# A Comparison of Variable Selection Techniques for Credit Scoring

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*

Selecting new and more predictive variables is fundamental for scorecards to perform well. This study makes use of a very large set of credit scoring data and investigates the application of several variable selection techniques for scorecard development. Among the four different techniques used, stepwise regression, which is currently the most popular technique used in practice, was found to perform best.

*Keywords*: Credit scoring; Variable selection

## 1. Introduction

Credit scoring is one of the most successful operational research technique used in the financial sector and the literature on the use of different statistical and intelligent system (IS) techniques to improve credit scoring models performance is vast. However, in our previous work [1], we found that the performance of scorecards, whether being developed by IS or statistical techniques, was not significantly different from each other. This was also supported by Hand and Henley [2] who believed that it is more likely for scorecards to have significant improvements from new and more predictive variables rather than the use of more sophisticated techniques.

While there have been some studies which focus on specific variable selection (VS) techniques, we could only find one [3] which had a discussion (no empirical work) on the use of several VS techniques in the field of credit scoring. Given the fact that better explanatory variables tend to improve

scorecard performance significantly, this study will investigate and compare several VS techniques based on a large set of credit scoring data which was obtained from a major Australian bank. The dataset consists of 138 variables with 38,766 records.

## 2. A Review of Variable Selection Techniques in Credit Scoring

VS is defined as the process of finding the best subset of features from the original set of features in a given dataset [4]. It becomes important when there is a need to model the relationship between a predictor variable and a subset of potential explanatory variables, but there is uncertainty in regards to the subset to be used [5]. There are many different methods for selecting variables; however, the literature on this subject in the field of credit scoring is fairly limited. This is probably because 1) credit scoring data is difficult to obtain and 2) most of these studies made use of either benchmark dataset (e.g. UCI Australian and German dataset [6]) or real datasets containing only the most relevant variables.

Morrison [3] gave a good discussion of several VS techniques for credit scoring. In this study, we extend the work of Morrison in that we do some empirical investigations on these VS techniques. The latter are stepwise regression, factor analysis, variable clustering and partial least square.

## 3. Variables Selection Techniques

The dataset was obtained from a leading Australian bank and it consists of different kinds of attributes from the personal details to the credit history of applicants. Before applying the different VS techniques to the dataset, it was first cleaned at both record and attribute levels. The new set of clean data consists of 50 attributes with 15,576 records.

### 3.1. *Stepwise Regression*

Stepwise selection is one of the most popular procedures in the VS literature [7]. Predictor variables are added or deleted to the prediction equation, generally one at a time [8] by making use of the F-statistics which reflect the variables' contribution to the model if they are used.

Stepwise regression is probably the technique of choice in the field of credit scoring, being widely used in practice. As such, it will be used as the benchmark method against which the other VS techniques will be compared.

Based on the stepwise regression technique, 20 variables were selected for inclusion in the scorecard development process. While only the names of the variables are given in Appendix A, the results obtained show that all the 20 variables are statistically significant at 5% confidence level, suggesting that all the variables are useful to the model.

### 3.2. *Factor Analysis*

Factor analysis is a standard statistical technique that can reduce a large number of variables into a smaller set of factors that summarises the essential information contained in the variables. It is frequently used as an exploratory data reduction technique and is an appropriate tool for variable selection. Indeed, a survey of a recent two-year period in PsycINFO yielded over 1700 studies that used some form of exploratory factor analysis [9].

The assignment of the variables to the factors can be improved by performing a factor rotation [10]. While unrotated factor solutions achieve the objective of data reduction, it does not provide adequate information on the interpretation of the variables involved. Factor rotation simplifies the factor structure, thereby improving its interpretation by removing some of the ambiguities that are often present in the initial unrotated factor solutions [11].

Factor analysis was performed using the principle axis factors as the extraction method and the varimax as the rotation method. The Kaiser criterion (i.e. all factors with eigenvalues greater than one) was used to obtain the groups of variables. As for the cut-off value of the factor loading which is used to evaluate the factor patterns, Tabachnick and Fidell [12] suggest 0.32 as a good rule of thumb for the minimum loading of an item, which equates to approximately 10% overlapping variance with the other items in that factor. We used a cut-off value of 0.40 because too many factors and variables were extracted.

The results obtained indicate that 12 factors were extracted, containing 31 variables altogether. Compared to the stepwise regression technique in which the number of variables was reduced to 20, factor analysis generates 31 variables to be fitted into the model. In order to reduce the high number of variables, the composite value of each factor was calculated by combining the variables of each factor into a single composite measure. A composite scale provides two specific advantages [11]. Not only is it able to represent multiple aspects of a concept in a single measure, thereby simplifying interpretation of results, but it also provides a means of overcoming, to some extent, the measurement error inherent in all measured variables. By do-

ing so, only 12 variables were obtained for inclusion in the model. The 12 variables selected are shown in Appendix A.

### 3.3. *Variable Clustering*

Variable clustering is very similar to cluster analysis in that it splits a set of variables with similar characteristics using a set of the subject data [13]. Therefore, variable clustering results in groups of variables where variables in a group or cluster are similar to other variables in the same cluster and as dissimilar as possible to variables in another cluster.

The variable clustering algorithm used in this study is the VARCLUS algorithm which can be obtained from the SAS package. It is both a divisive and iterative one in that it first starts with one cluster containing all the variables and recursively splits existing clusters into two sub-clusters if the second eigenvalue for the cluster is greater than a specified threshold. It does so by computing the first two principal components and assigning each variable to the component with which it has the highest squared correlation. A testing procedure is then performed to check if assigning each variable to a different cluster increases the amount of variance explained. If a variable is assigned to different cluster, the components of the two clusters involved are recomputed before the next variable is tested [14]. The algorithm stops when the maximum number of clusters is attained or when a certain percentage of variation explained is reached.

In the clustering literature, a rule exists for selecting the cluster representative. That rule dictates to select the variable with the minimum $(1 - R^2_{ratio})$ as the cluster representative. It is defined as:

$$(1 - R^2_{ratio}) = \left( \frac{(1 - R^2_{own})}{(1 - R^2_{nearest})} \right) \tag{1}$$

Intuitively, there is a need to select the cluster representative which is as closely correlated to its own cluster $(1 - R^2_{own})$ and as uncorrelated as possible to the nearest cluster $(1 - R^2_{nearest})$. Therefore, the optimal representative of a cluster is variable where $(1 - R^2_{ratio})$ tends to zero [13]. 18 variables were selected and are shown in Appendix A.

### 3.4. *Partial Least Square*

PLS was developed in the 1970s by Herman Wold [15] as an econometric technique. It has since been used in a variety of disciplines including chemistry, medicine, education, marketing, and the social sciences where

predictive linear modelling, especially with a large number of predictors, is necessary [16].

PLS is a predictive technique and it is particularly useful when the predictor variables are highly correlated. It combines features of principal components and multiple regression techniques. While PLS may be implemented as a regression model, predicting one or more dependents from a set of one or more independents; or implemented as a path model, akin to structural equation modelling, in this study, PLS will be used mainly as an exploratory analysis tool to select suitable predictor variables and to identify outliers. Its ability to handle large numbers of correlated predictors and this feature makes PLS quite suitable for credit scoring.

The outcomes of PLS are very similar to those of FA; however, since in this study the interest is more on finding the most appropriate predictor variables (in their original form) rather than the PLS factors, the Variable Importance for Projection (VIP) criteria, which offers a good way for recommendations to be made regarding variable selection, will be used. Indeed, the VIP coefficients, obtained by the partial least squares (PLS) regression, have had an increasing attention these days as an importance measure of each explanatory variable or predictor. As a rule of thumb advanced by Wold [17], any independent variable with (a) a small VIP coefficient ($<$ 0.8) and (b) a small regression coefficient in absolute size becomes prime candidate for deletion. As such, credit analysts can select a subset of the predictors for inclusion in the scorecards based on the VIP coefficients.

The variables (23 in total) with VIP coefficient $> 0.8$ were selected and are shown in Appendix A. The VIP value corresponds to the significant effect of a corresponding independent variable on the predictor variable.

## 4. Analysis and Results

The 4 different sets of data obtained as a result of the 4 VS techniques were each divided into an 80% development (training) sample and a 20% holdout (testing) sample using a stratified sampling method. Logistic regression, which is the most popular method used by financial institutions for scorecard development, was used. The performance of the 4 scorecards was tested using the Gini (G) coefficient, which is the main measure of performance used in practice in the field of credit scoring.

The results obtained are shown in Table 1 and show that stepwise regression is the best VS technique, having achieved the highest G coefficient for both the development and holdout sample. This is probably why this technique is still being used widely in most financial institutions. Factor

analysis recorded the worst G coefficient for the development sample. One possible reason for this is because the input data are composite values of several variables, thereby reducing the explanatory power of the variables.

Table 1. Gini Coefficient

| Technique | Development | Holdout |
|---|---|---|
| Stepwise Regression | 0.544 | 0.580 |
| Factor Analysis | 0.476 | 0.524 |
| Variable Clustering | 0.486 | 0.522 |
| Partial Least Square | 0.536 | 0.524 |

## 5. Conclusion

Selecting more relevant variables seems to have a more significant positive effect on the performance of scorecards compared to using more sophisticated classification techniques. However, the current credit scoring literature on VS is fairly limited. In this study, we performed an empirical investigation and comparison of several VS techniques for use in developing scorecards. A large credit scoring dataset was used and it was found that stepwise regression, which is widely used in the corporate world, has the best performance compared to the other three techniques.

## Appendix A. Variables Selected

The selected variables for each VS technique are shown in Table A1.

## References

1. K. Leung, F. Cheong and C. Cheong, A comparison of conventional and simple artificial immune system (SAIS) techniques in consumer credit scoring, Under review, (2008).
2. D. J. Hand and W. E. Henley, *Journal of the Royal Statistical Society* **160**, 523 (1997).
3. J. S. Morrison, *Variable selection in model development*, tech. rep., TransUnion (Atlanta, 2004).
4. K. J. Cios, W. Pedrycz and R. W. Swiniarski, *Data Mining Methods for Knowledge Discovery* (Kluwer Academic Publishers, Boston, 1998).
5. E. I. George, *Journal of the American Statistical Association* **95**, 1304 (2000).
6. C. L. Blake and C. J. Merz, Uci repository of machine learning databases (1998), `http://www.ics.uci.edu/~mlearn/MLRepository.html`.
7. T. H. Fan and K. F. Cheng, *Data and Knowledge Engineering* **63**, 811 (2007).

Table A1. Variable Selected from the VS Techniques

| Stepwise Regression | Factor Analysis | Variable Clustering | Partial Least Square |
|---|---|---|---|
| Age of additional CH | Account | % HL owing | Amt owing on HL |
| Age of bureau file | Address | Applicant age | Assets to income |
| Amt owing on HL | Age | Balance 1 income | Assets to liabilities |
| Balance 1 income | Assets | Balance tranfer amt | Balance 1 income |
| Drivers licence ind | Balance | Balance for customer | Balance for customer |
| Home phone ind | Credit | Credit limits | Bank customer |
| Mobile phone ind | HL | Drivers licence ind | Drivers licence ind |
| Num of add changes | Income | Gross annual income | HL ratio |
| Num of dependants | Inquiry | HL value | Inq 180-365 days |
| Num of loans | Payment | Mth HL payment | Inq 30-90 days |
| Num of searches | Residence | Mth payment on loan | Inq 90-180 days |
| Occupational group | Saving | Num of add changes | Mth HL payment |
| Other credit limits | | Num of dependants | Num of add changes |
| Referee phone ind | | Num of searches | Num of dependants |
| Savings 1 - yrs open | | Savings balance 1 | Num of searches |
| Savings 2 - yrs open | | Sum of balances | Occupational group |
| Savings balance 2 | | Total income | Other credit limits |
| Sum of balances | | Value of home | Savings 1 - yrs open |
| Time at current add | | | Savings balance 1 |
| Total search | | | Savings balance 2 |
| | | | Sum of balances |
| | | | Time at current add |
| | | | Total search |

8. 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.
9. J. W. Osborne and A. B. Costello, *Practical Assessment Research and Evaluation* **10** (2005).
10. I. T. Jolliffe, *Principle Component Analysis* (Springer-Verlag, New York, 2002).
11. J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson and R. L. Tatham, *Multivariate Data Analysis* (Pearson Prentice Hall, 2006).
12. B. G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics* (Allyn and Bacon, Boston, 2001).
13. R. Sanche and K. Lonergan, *Casualty Actuarial Society Forum* , 89 (2006).
14. SAS Institute Inc., *The VARCLUS procedure*, in *SAS/STAT Users Guide*, 1999, pp. 3593–3620.
15. H. Wold, Partial least squares, in *Encyclopedia of Statistical Sciences*, (New York, 1985).
16. M. Escabias, A. M. Aguilera and M. J. Valderrama, *Computational Statistics and Data Analysis* **51**, 4891 (2007).
17. S. Wold, Pls for multivariate linear modeling, in *Proc. QSAR: Chemometric Methods in Molecular Design*, (Weinheim, Germany, 1994).