

USING DATA MINING TO DETERMINE EFFICIENT LOSN GRANTING POLICIES FOR CREDIT UNIONS

Abbas Heiat, College of Business, Montana State University-Billings, 1500 University Drive, Billings, MT 59101, 406-657-1627

ABSTRACT

The goal of this study is to determining the important variables that are significant in building a model for classifying good loans from bad loans for eight credit unions and the implications for changing policy regarding granting loans. Based on this analysis it seems the policy used by credit unions for granting loans is conservative and actually results in less return than an alternative policy.

INTRODUCTION

In the financial industry, consumers regularly request credit to make purchases. The risk for financial institutions to extend the requested credit depends on how well they distinguish the good credit applicants from the bad credit applicants. However, many credit unions use fixed conservative criteria to grant loans that results in less than optimum return on loans. An accurate estimation of risk, and its use in loan granting risk models, could be translated into a more efficient use of resources and more profit. One important approach to achieve this goal is to find more efficient predictors of loan risk in the credit portfolios of credit unions. One widely adopted approach for finding efficient predictors is the application of asset of decision models and their underlying techniques.

The goal of this study is to determining the important variables that are significant in building a model for classifying good loans from bad loans for eight credit unions and what are the implications for changing policy regarding granting loans.

DATA

In this research I have used the data set with information pertaining to customers who borrowed from eight credit unions. The data set contains information related to the customers' financial standing, reason to loan, loan type, balance status, loan rate, return on loans, total savings, delinquency status, employment, and demographic information. The data set contains observations on 31 variables for 4171 past applicants who applied for credit. Each applicant has rated as delinquent or non-delinquent. The original data set had a number of categorical variables, some of which have been transformed into a series of binary variables so that they can be appropriately handled by the data mining software (Shmueli, Patel, and Bruce, 2010).

METHODOLOGY

This study focused on determining the important variables that are significant in building a model for classifying good loans from bad loans for eight credit unions. We used IBM SPSS Modeler data mining software in this study. The followings are the seven algorithms used in this study: Support Vector Machine (SVM), Logistic Regression, Decision Tree, C5, C & R Tree, CHAID, Artificial Neural Network (ANN), and Bayes Network. We used IBM SPSS Modeler data mining software in this study.

RESULTS OF THE STUDY

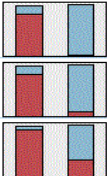


Statistical feature selection in the SPSS Modeler is used to establish the most important variables that are correlated to the target variable STATUS, (paid / good) loans from defaulted /bad loans). Figure 1 shows the result of this step. Seven variables are listed as statistically the most important: LOAN TYPE, AGE, EMPLOYMENT, LOAN AMOUNT, TERM OF LOAN, LOAN RATE, and GENDER.

Figure 1. The most important variables determined by Feture selection of SPSS Modeler

	Rank ▲	Field	Measurement	Importance	Value
<input checked="" type="checkbox"/>	1	LOAN TY...	Nominal	Important	1.0
<input checked="" type="checkbox"/>	2	AGE	Continuous	Important	1.0
<input checked="" type="checkbox"/>	3	Employed	Flag	Important	1.0
<input checked="" type="checkbox"/>	4	LOAN-AMT	Continuous	Important	1.0
<input checked="" type="checkbox"/>	5	TERM	Continuous	Important	1.0
<input checked="" type="checkbox"/>	6	LRATE	Continuous	Important	1.0
<input checked="" type="checkbox"/>	7	GENDER	Flag	Important	1.0

In next step after removing un-important variables from the data set, balancing or boosting is used to establish a balance between number of good loans and defaulted loans. Next the data was partitioned to training and testing data sets. The applied models are built based on training data set but they are evaluated using a new data set (testing) in order to avoid bias. Auto Classifier feature is used to determine the best algorithms (Models). The results in Figure 2 indicate that Decision Tree (CHAID), Logistic Regression, and Support Vector Machine (SVM) are the most reliable models for this data set.

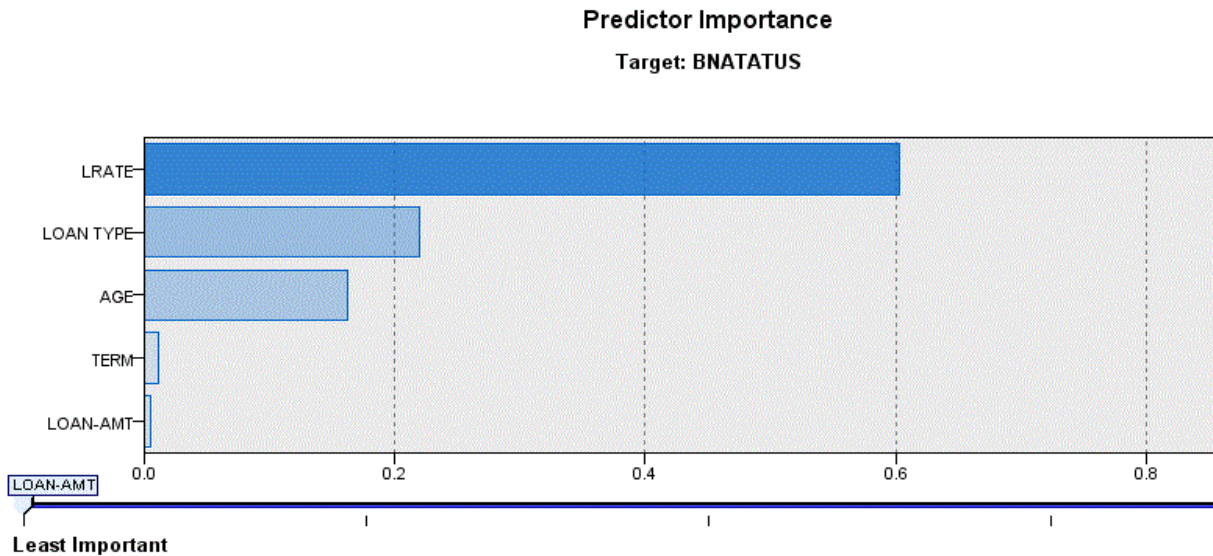
Figure 2. The most accurate algorithms found by SPSS Modeler: CHAID (decision Tree), Logistic Regression, and Support Vector Machine (SVM)

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in	Lift(To... ▾)	Overall Accuracy	No. Fields Used	Area Under Curve
<input checked="" type="checkbox"/>		C...	< 1	8,003.73	56	1.929	91.317	5	0.967
<input checked="" type="checkbox"/>		L...	< 1	7,650	52	1.867	88.422	7	0.949
<input checked="" type="checkbox"/>		S...	< 1	6,770	46	1.824	80.287	7	0.886

M)

First the Decision Tree model (CHAID) is created to study its' results. Figure 3 lists the most important variables that are determined by CHAID for classifying good loans from bad loans. These are LOAN RATE, LOAN TYPE, AGE, LOAN TERM, and LOAN AMOUNT.

Figure 3. The important variables determined by CHAID



The Decision Tree model generated decision rules for deciding whether a loan would be a good loan or bad. In the data set 1 stands for a good loan and 2 for a defaulted loan. For example, looking at the rules for 2, the first rule states that if the loan rate is equal or less than 7% and loan type is L10, then the loan probably is going to be a bad one. Of course these rules must be studied by a loan officer(s), and in case they do not make sense, they should be revised or dropped.

Confusion matrix (Figure 4) for the CHAID results show the evaluation and accuracy of the Decision Tree Model (CHAID) based on the test data. On the whole the model 90.72 % of the time correctly classifies good loans from bad which is a very good. However, the model based on the test data (second part of the Figure 7) $(1615) / (1615+288) = 84.86\%$ of the time is accurate in classifying good loans and $(1864) / (68+1864) = 97\%$ of the time is accurate in classifying bad loans. The latter, depending on the loan policy, may be more important in granting loans.

Figure 4. CHAID Confusion Matrix

Comparing \$R-BNATATUS with BNATATUS

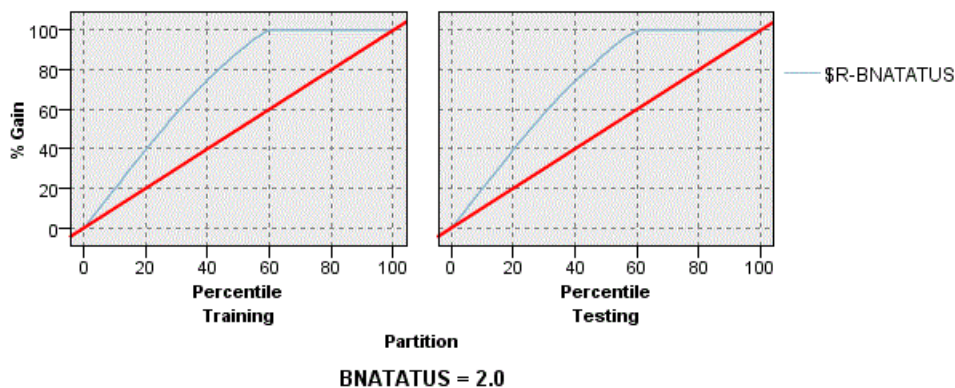
'Partition'	1_Training		2_Testing	
Correct	7,938	90.23%	3,479	90.72%
Wrong	860	9.77%	356	9.28%
Total	8,798		3,835	

Coincidence Matrix for \$R-BNATATUS (rows show actuals)

'Partition' = 1_Training	1.000000	2.000000
1.000000	3,711	702
2.000000	158	4,227
'Partition' = 2_Testing	1.000000	2.000000
1.000000	1,615	288
2.000000	68	1,864

The lift chart (Figure 5) shows how efficient is using the developed model versus using no model. The distance or gap between the red line (no model) and the blue line indicates the improvement by using a model. In this case it is almost 40%.

Figure 5. CHAID lift Curve



CONCLUSION

Based on this analysis it seems the policy used for granting loans is conservative and actually results in less return than an alternative policy. Simple calculations and cost benefit analysis using the average return on loans (Original amount of loan times Loan Term times Loan rate and loss minus 10% recovery) were used in this study yielding the following:

$288(\text{number of good loans classified as bad}) \times 1546(\text{average return on good loans}) = 148416$

$68(\text{number of bad loans classified as good}) \times (1546 \times 0.9) (\text{Average loss on bad loans}) = 94615.2$

These calculations show that even using a reliable model based on existing loan granting policy results in less return (14816 vs. 94615.2 on average). It may be a good idea to consider risk adjusted loan rate policy in granting loans. As the Decision Tree model indicates LOAN RATE, LOAN TYPE, AGE, LOAN TERM, and LOAN AMOUNT, and specially Loan Rate should be considered carefully in granting loans.

Next step in the study is to repeat analysis for each credit union separately. There are some problems to take into account. For example, the sample size for some credit unions is small. Another problem I noticed is there is no defaulted loan for one credit union. It is also possible to develop an optimization return model based on loan type or/and loan rate.

REFERENCES

- Arminger, G., D., and Bonne, T., 1997, "Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis and feed- forward networks", *Computational Statistics*, Vol. 12, pp. 293-310.
- Caudill, M., and Butler, A., 1990, *Naturally Intelligent Systems*, the MIT Press Cambridge, Massachusetts.
- Darwiche, Adnan, 2003, "A differential approach to inference in Bayesian networks", *Journal of the ACM*, Vol. 50, Issue 3, pp. 280-305.

Hand, D. J. and Henley, W. E., 1997, "Statistical Classification Methods in Consumer Credit Scoring: a Review", *Journal of Royal Statistics Society*, Part 3, pp. 523-541.

Hand, D. J., Oliver, J. J., and Lunn, A. D., 1998, "Discriminant analysis when the classes arise from a continuum", *Pattern Recognition*, Vol. 31, pp. 641-650.

Henley, W. E., 1995, *Statistical Aspects of Credit Scoring*, PhD theses, The Open University, Milton, Keynes, UK.

Sabzevari, H., Soleymani, M., Noorbakhsh, E., "A comparison between statistical and Data Mining methods for credit scoring in case of limited available data", *CRC Conference*, 2007.

Shmueli, G., Patel, N., and Bruce, P., 2010, *Data Mining for Business Intelligence*, Wiley, New Jersey.

Wiginton, J. C., 1980, "A note on the comparison of logit and discriminant models of consumer credit behavior", *Journal of Financial Quantitative Analysis*, Vol. 15, pp. 757-770.

Williamson, S., 1987, "Costly monitoring, loan contracts and equilibrium credit rationing", *Quarterly Journal of Economics*, VOL.102 (1), pp. 135-145.