decision outcomes are 'unknown'.

The data was then cleaned at an attribute-level. Attributes, such as the gender of the applicant, which cannot be included in the development sample due to legal reasons, were removed. Other attributes, for example application number and customer number, which are not likely to add any additional efficiency gains to the scorecard, were also removed. Variables with a high percentage of missing values were also excluded.

3.2. *Data Discretization/Classing*

Due to the fact that the range of some continuous variables was so large and because of the presence of outliers, the dataset was discretized. Discretization divides the interval of the values of a numeric attribute into a number of intervals, whereby each interval can be treated as one value of a categorical attribute. It can help understand the relationship between the attributes and the dependent variable and several studies [10,11] found that discretization of numeric attributes often leads to better decisions.

Many studies [12,13] made use of discretization algorithms; however, in this study, discretization was performed based on the recommendations from credit scoring practitioners. The discretization process, in this study, was performed in two separate, but iterative stages:

(i) Fine Classing
The first stage is called fine classing, whereby the raw data is examined

for its reliability and suitability; and then categorized into smaller groups. An example of fine classing is: if 95% of the 'income' variable were between $30,000 and $200,000, then they could be categorized into groups of $5,000. Missing values are also categorized into separate classes.

(ii) Coarse Classing
The second stage of discretization is called coarse classing. Data is aggregated into stable and statistically significant groups. The data is converted into standardized good/bad ratio, also known as the weight of evidence (WOE). It is calculated as follows:

$$WOE = \ln\left(\frac{p(value = good)}{p(value = bad)}\right) \tag{1}$$

with $p(value = good)$ being the number of 'good' that have this value for the attribute divided by the total number of 'good' and $p(value = bad)$ being the number of 'bad' having this value for the attribute divided by the total number of 'bad'.
Coarse classing is performed on each attribute with the goal of minimizing the drop in its information value without breaching coarse classing standards. Usually, most financial institutions would have their own classing standards and one example is to have a minimum of 5% 'bad' for each group.
The process of fine and coarse classing is recursively performed on each attribute until the coarse classing standards are satisfied. This will result in a new set of clean data in which the values of the attributes are represented by their corresponding weights of evidence.

3.3. *Variable Selection*

Model validity requires all of the variables to be included; however, practical application requires that the number of variables to be of a reasonably small value. It is therefore important to have parsimonious scorecards that only consider a small number of attributes to make the credit granting decision. Not only is there a potential of over-fitting the data with a lot of variables [14], but scorecard efficiency is also reduced with too many variables included.

In practice, stepwise regression analysis is used as the main variable selection technique. It makes use of a sequence of F-statistics to control the inclusion and exclusion of variables [15].

The result obtained from the stepwise iteration process is shown in Table 1. 20 variables were selected for inclusion in the model. The coefficient of the variables is shown in column B. The Wald statistic is a test used to check whether the relationship between the dependent and independent variable is statistically significant. All 20 variables are statistically significant at 5% confidence level, suggesting that all the variables are useful to the model.

Table 1. Stepwise Regression Results.

Num	Variable	B	Wald	Sig
1	Age of bureau file	5.16	14.983	0.000
2	Time at current address	0.739	14.918	0.000
3	Savings balance 2	0.657	14.397	0.000
4	Savings 1 - years open	0.372	11.533	0.001
5	Savings 2 - years open	0.538	3.862	0.049
6	Amount owing on home loan	0.818	17.826	0.000
7	Balance 1 income	0.748	43.975	0.000
8	Other credit limits	0.8	8.521	0.004
9	Sum of balances for customer	0.31	9.048	0.003
10	Number of dependants	0.713	29.164	0.000
11	Drivers licence indicator	1.078	25.155	0.000
12	Number of searches last 6 months	0.667	18.269	0.000
13	Number of loans	4.135	8.329	0.004
14	Total search	0.861	19.623	0.000
15	Number of address changes	0.395	4.358	0.037
16	Home phone indicator	1.291	5.081	0.024
17	Mobile phone indicator	0.636	4.102	0.043
18	Referee phone indicator	4.703	3.949	0.047
19	Occupational group	1.001	83.261	0.000
20	Age of additional card holder	1.914	5.452	0.020

3.4. *Development and Holdout Sample Generation*

Once the data is cleaned and the proper variables selected, the development and holdout sample can be generated. The development sample is used to develop the model while the holdout sample is used for model testing. Usually a stratified sampling method is applied as it not only ensures that the sample is randomly chosen, but it is also made to reflect the population in some specific characteristics. The new set of clean data consists of 20 variables with 6.7% 'bad' and 93.3% 'good' instances. Using the stratified sampling method, the dataset was divided into an 80% development and a 20% holdout sample, each containing 6.7% 'bad' and 93.3% 'good' instances.

It should be noted that the generated samples are built from accepted

applicants who are either 'good' or 'bad' applicants. However, no information is available on applicants who were denied credit. Consequently, this phenomenon introduces some bias in the samples. The idea of reject inference has been suggested to cater for this problem. It is the process of deducing how a rejected applicant would have behaved had he/she be granted credit. In practice, this data is also included in the development sample in order to have a complete picture of the population applying for credit. In this study, the scorecard developed was not inferred with information of rejected applicants simply because it was not available.

3.5. *Model Development and Validation*

Logistic regression is the most widely used technique by most financial institutions for credit scoring. Indeed, it is probably the most suitable statistical approach since the outcome of credit scoring is binary, i.e. grant or refuse credit. Thus, a scorecard is developed by applying logistic regression on the development sample. By using the holdout sample, it is then tested using the Gini (G) coefficient, which is the main performance measure used by financial institutions. If the difference in the G coefficient between the development and holdout samples is less than 10%, then the model is not over-fitted and is thus valid.

The G coefficient obtained for the development and holdout samples were 54.4% and 58.0% respectively. The results clearly indicate that the model is valid and not over-fitted, having a difference in G coefficient of less than 10%. The results also show that the scorecard performs quite well since a typical scorecard has a G coefficient ranging from 40%-70%.

The results also show that the G coefficient for the holdout sample is higher than that of the development sample, suggesting that the scorecard performs better on the data that was not used to develop it. While this scenario is unusual, it is not impossible and the most likely reason for that is due to the small volume of 'bad' in the holdout sample.

4. Conclusion

In this study, a valid and well-performing scorecard was developed from a large set of raw data. All the processes from data cleaning to the development and validation of the scorecard were explained in detail. The literature describing these processes is fairly limited probably because of the scarcity of real data in that area. These processes explained in this study are very similar to what are usually performed in practice. Two minor points worth

mentioning are: 1) reject inference was not included in the model development and 2) the last step in building a scorecard in practice is model approval, which could not be applied to this study. The model must be approved by the financial institution and usually the model is subject to small adjustments depending on the business rules used by the institution.

References

1. L. C. Thomas, D. B. Edelman and J. N. Crook, *Credit Scoring and Its Applications* (Elsevier Science Publishers, North-Holland, Amsterdam, 2002).
2. C. L. Blake and C. J. Merz, Uci repository of machine learning databases (1998), `http://www.ics.uci.edu/~mlearn/MLRepository.html`.
3. B. Baesens, Developing intelligent systems for credit scoring using machine learning techniques, PhD thesis, Katholieke Universiteit Leuven2003.
4. N. Boonyanunta, Improving the predictive power of consumer credit classification modeling, PhD thesis, RMIT University, (Melbourne, Australia, 2005).
5. M. Boyle, J. N. Crook, R. Hamilton and L. C. Thomas, *Methods for credit scoring applied to slow payers*, in *Credit Scoring and Credit Control*, eds. L. C. Thomas, J. N. Crook and D. B. Edelman (Oxford University Press, 1992), pp. 75–90.
6. M. B. Yobas, J. N. Crook and P. Ross, *IMA Journal of Mathematics Applied in Business and Industry* **11**, 111 (2000).
7. R. Malhotra and D. K. Malhotra, *The International Journal of Management Science* **31**, 83 (2003).
8. T. Lee and I. F. Chen, *Expert Systems with Applications* **28**, 1743 (2005).
9. M. A. Hall and G. Holmes, *IEEE Transactions On Knowledge and Data Engineering* **15**, 1437 (2003).
10. J. Dougherty, R. Kohavi and M. Sahami, Supervised and unsupervised discretization discretization of continuous features, in *Proc. Twelfth International Conference on Machine Learning*, (Los Altos, CA, 1995).
11. P. Perner and S. Trautzsch, Multi-interval discretization methods for decision tree learning, in *Proc. Advances in Pattern Recognition*, 1998.
12. R. M. kirby, Z. Yosibash and G. E. Karniadakis, *Journal of Computational Physics* **223**, 489 (2007).
13. U. M. Fayyad and K. B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in *Proc. Thirteenth International Joint Conference on Artificial Intelligence*, (Chambery, France, 1993).
14. R. Sanche and K. Lonergan, *Casualty Actuarial Society Forum* , 89 (2006).
15. R. L. Mason, R. F. Gunst and J. L. Hess, *Variable selection techniques*, in *Statistical Design and Analysis of Experiments: With Applications to Engineering and Science*, (John Wiley and Sons, 2003), pp. 659–687.