Improving Personal Credit Scoring with HLVQ-C

A. S. Vieira[1], João Duarte[1], B. Ribeiro[2] and J. C. Neves[3]

[1] ISEP, Rua de S. Tomé, 4200 Porto, Portugal
[2]Department of Informatics Engineering, University of Coimbra, P-3030-290 Coimbra, Portugal
[3]ISEG - School of Economics, Rua Miguel Lupi 20, 1249-078 Lisboa, Portugal
{asv@isep.ipp.pt, bribeiro@dei.uc.pt, jcneves@iseg.utl.pt}

Abstract. In this paper we study personal credit scoring using several machine learning algorithms: Multilayer Perceptron, Logistic Regression, Support Vector Machines, AddaboostM1 and Hidden Layer Learning Vector Quantization. The scoring models were tested on a large dataset from a Portuguese bank. Results are benchmarked against traditional methods under consideration for commercial applications. A measure of the usefulness of a scoring model is presented and we show that HLVQ-C is the most accurate model.

Keywords: Credit Scoring, Neural Networks, Classification, Hidden Layer Learning Vector Quantization.

1 Introduction

Quantitative credit scoring models have been developed for the credit granting decision in order to classify applications as ‘good’ or `bad’, the latest being loosely defined as a group with a high likelihood of defaulting on the financial obligation.

It is very important to have accurate models to identify bad performers. Even a small fraction increase in credit scoring accuracy is important. Linear discriminant analysis still is the model traditionally used for credit scoring. However, with the growth of the credit industry and the large loan portfolios under management, more accurate credit scoring models are being actively investigated [1]. This effort is mainly oriented towards nonparametric statistical methods, classification trees, and neural network technology for credit scoring applications [1-5].

The purpose of this work is to investigate the accuracy of several machine learning models for the credit scoring applications and to benchmark their performance against the models currently under investigation.

The credit industry has experienced a rapid growth with significant increases in instalment credit, single-family mortgages, auto-financing, and credit card debt. Credit scoring models, i.e, rating of the client ability to pay the loans, are widely used by the financial industry to improve cashflow and credit collections. The advantages of credit scoring include reducing the cost of credit analysis, enabling faster credit decisions, closer monitoring of existing accounts, and prioritizing collections [4].

Personal credit scoring is used by banks for approval of home loans, to set credit limits on credit cards and for other personal expenses. However, with the growth in financial services there have been mounting losses from delinquent loans. For instance, the recent crises in the financial system triggered by sub-prime mortgages have caused losses of several billion dollars.

In response, many organizations in the credit industry are developing new models to support the personal credit decision. The objective of these new credit scoring models is increasing accuracy, which means more creditworthy applicants are granted credit thereby increasing profits; non-creditworthy applicants are denied credit thus decreasing losses.

The main research focuses on two areas: prediction of firm insolvency and prediction of individual credit risk. However, due to the proprietary nature of credit scoring, there is few research reporting the performance of commercial credit scoring applications.

Salchenberger et al. investigate the use of a multilayer perceptron neural network to predict the financial health of savings and loans [6]. The authors compare a multilayer perceptron neural network with a logistic regression model for a data set of 3429 S&L's from January 1986 to December 1987. They find that the neural network model performs as well as or better than the logistic regression model for each data set examined.

The use of decision trees and multilayer perceptrons neural network for personal credit scoring were studied by several authors. West tested several neural networks architectures on two personal credit datasets, German and Australian. Results indicates that multilayer perceptron neural network and the decision tree model both have a comparable level of accuracy while being only marginally superior to tradition parametric methods [7].

Jensen [5] develops a multilayer perceptron neural network for credit scoring with three outcomes: obligation charged of (11.2%), obligation delinquent (9.6%), and obligation paid-of. Jensen reports a correct classification result of 76 - 80% with a false positive rate (bad credit risk classified as good credit) of 16% and a false negative rate (good credit risk classified as bad credit) of 4%. Jensen concludes that the neural network has potential for credit scoring applications, but its results were obtained on only 50 examples.

The research available on predicting financial distress, whether conducted at the firm or individual level suggests that recent non-parametric models show potential yet lack an overwhelming advantage over classical statistical techniques. Recently we have successfully applied new data mining models like Hidden Layer Learning Vector Quantization (HLVQ-C) [8] and Support Vector Machines (SVM) [9] for bankruptcy prediction where they clear outperformed linear methods. However, the major drawback for using these models is that they are difficult to understand and the decisions cannot be explicitly discriminated.

This paper is organized as follows. Section 2 discusses the dataset used, the pre-processing of the data and feature selection. Section 3 presents the models and the usefulness measure. In Section 4 the results are discussed and finally section 5 presents the conclusions.

2 Dataset

The database contains about 400 000 entries of costumers who have solicited a personal credit to the bank. The valued solicited ranges from 5 to 40 kEuros and the payment period varies between 12 to 72 months.

Table 1 presents the definitions of the eighteen attributes used by the bank. Eight of these attributes are categorical (1, 2, 3, 4, 5, 8, 9 and 10) and the remaining continuous. Most of the entries in the database have missing values for several attributes. To create a useful training set we select only entries without missing values.

The database also contains the number of days that each client is in default to the bank concerning the payment of the monthly mortgage – in most cases this number is zero. We consider a client with bad credit when this number is greater than 30 days. We found 953 examples in the database within this category. To create a balanced dataset an equal number of randomly selected non-default examples were selected, reaching a total of 1906 training cases. We call this dataset 1.

We also created a second dataset where the definition of bad credit was set to 45 days of delay. This dataset is therefore more unbalanced containing 18% of defaults and 82% non-defaults. This is called dataset 2.

Table 1: Attributes used for credit scoring. Marked bold are the selected attributes.

#	**Designation**	#	**Designation**
1	Professional activity	10	Nationality
2	Previous professional activity	**11**	Debt capacity
3	Zip code	12	Annual costs
4	Zip code – first two digits	13	Total income
5	Marital status	14	Other income
6	Age	15	Effort ratio
7	Number of dependents	16	Future effort ratio
8	Have home phone	**17**	Number of instalments
9	Residential type	**18**	Loan solicited

2.1 Feature selection

Several feature selection algorithms were used to exclude useless attributes and reduce the complexity of the classifier. Due to the presence of many categorical attributes, feature selection is difficult. Several methods were used to test the consistency of the selection: SVM Attribute Evaluation, Chisquared and GainRatio. Each method selected slightly different sets of attributes. We choose the following set of six attributes with the highest consensus among all rankers: 1, 3, 4, 11, 17 and 18.

3. Models used

The data was analysed with five machine learning algorithms: Logistic, Multilayer Perceptron (MLP), Support Vector Machine (SVM), AdaBoostM1 and Hidden Layer Learning Vector Quantization (HLVQ-C).

For MLP, we used a neural network with a single hidden layer with 4 neurons. The learning rate was set to 0.3 and the momentum to 0.2. The SVM algorithm used was the LibSVM [12] library with a radial basis function as kernel with the cost parameter C = 1 and the shrinking heuristic. For AdaBoostM1 algorithm we used a Decision Stump as weak-learner and set the number of iterations to 100. No resampling was used. The HLVQ-C algorithm implementation is described elsewhere [8].

3.1 Usefulness of a classifier

Accuracy is a good indicator, but not the only criteria, to choose the apropriate classifier. We introduce a measure of the usefulness of a classifier, defined by:

$$\eta = E - L\,,$$

where E is the earnings obtained by the use of the classifier and L the losses incurred due to the inevitable misclassifications.

Earning, for the bank point of view, results from refusing credit to defaults clients, and can be expressed as:

$$E = NV(1-e_I)x$$

where N is the number of loans applicants, V the average value of a loan, e_I the type I error and x the typical percentage of defaults in the real sample. For simplicity we are assuming a Loss Given Default (LGD) of 100%.

Losses results from excluding clients that were incorrectly classified as defaults. In a simplified way they can be calculated as:

$$L = mNV(1-x)e_{II}$$

where m is the average margin typically obtained by the bank in a loan. The net gain in using a classifier is:

$$\eta = NV\left[x(1-e_I) - (1-x)e_{II}m\right].$$

To have $\eta > 0$ we need:

$$\frac{x}{1-x} > mG\,,$$

where $G = \frac{e_{II}}{1-e_I}$, is a measure of the efficiency of the classifier. This quantity should be the lowest possible. Assuming *x* small and e_I = 0.5 , we should have $x > 2me_{II}$.

4. Results

In table 2 we compare the efficiency of the classifiers on two datasets using 10-fold cross validation. For dataset 1, most classifiers achieve a good accuracy in detecting defaults but at the cost of large type II errors. Since real data is highly unbalanced, most cases being non-defaults, this means that more than half of clients will be rejected. SVM is the most balanced classifier while HLVQ-C achieved the highest accuracy on both datasets.

Since dataset 2 is more unbalanced and the default definition more strict error type II decreased considerably while error type I increased. More important, the usefulness of the classifier, measured by *G*, improved substantially. The HLVQ-C is again the best performer, either on accuracy and usefulness, and AdaboostM1 the second best. Logistic is the worst performer.

Following our definition, for the classifier to be useful the dataset has to have about 6% defaults, considering the best model (HLVQ-C), and as much as 11% for the Logistic case (setting m = 0.5).

To increase the usefulness, i.e. lower *G*, error type II should decrease without deteriorating error type I. This can be done either by using a more unbalanced dataset or applying different weights for each class. The exact proportion of instances in each class in the dataset can be adjusted in order to minimize G.

Table 2. Accuracy, error types and usefulness of different models in the two datasets considered.

	Classifier	**Accuracy**	**Type I**	**Type II**	**G**
Dataset 1	Logistic	66.3	27.3	40.1	54.8
	MLP	67.5	8.1	57.1	61.1
	SVM	64.9	35.6	34.6	52.3
	AdaboostM1	69.0	12.6	49.4	55.7
	HLVQ-C	**72.6**	5.3	49.5	52.3
Dataset 2	Logistic	81.2	48.2	11.0	21.2
	MLP	82.3	57.4	9.1	20.1
	SVM	83.3	38.1	12.4	19.3
	AdaboostM1	84.1	45.7	8.0	14.7
	HLVQ-C	**86.5**	48.3	6.2	**11.9**

5 Conclusions

In this work we compared the efficiency of several machine learning algorithms for credit scoring. Feature selection was used to reduce the complexity and eliminate useless attributes. From the initial set of 18 features only 6 have been selected.

While MLP slightly improves the accuracy of Logistic regression, other methods show considerable gains. AdaboostM1 boosts its accuracy by 3% and HLVQ-C up to 5%.

The price to be paid for the accurate detection of defaults is a high rate of false positives. To circumvent this situation an unbalanced dataset was used with a more strict definition of default. A measure of the usefulness of the classifier was introduced and we showed that it improves considerably on this second dataset.

References

1. Brill J. The importance of credit scoring models in improving cashflow and collections. Business Credit **7**, 1 (1998**)**
2. Tam KY, Kiang MY. Managerial applications of neural networks: the case of bank failure predictions. Management Science **47**, 926 (1992).
3. Davis RH, Edelman DB, Gammerman AJ. Machine learning algorithms for credit-card applications. IMA Journal of Mathematics Applied in Business and Industry **51**, 43 (1992).
4. Desai VS, Crook JN, Overstreet GA. A comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research 37, 24 (1996).
5. Jensen HL. Using neural networks for credit scoring. Managerial Finance **26**, 18 (1992).
6. Salchenberger LM, Cinar EM, Lash NA. Neural networks: a new tool for predicting thrift failures. Decision Sciences **23**, 899 (1992).
7. West D, Neural Network credit scoring models, Computers & Operations Research **27**, 1131 (2000).
8. A. S. Vieira and J. C. Neves, “Improving Bankruptcy Prediction with Hidden Layer Learning Vector Quantization”, Eur. Account. Rev. **15** (2), 253-275. (2006)
9. B. Ribeiro, A. Vieira and J. C. Neves: Sparse Bayesian Models: Bankruptcy-Predictors of Choice?, Int. J. Conf. Neural Networks, Vancouver, Canada 3377-3381 (2006).