

USING CLASSIFICATION TO DETERMINE LOAN REPAYMENT

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

ABSTRACT

This research uses classification and other data mining techniques to identify the factors that are relevant to determining loan repayments using data from over 1,200 small loans that were granted for a variety of purposes for the period 2006 to 2011 by a regional bank located in the United States of America. We use quarterly data to identify loans that were in default, and IBM SPSS modeler data mining software to isolate factors that seem to influence loan default.

INTRODUCTION

One of the macro goals of almost all governmental economic policy is poverty alleviation. In recent decades governmental policies have drifted towards achieving this goal through the market system. Specifically, micro loans are being increasingly used to uplift the very poor by increasing their incomes through marketable skills. However, the capital required to market those skills and derived products are provided through both, subsidized and unsubsidized direct loans. The loan principals are usually modest amounts that are not backed by any collateral other than the fixed and working capital obtained through the loan. These loans are risky because of a variety of reasons, many of them beyond the control of individual borrowers. Most studies use Logistic model to determine important default factors, however, this model was not superior to the least squares and was not cost effective. The goal of this paper is to use data mining techniques to isolate factors that are important in leading to loan defaults. The sample data consists of consumption loans rather than loans for income generation purposes.

LITERATURE REVIEW

One of the major problems with microcredit is the high loan default rate (Robinson, 2001). Many papers have been devoted to studying the loan repayment problem for microloans Mokhtar, Nartea and Gan; Acquah and Addo (2011), Ugbomeh, Achoja, Ideh and Ofuoku (2008), Okorie (1986), Njoku and Odii (1991). All of these papers are specific to a particular region or industry in the developing world where microloans are more popular. Mokhtar, et al (2011) study the problem for two institutions in Malaysia. They use a Logit regression model to identify factors like gender, age, business type, repayment period, repayment amount that influenced loan repayments. They found that male borrowers, seasonal businesses, frequent repayment periods that were not in consonance with the revenue cycle were the primary factors that were responsible for a loan going bad. Ugbomeh, et. al. (2008) found that household size and interest rates were some of the factors that negatively affected loan repayments for women in Nigeria. Acquah and Jaddo (2011) found that defaults among fishermen in Ghana increased with the age of the borrower and the amount of the initial investment. The predominant technique used to evaluate loan defaults was the use of the least squares model, Ugbomeh, et.al. (2008) or the

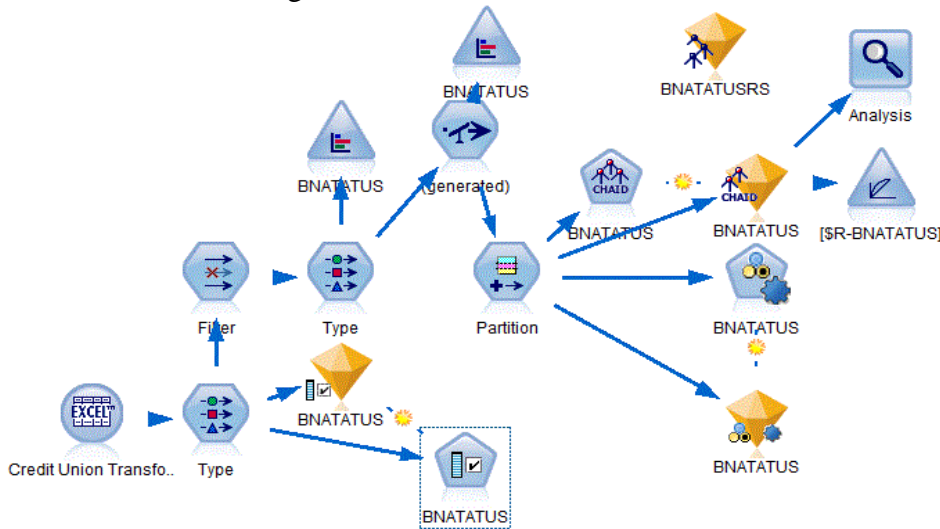
Logistic model, Mokhtar, et.al. However, Henley (1995) found that the logistic model was no better than linear regression and Wiginton (1980) found it to be not cost effective. The purpose of this paper is to use data mining techniques to find efficient predictors of consumer loan risk in credit portfolios. The goal is to determine the important variables that are significant in building a model for classifying good loans from bad loans. This determination will have policy implications for individual credit institutions.

RESEARCH METHOD

The data was obtained from multiple credit unions with a large geographic spread. The data set contains information related to the customers’ financial standing, employment, age, gender, loan type, balance, loan rate, total savings and delinquency status. The data set contains quarterly observations for 1200 accounts for a two year period from 2010 to 2012. This study focuses on determining the important variables that are significant in building a model for classifying good loans from bad loans. We use an IBM SPSS Modeler data mining software for this study. The following eight algorithms are used in this study: Support Vector Machine (SVM), Logistic regression, Decision Tree, C5, C&R Tree, CHAID, Artificial Neural Network (ANN), Bayes Network.

This study focuses on determining the important variables that are significant in building a model for classifying good loans from bad loans. We used IBM SPSS Modeler data mining software in this study. Figure 1 shows the created model for analysis in this study:

Figure 1: Model with Boosting



Statistical feature selection in the SPSS Modeler is used to establish the most important variables that are correlated to the target variable STATUS (pair or good loans from bad or defaulted loans).

RESULTS

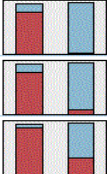
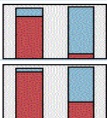

The results are shown in figure 2. The seven variables listed as statistically the most important are LOAN TYPE, AGE, EMPLOYMENT, LOAN AMOUNT, TERM OF LOAN, LOAN RATE and GENDER.

Figure 2: The most important variables determined by Feature 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 the not 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 3 indicate that Decision Tree (CHAID), Logistic Regression, and Support Vector Machine (SVM) are the most reliable models for this data set.

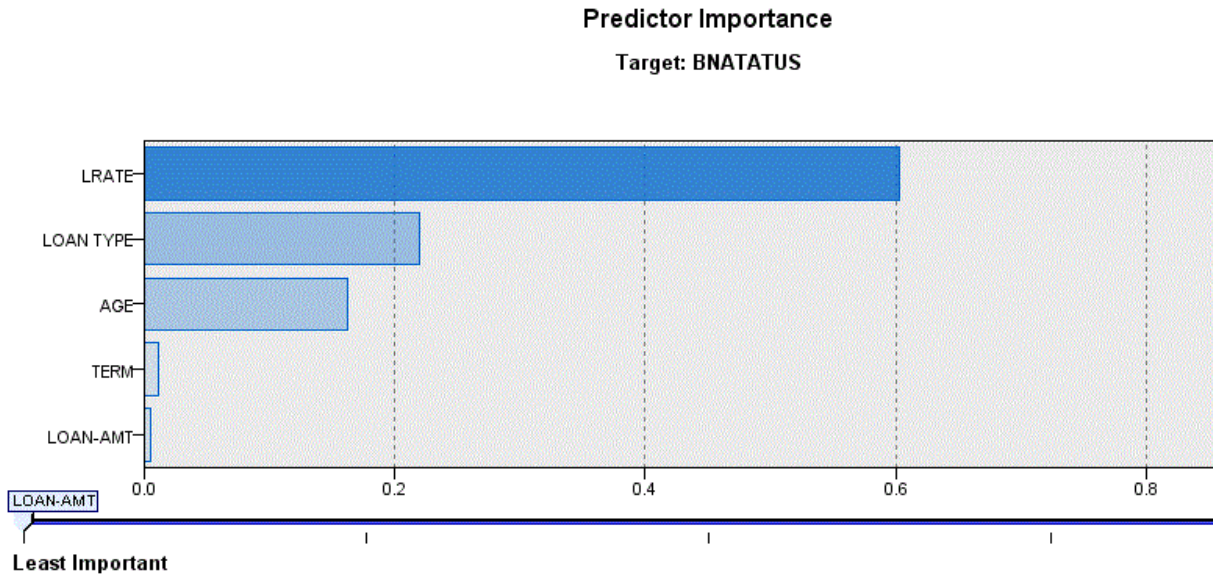
Figure 3: The most important algorithm found: CHAID, Logistic Regression and SVM

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

First, CHAID is created to study its results. Figure 4 list 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.

The Decision Tree model generated decision rules for deciding whether a loan would be a good loan or bad. However, due to space limitation it is not displayed in this paper. In the data set, 1 stands for a good loan and 2 for a defaulted loan. For example, considering 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. Figure 5 is a graphic representation of the generated rules.

Figure 4: The most important variables determined by CHAID



Confusion matrix (Figure 5) 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 5: CHAID Confusion Matrix

Comparing \$R-BNATATUS with BNATATUS

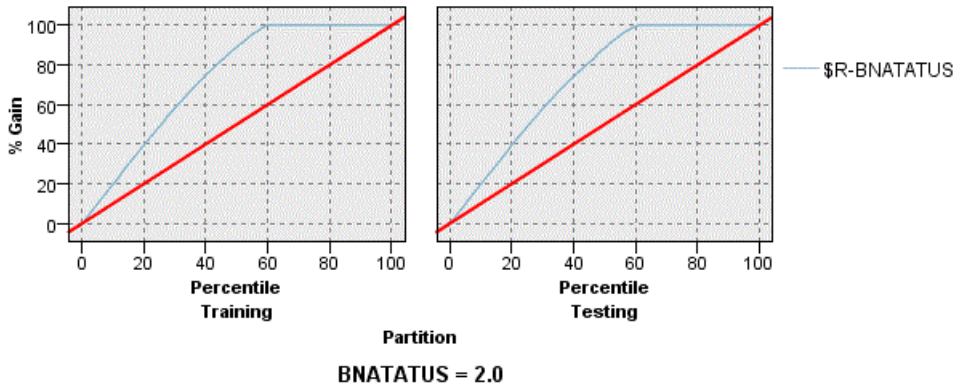
'Partition'	1_Training		2_Testing	
Correct	7,938	90.23%	3,479	90.72%
Wrong	860	9.77%	356	9.28%
Total	8,798		3,835	

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 6) 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 6: CHAID lift Curve



CONCLUSION

The sample data shows that the most important variables in determining default are in descending order, loan rate, loan type, age, the term of the loan and the loan amount. The higher the loan rate, the greater the age of the borrower, the shorter the term of the loan and the smaller the loan amount are all factors that seem to increase the probability of default. This is similar to some of the findings by prior studies that found that interest rates, age and term of the loan had an influence on default rates.

REFERENCES Available upon request.