Variable Selection in the Credit Card Industry

Moez Hababou, Alec Y. Cheng, and Ray Falk, Royal Bank of Scotland, Bridgeport, CT

ABSTRACT

The credit card industry is particular in its need for a wide variety of models and the wealth of data collected on customers and prospects. We propose a methodology to select variables for predictive modeling purposes out of the plethora of data available using a combination of Oblique Component Analysis (PROC VARCLUS), Information Value (IV) and Weight Of Evidence (WOE) analysis, and business intelligence. Our tools enable us to quickly identify the most informative variables for logistic regression models.

INTRODUCTION

Data mining has become central to the financial services industry as the competition for consumers has intensified and increased in recent years. As a result, there is an increasing and growing plethora of data collected on consumers. The three major bureaus (Equifax, TransUnion, and Experian) dispose now of thousands of variables that can be used for analytical purposes. For instance, Equifax provides over 1,200 credit and demographic attributes which can be used for various modeling and analytical projects. In addition, thanks to powerful data warehouses, financial institutions have managed to collect tons of data on customers and prospects which can be used for various purposes (direct marketing, retention, fraud, risk management, customer segmentation, revenue and profit forecasts, etc.). Such a wealth of data can be problematic as modelers need to sift through all these variables. It is thus important to develop mechanisms and processes which assist analysts and modelers to navigate through the maze of data, and identify a smaller set of variables.

Models can be built in several different ways but there are several common major phases in model development process: variable reduction and transformation, and model development as highlighted in Figure 1.

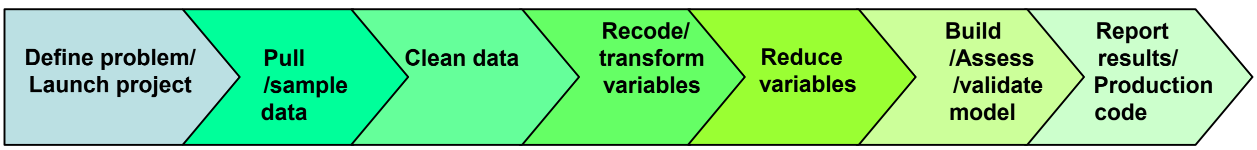

Figure 1. Typical phases of model development process

In this paper, we are concerned with Step 5, that is reducing the number of variables to a smaller manageable set that the analyst or modeler can further investigate.

THE VARIABLE REDUCTION PROCESS

Our variable reduction serves as the generic initial stage of variable selection process, which fuses with model building by itself. The aim of variable reduction is to maintain a compact set of predictors that can help to accelerate model building but not loosing potential predictive powers. To this end, we use a variant of the Oblique Component Analysis (OCA), the PROC VARCLUS facility in SAS/STAT to group predictors into clusters.

Unlike the classical principal component analysis which diagonalizes the variance-covariance matrix formed by the underlying random vector, PROC VARCLUS attempts to block-diagonalize the variance-covariance matrix through row permutations only. As a result, variables in each block (cluster) retain best similarity while correlations between blocks (clusters) are minimized. As in credit card business, model interpretation and stability is rather important, PROC VARCLUS can help to retain the maximum interpretability of variables. Intuitive variables in each cluster will be further picked up as candidate predictors based on their Information Value (IV) or correlation with the variable to be predicted/modeled. The appendix section of this paper discusses more of the Information Value concept

Relatively newer approaches use Support Vector Algorithm as a way to reduce variables. The approach to feature selection outlined in this presentation is largely a product of conventional practice and is geared to the types of models easily and naturally implemented as additive scorecards. Implications for feature selection with modern "gold standard" methods such as support vector algorithms have implications for the underlying feature selection. (In broad terms, support vector algorithms maximize the margin between groups or minimize complexity of a model subject to perfectly fitting the training data.) In particular, Forman (2005) reports that support vector models reach and maintain a threshold of classification accuracy, while conventional methods reach an (inferior) maximum classification accuracy, which decays as excess predictors are added to the series of models. In addition, support vector algorithms typically employ a nonlinear (kernel) function of feature profiles for pairs of observations to expand the predictive feature space nonlinearly. Of interest for direct feature selection is that the effectiveness of this expansion diminishes significantly as directly predictive features are made available.

PROPOSED APPROACH

The complete variable reduction/selection process can be briefly described in Figure 2

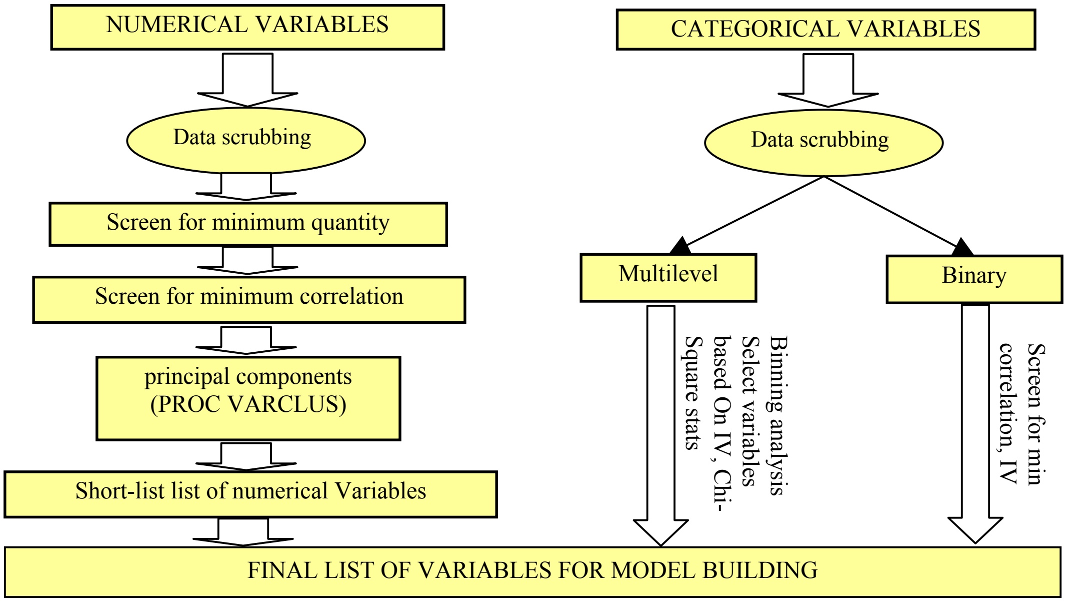

Figure 2. Flow-chart of variable reduction/selection process

Numeric and categorical variables are processed separately. Numeric variables are processed in six steps:

1. Screen out the candidate variables with more than X% of missing values.
2. Screen out the candidate variables with minimal correlation with performance variable
3. Apply Oblique Component Analysis to remaining variables, and select variables with highest information value and lowest R-square in each component.
4. Variable binning and classing based on performance variable. Final binning variables and classing variables are selected based on Weight of Evidence and Information Value
5. Variable transformation. Final variable transformations are based on chi-square statistics.
6. Missing indicators are created for all remaining variables with missing values.

Categorical variables are processed in three steps:

1. Screen out the candidate variables with minimal correlation with binary performance variable
2. Variable binning based on performance variable
3. Final binning variables are selected based on Weight of Evidence and Information Value.

NUMERICAL EXAMPLE

In this numerical example, the PROC VARCLUS analysis results in 40 groups. The next step is to choose representatives from each cluster to be further evaluated for modeling and analysis purposes. We recommend to this end to sequentially look at three criteria:

1. Maximize the Information Value of the variable
2. Minimize the 1-R2 ratio
3. Incorporate any additional business priorities

This logic will ensure that the most predictive variables are selected while taking into account the “uniqueness” of each predictor. The main purpose is to build a model (similar to scorecard modeling) which includes a wide variety of variables. To this effect, a model which draws its predictive power from multiple predictors is preferred to a model whose predictive power is drawn from few predictors. For instance, we can end up with two models with very similar predictive power. This is the main rationale behind the use of Factor Analysis for grouping variables. It is also reflected at the variable selection stage as we pick representative variables for the different clusters.

For Cluster 1, the main theme is the number of bank card trades. In this particular example, we retained Bankcard_HiBal_1, #BKCRD TRD, and #BKCRD REPORTED W/IN 3 MOS as representatives of Cluster 1 based in the IV and 1-R2 criteria. Some attributes may be included because of business priority (such as #TRD ALWAYS

SATIS). Cluster 14 includes variables addressing the number of revolving trades and as such #TRD ALWAYS SATIS was chosen as representative. Finally, Inquiries_6mos_ were chosen as representative of Cluster 18.

Business priorities and general diversity may be accommodated in some instances by scoring the selected variables over selected attributes (5 counts of trades, 1 balance amount, and 2 inquiries were selected); this may be compared with the overall availabilities of these attribute types. In particular, cumulative tallies of these meta-attributes may be compiled as variable selections are entered on the spreadsheet.

40 Clusters		*R-squared with*					
Cluster	***Variable***	***Own Cluster R**2***	***Next Closest R**2***	***1-R**2 Ratio***	***IV***	***Business priority***	***Selected***
Cluster 1	Bankcard_HiBal_1	0.8650	0.6216	0.3568	0.0959		1
Cluster 1	#BKCRD TRD	0.6671	0.6389	0.9218	0.0594		1
Cluster 1	#BKCRD TRD ALWAYS SATIS	0.6666	0.6363	0.9167	0.0587		
Cluster 1	Bankcard_num_opn_1	0.9404	0.6264	0.1595	0.0474		1
Cluster 1	Current value for Bankcard_num_opn_	0.9425	0.6303	0.1556	0.0456		
Cluster 1	#BKCRD TRD ALWAYS SATIS W/IN 6MTH	0.9352	0.6141	0.1679	0.0378	E	1
Cluster 1	#BANKCARD REPORTED W/6 MOS	0.9352	0.6144	0.1681	0.0374		
Cluster 1	#BKCRD TRD ALWAYS SATIS W/IN 3MTHS	0.9272	0.6090	0.1863	0.0350		
Cluster 1	#BANKCARD REPORTED W/IN 3 MOS	0.9271	0.6090	0.1864	0.0346		1
Cluster 14	#TRD ALWAYS SATIS	0.8062	0.3676	0.3064	0.0668		1
Cluster 14	#REVL ALWAYS SATIS	0.9438	0.5263	0.1186	0.0492		
Cluster 14	#OPN TRD	0.8839	0.6210	0.3064	0.0435		1
Cluster 14	#TRD ALWAYS SATIS W/IN 6MTHS	0.8446	0.6547	0.4501	0.0374		
Cluster 14	#TRADES REPORTED W/6 MOS	0.8441	0.6527	0.4489	0.0373		
Cluster 14	#TRADES REPORTED W/IN 3 MOS	0.8343	0.6466	0.4689	0.0365		
Cluster 14	#TRD ALWAYS SATIS W/IN 3MTHS	0.8346	0.6478	0.4697	0.0361		
Cluster 14	#OPN REVL (3146)	0.8896	0.6645	0.3291	0.0346		
Cluster 18	Inquiries_24mos_	0.8610	0.1831	0.1702	0.0802		
Cluster 18	Inquiries_6mos_	0.9192	0.1930	0.1001	0.0760		1
Cluster 18	Inquiries_3mos_	0.8982	0.1837	0.1247	0.0650		
Cluster 18	inq_12mos_4	0.8603	0.2176	0.1785	0.0285		

Table 1. Numerical example for PROC VARCLUS coupled with IV consideration

CONCLUSION

In this paper, we briefly presented an approach to reduce the complexity of the data and reduce it to a more manageable size. Our approach combines Oblique Component Analysis with the concept of Information Value in order to group attributes into likewise clusters and choose representatives to be further looked at the modeling stage. This approach yields models and scorecards which have several advantages:

- They are not dependent on a handful of predictors (hence less sensitive to population changes and data issues).
- They are easier to sell to management and business users
- They use a wider selection of predictors; hence capture a broader range of dimensions which influence the target variable.

This approach can also be used to zoom in on fewer predictors for advanced ad-hoc analytical work.

REFERENCES

Forman, G. (2005) "Feature Selection: We've barely scratched the surface", IEEE Intelligent Systems Magazine 20(6): 74-76.

Forster, D. and Stine, R. (2004), "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy", working paper A01-028-R2, Wharton School of the University of Pennsylvania.

Nelson, B.D. (2001) "Variable Reduction for Modeling using PROC VARCLUS", Conference Proceedings SAS Users Group International

SAS Institute, Inc.(2000) “Predictive Modeling Using Logistic Regression (Training Course Material)”, SAS institute.

SAS Institute, Inc. SAS/STAT User's Guide, Version 9 (online at http://support.sas.com/), SAS Institute Inc.

Siddiqi, N. (2005), “Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring”, published by John Wiley Sons Inc.

Tasche, D. (2002), “Remarks on the Monotonicity of Default Probabilities”, Deutsche Bank Publication

APPENDIX: INFORMATION VALUE AND WEIGHT OF EVIDENCE

Weight of evidence (WOE) and information value are increasingly popular in the analytical and modeling community as they represent good alternatives to approximate non-linearity in the data. WOE is subsequent to binning potential predictors into meaningful bins.

In its continuous form, Information Value (IV) is expressed as

$$IV = \int (f_G - f_B) \log \frac{f_G}{f_B} dx$$

where f_G and f_B are conditional probability densities of the predictor variable either when the ‘response’ is good or bad. In discrete form, we compute within each interval the percentage of goods (zeros) and bads (ones). The WOE and Information Value (IV) are computed with the following as:

WOE = Log(Distribution Good/Distribution Bad)

IV = $\{\Sigma$(Dist Good – Dist Bad) x WOE$\}$

The empirical rule of thumb for assessing the IV is as follows (See Siddiqi 2004): Less than 0.02: the variable is not predictive; 0.02 to 0.1: the variable has weak predictive power; 0.1 to 0.3: the variable has medium predictive power; 0.3+ : the variable has strong predictive power.

We provide a small numerical example to compute WOE and IV.

# Inquiries 0-5 months excluding last 7 days	*% Good*	*% Bad*	*WOE*	*IV contribution*
missing	5%	6.50%	-.2624	0.00
0	5%	2%	.9163	0.03
1	20%	15%	.2877	0.01
2	30%	25%	.1823	0.01
3	24%	27%	-.1178	0.00
4+	16%	25%	-.4261	0.04
total	100%	100%		0.09

Table 2. Numerical example for WOE/IV calculation

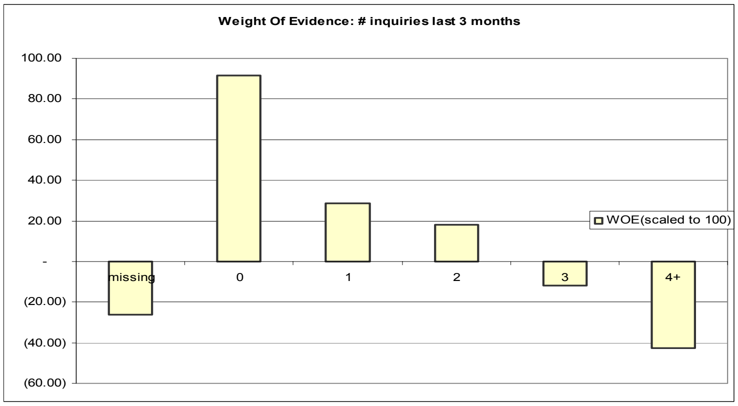

Figure 3. Example of WOE pattern

ACKNOWLEDGEMENTS

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.

CONTACT INFORMATION

Send your comments to

Moez Hababou, Ph.D.
VP Predictive Modeling
Royal Bank of Scotland National Bank
1000 Lafayette Blvd, Bridgeport CT 06854
Phone: 203-551-3055
Fax: 203-551-2011
E-mail: moez.hababou@rbsnb.com
Web: www.rbsnb.com

or

Alec Cheng
Royal Bank of Scotland
1000 Lafayette Blvd
Bridgeport, CT 06854
Phone: 203-551-5059
Fax: 203-551-2013
E-mail: yu.y.cheng@rbsnb.com