Data Mining for loan evaluation at ABN AMRO: a case study

A.J. Feelders and **A.J.F. le Loux**
University of Twente
Knowledge based systems group
PO Box 217, 7500 AE Enschede
The Netherlands
e-mail: feelders@cs.utwente.nl

J.W. van 't Zand
ABN AMRO bank
Operations Research Department
PO Box 669, 1000 EG Amsterdam
The Netherlands

Abstract

We describe a case study in data mining for personal loan evaluation, performed at the ABN AMRO bank in the Netherlands. Historical data of clients and their pay-back behaviour are used to learn to predict whether a client will default or not. It is shown that, due to the pre-selection by a credit scoring system, the data base is a sample from a different population than the bank is actually interested in; this necessarily restricts inference as well. Furthermore we point out the importance of integrity and consistency checking when the data are entered into the system: noise is a serious problem.

The actual experimental comparison involves a "classical" statistical method, linear discriminant analysis, and the classification tree algorithm C4.5. Both methods use one and the same training set, drawn from the historical database, to learn a classification function. The percentages of correct classifications on an independent test set are 71.4% and 73.6% respectively. McNemar's test shows that the null hypothesis of equal performance has a p-value of 0.1417.

The classification tree constructed by C4.5 uses 10 out of 38 attributes to distinguish between defaulters and non-defaulters, and is consistent with the available theory on credit scoring. The linear discriminant function uses 17 variables to make the classification. Both from the viewpoint of predictive accuracy and comprehensibilty, the classification tree performs better in this study.

To make furhter progress, the level of noise in the data has to be reduced, and data has to be collected on loans that are rejected by the credit scoring system.

Introduction

Banks and other financial institutions are among the most information intensive organizations in the business community. They possess enormous amounts of data on customers and transactions, and if properly analysed these data bases may yield valuable knowledge for future decision making.

In this paper we describe a case study that was performed at the ABN AMRO bank in the Netherlands. This case study involves the application of data mining to the evaluation of personal loans. Data mining is performed with the well-known classification tree algorithm C4.5. In order to assess predictive performance and ease of interpretation, we compare the results of the classification tree to those obtained with linear discriminant analysis. We selected linear discriminant analysis, because it is a popular statistical classification technique that is often applied to credit scoring problems.

First we describe the basics of personal loan evaluation, as well as the credit-scoring system that ABN AMRO currently uses for this purpose. Then we specify the scope of this study, and note the problem of selection in data bases. Subsequently, we describe the data base that is used in this study, and give a short description of linear discriminant analysis and classification trees. After that, we discuss the experimantal results of both methods on the credit data. Finally we draw a number of conclusions from this case study, and indicate a number of possibilities for future research.

The current credit scoring system

Personal loans are supplied by credit institutions to private persons, the credit takers. The credit institution supplies a certain amount of money to the credit taker, who is then obliged to pay a monthly installment during a period of time.

The credit institution will not give money to everybody who asks for it. An assessment has to be made of whether the loan will be payed back properly. The financial data of the client are used to evaluate whether his financial position is adequate to get a loan, and if so, what the maximum amount of the loan may be.

Subsequently, ABN AMRO uses a quantitative credit evaluation system, also called a *credit scoring* system. A credit scoring system awards points to particular relevant attributes of the credit taker. The scores on the relavant attributes are added, and the total score is used as an indication of how likely the credit taker will pay back the loan properly. Table 1 shows which attributes determine the score of an individual.

If the score is below a certain bound, the credit is not granted; otherwise the credit is granted. If a credit has been granted, data concerning the number of arrears

Attribute	Value	
Age	< 22	–
	22–24	–
	≥ 45	+
Telephone	not reported	–
Owns house?	yes	+
Employment	self-employed/part-time	–
	unemployed/not reported	–
Length of employment	more than 15 years	+
Mode of payment	not automatic collection	–
Client of ABN AMRO?	no	–
Number of BKR-lines	4 or more	–
Number of current personal loans	1	–
	2 or more	–
Number of current own contracts	1 or more	+
Number of arrears (BKR)	1 or more	–
Number of current arrears	1 or more	–
Number of coded arrears	1	–
	2 or more	–

Table 1: Attributes and their influence on the credit score

and duns of a client are recorded in the data base. A client is marked as a defaulter if there has been an arrear of 60 days or more. This criterion is used because the bank is legally obliged to report such arrears to the Dutch Credit Registration Bureau (BKR).

Data base analysis shows that the percentage of defaulters decreases as the credit score increases.

Databases and selection

We're interested in the population of people who apply for a loan at ABN AMRO bank. We would like to know of these persons whether they will become defaulters or not. In order to obtain reliable information on this issue, one could draw a random sample from the population, where every element has a non-zero chance of being in the sample. The data base of ABN AMRO clearly does not fullfil these requirements. There has been a systematic selection: the data base only contains data on persons to whom a loan has been given.

People whose credit-score was not high enough had a zero chance of entering the sample. We have no way of knowing whether these people would have become defaulters or not. It is easy to see how the systematic selection may lead to unwarranted conclusions. Suppose that having the value "no" for binary attribute A has such a negative impact on the credit-score that it can not be compensated by favourable values for other attributes. In that case all accepted loans have the value "yes" for attribute A. This means that there is no association between A and whether or not someone defaults in the database. However, in the population of applicants there probably is a very strong association. Correlations can be influenced in more subtle ways by selection, see e.g. (Fleiss 1973; Cooper 1995).

This problem is of course not specific to the application we are discussing here. In general, the data present in data bases were not collected in order to answer the question we may have in mind. This may, among other things, mean that they come from a different population than the one we are actually interested in.

The current study is consequently restricted to learning to tell the difference between defaulters and non-defaulters, *within the population of people who would be accepted on the basis of their credit-score.*

The data base used

Apart from the personal and financial data that are supplied by the applicant, data about the credit history of the applicant are obtained from the Dutch Credit Registration Bureau (BKR), and recorded in the data base. Table 1 shows that this last group of data is rather important to the decision whether or not to grant the credit.

The quality of data in the data base is evidently of crucial importance to the ability to detect useful patterns. The data on which this study is based, were entered at ABN AMRO offices on several locations in the Netherlands. In the absence of clear guidelines, this geographical spread may contribute to inconsistencies and differences in interpretation. Study of the program used to enter the data revealed that rather simple consistency checks on the data entered into the data base are missing. As a consequence, some attributes have a rather large amount of unknown, or evidently incorrect values. To give one example, the attribute *norm amount for cost of living* is defined in such a way, that it minimally amounts to *ƒ*950, –. In 40% of the cases, however, a lower amount has been entered, and in 10% of the cases the value *zero* has been entered. It is very difficult to establish afterwards how these figures shoud be interpreted, for example does *zero* mean *unknown*, does it mean the same thing at every ABN AMRO office? We were not able to get unambiguous answers to these questions.

The data base at our disposal consisted of 52,569 records (loans), of which 1039 (±2%) were defaulters. We added three attributes to the data base, because we suspected - on the basis of theoretical considerations - that they would have considerable explanatory power. The attributes added can of course all be derived from those originally present in the data base. The attribute *Bufferspace* was added to indicate the excess of monthly income of the applicant over his expenses and monthly installment. Because this attribute is derived from a total of 10 other attributes, often one of these 10 did not have a value. Furthermore *norm amount for cost of living* is part of the definition of *Bufferspace*, and we already discussed the problems with the former attribute. As a less noisy alternative, we added

$$\frac{\text{Monthly installment}}{\text{Net monthly income}}$$

which indicates the proportion of monthly income that has to be spent by the credit taker on paying the installment. Finally, we also added an attribute to indicate whether

1. The applicant at some time in the past received a loan.
2. All loans received in the past had been payed back correctly.

If both conditions are fullfilled, the attribute has value *one*, otherwise *zero*. This last attribute could be discovered by a classification tree algorithm, so it only attemps to "shortcut" the learning process. We also removed six attributes from the data base, because they are not known when the loan is applied for; they record information on the execution of the loan, such as the number of duns a credit taker has received. All in all, each credit taker is described by 38 attributes.

Classification methods

In this section two methods for learning classification functions are described very briefly. One method, linear discriminant analysis, represents the "classical" statistical approach to classification. The second, classification trees, represents a group of flexible classification methods, that is typical for the data mining approach.

Linear Discriminant Analysis

Linear Discriminant Analysis is probably the most popular "classical" statistical method of classification. It is most easily viewed as a special case of the well-known Bayes' criterion. This criterion states that, assuming one wants to minimize the number of misclassifications, to assign an object to group C_j if

$$P(C_j)\,P(\mathbf{x} \mid C_j) > P(C_i)\,P(\mathbf{x} \mid C_i) \qquad \forall i \neq j$$

where $P(C_j)$ is the relative frequency of C_j in the population, and $P(\mathbf{x} \mid C_j)$ is the conditional probability of observing attribute vector $\mathbf{x}$, given that an object belongs to group C_j.

Although this rule is optimal, its straightforward application requires the estimation of many conditional probabilities. In order to reduce this problem, statisticians have introduced additional assumptions. If it is assumed that $\mathbf{x}$ follows a multivariate normal distribution in all groups, and all groups have identical covariance-matrices, then the above rule can be replaced by assigning to group C_j if

$$f_j(\mathbf{x}) + \ln(P(C_j)) > f_i(\mathbf{x}) + \ln(P(C_i)) \qquad \forall i \neq j$$

where

$$f_i(\mathbf{x}) = \mu_i^t \Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_i^t \Sigma^{-1}\mu_i$$

where Σ denotes the covariance-matrix common to the multivariate normal distribution of $\mathbf{x}$ in all groups. Note that $f_i(\mathbf{x})$ is a linear function.

If it is additionally assumed that objects are to be classified into one of two groups, the optimal rule is to classifiy into group 1 if

$$z(\mathbf{x}) > \ln(P(C_2)) - \ln(P(C_1))$$

and to group 2 otherwise, where $z(\mathbf{x}) = f_1(\mathbf{x}) - f_2(\mathbf{x})$. This is the form in which most discriminant functions are presented in the literature.

In practice the population parameters μ_i and Σ are unknown and have to be estimated from the sample. The discriminant function is estimated using the sample estimates $\bar{\mathbf{x}}_i$ and the pooled covariance matrix S^{pooled}. This is often called the "plug-in" estimate of the discriminant function.

Classification Trees

We already noted that linear discriminant analysis is based on a number of assumptions that may not always be realistic, for example when attributes are measured on a nominal or discrete ordinal scale. In the credit evaluation domain many such attributes exist. The attribute *Has Telephone?*, for example, has value zero when no private telephone number has been given by the applicant, and value *one* otherwise. The distribution is very skewed: about 80% of the cases belongs to the latter group. A normal distribution clearly does not yield a good approximation in this case.

A number of alternative methods have been developed, stimulated by increased computer power, which make no such a priori assumptions. Examples of such *non-parametric* methods are classification trees, neural networks and k-nearest neighbour.

The basic idea of classification trees is as follows. A tree is fitted to the training sample by "recursive partitioning", which means that the training sample is succesively split into increasingly homogeneous subsets, until the leaf nodes contain only cases from a single class, or some other reasonable stopping criterion applies. Well-known examples of classification tree algorithms are CART (Breiman *et al.* 1984), ID3 (Quinlan 1983), and its successor C4.5 (Quinlan 1993). Classification trees approximate the optimal Bayes classifier by making a partition of the sample space, and estimating the posterior probabilities (and hence the Bayes rule) within each cell of the partition by the relative frequencies of the classes of the training cases which fall within that cell (Ripley 1994).

A crucial decision in building a classification tree, is selecting the variable on which to make the next split. Since one strives towards homogeneous subsets, most algorithms employ some measure to indicate the "impurity" of a set of cases, i.e. the extent to which a node contains training cases from multiple classes. Such a measure should take its largest value when all classes are equally represented, and its smallest value when the node contains members of a single class. For example, CART employs the so-called Gini-index

$$i(t) = \sum_{j \neq k} P(j \mid t) P(k \mid t)$$

where $i(t)$ denotes the impurity at node t, and $P(j \mid t)$ denotes the proportion of cases from class j at node t. On the other hand, ID3 uses the *entropy* of a node for the same purpose

$$i(t) = -\sum_{j=1}^{n} P(j \mid t) \times \log_2(P(j \mid t))$$

where n is the number of classes at node t. The "optimal" split is the one for which

$$\sum_{k=1}^{m} P(t_k) i(t_k)$$

is minimal, where m is the number of subsets resulting from the split, and $P(t_k)$ is the proportion of cases in subset t_k. C4.5 has a slightly adjusted splitting criterion, called the *gain ratio*. The gain ratio has been introduced by Quinlan, to correct for the bias of the above criteria in favor of splits with many outcomes. For a description of the gain ratio criterion, we refer to (Quinlan 1993).

Another critical issue in learning classification trees and other flexible classification methods, is the prevention of overfitting on the training sample. To this end, one usually adopts a pruning strategy, whereby the tree is first grown to its full size, and then leaf nodes are merged back or "pruned" to produce a smaller tree. There are basically two approaches to pruning (Quinlan 1993), one which predicts the error rate of a tree and its subtrees from an independent test set. Although this approach will yield unbiased estimates of the error rates, it also forces one to exclude part of the available data from the training set. C4.5 uses a second approach, that estimates the error rates using the training set from which the tree was built. The basic idea is that the error rate estimated from the training data is somewhat increased (in a rather heuristic way) to correct for the known optimistic bias of in-sample error rate estimation. For a detailed description of the pruning mechanism of C4.5, we again refer the reader to (Quinlan 1993).

Experimental comparison

In order to perform an experimental comparison of linear discriminant analysis and classification trees, we extracted a training set and test set from the data base as follows. We randomly drew 700 cases from the set of 1039 defaulters, and put them in the training set. The remaining 339 defaulters were put into the test set. Subsequently we randomly drew 1000 and 450 cases respectively from the set of non-defaulters and put them in the training set and test set respectively. Thus the balance between the two classes has been somewhat restored; using the relative frequencies that occur in the population (2% defaulters) would hamper the learning process.

	$I_{c4.5}$	$C_{c4.5}$	Total
I_{lda}	150	76	226
C_{lda}	58	505	563
Total	208	581	789

Table 2: Incorrect and Correct classifications of lda and c4.5

Linear discriminant analysis was performed using the statistical package SPSS 6.0 for Windows; we used the Mahalonobis method for variable selection. Where necessary, variables were recoded in order for LDA to be able to deal with them optimally. The estimated discriminant function contains 17 variables, and classifies 71.4% of the test set correctly.

As a representative of the classification tree algorithms we used C4.5. The best classification result was obtained using a high amount of pruning; each leaf of the tree was forced to contain at least 30 cases. The tree classifies 73.6% of the test set correctly, and uses 10 of the 38 variables to do so.

To determine the p-value of the difference observed between the two classifiers, we should take into account the correlation between their predictions. Therefore, a test that is often used in paired-sample designs, namely McNemar's test, is appropriate in this situation. In (Feelders & Verkooijen 1995) an experimental design involving the comparison of more that two methods is discussed. Table 2 summarizes the result of using the linear discriminant function (LDA) and C4.5 on the training set to classify the observations in the test set. The cells $(C_{lda}, C_{c4.5})$ and $(I_{lda}, I_{c4.5})$ respectively contain the number of cases classified correctly and incorrectly by both the linear discriminant function and C4.5. Since we want to test

$$H_0 : \mathrm{MPE}_{lda} = \mathrm{MPE}_{c4.5}$$

against

$$H_a : \mathrm{MPE}_{lda} \neq \mathrm{MPE}_{c4.5},$$

only the cells $(I_{lda}, C_{c4.5})$ and $(C_{lda}, I_{4.5})$ of this table are of interest. Here MPE is defined as the proportion of misclassifications a classifier makes on the population of interest. When an observation falling in the $(I_{lda}, C_{c4.5})$ cell of this table is defined as a success, then the number of successes is binomially distributed with $n = (I_{lda}, C_{c4.5}) + (C_{lda}, I_{c4.5}) = 134$ and $p = 0.5$, under the null hypothesis. Obtaining 76 successes out of 134, has a p-value of 0.1417 under the null hypothesis, according to an exact binomial test. We delibirately report the p-value and let the reader decide whether the difference found should be considered significant. We think the use of conventional significance

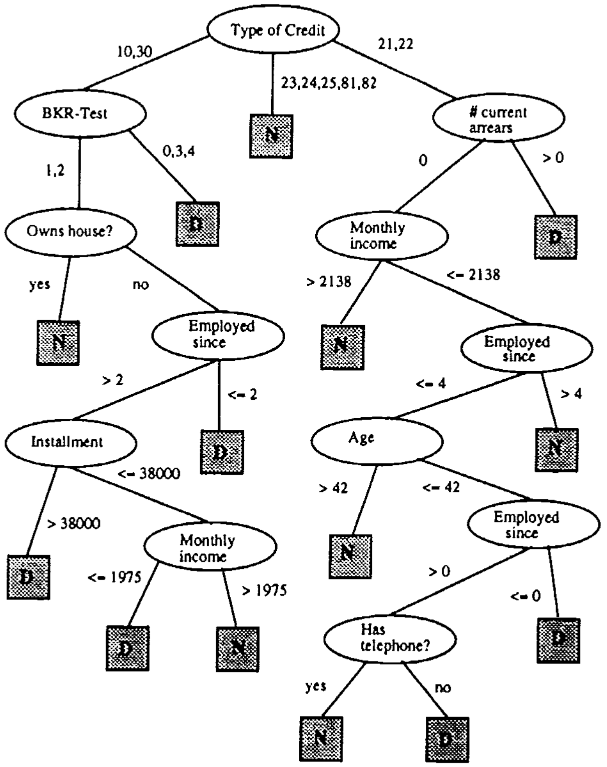

Figure 1: C4.5 classification tree (D = Defaulter)

levels is not appropriate in this situation, because it is not clear if making a Type I error is the most serious mistake.

The reason LDA is not doing that bad, despite the violation of normality assumptions, migth be that LDA as a classification rule does not always suffer from these violations. This is the case if the interaction structure of the attributes within the different classes is approximately equal (McLachlan 1992). Further study would have to show whether this is actually the case in our loan evaluation data base. Estimation of coefficients and posterior probabilities do become unreliable when normality assumptions are violated, whatever the interaction structure in the different classes (McLachlan 1992).

Figure 1 shows the tree that was generated using C4.5. It starts with a split on the type of credit. The types of credit (23,24,25,81,82) that immediately lead to the class Non-Defaulter, are all types of mortgage credit, where the collateral present gives a strong guarantee of payment. The attribute "employed since" (the number of years the credit taker works for his *current* employer) is used on three occasions in the tree. In all three cases, the split is consistent with the "theory" that the longer a person works for his current employer, the more likely he will pay back the loan correctly. All other splits that occur in the tree are consistent with earlier empirical findings, as well as the "theoretical conjectures" that can be found in the literature on credit scoring.

This tree should definitely not be used to classify people who apply for a loan at ABN AMRO. The attribute *Number of coded arrears* is not present in the tree, because applicants with coded arrears are "filtered out" by the credit scoring system. When someone has a coded arrear, it is virtually impossible to obtain a loan. Therefore the attribute does not have any discriminatory value in our database of accepted loans. In fact, the tree in figure 1 should only be used to classify applicants that have been accepted by the credit scoring system. It is clear that using the tree in addition to the credit scoring system always leads to a more restrictive policy, and thus less volume.

Conclusions and future research

Against our expextations, the predictive performance of linear discriminant analysis is not much worse than the performance of the classification tree algorithm C4.5, on the problem studied. From a *data mining* viewpoint we still prefer the classification tree, because it is more insightful, and also easier to adapt. The best classification tree uses only 10 out of 38 attributes to classify credit-takers into defaulters and non-defaulters; linear discriminant analysis uses 17 variables.

The best classification result on the test set, 73.6 %, appears to be rather low compared to other results reported in the literature. In these studies, however, one usually tries to predict the decision of an expert, namely accept or reject the loan. Because of the pre-selection of the credit scoring system, predicition becomes much harder. The defaulters that are "easy to recognize" have already been rejected by the credit scoring system.

Still, we think there are possibilities to increase the number of correct classifications. Firstly, the quality of the data can be improved by integrity and consistency checks when the data are entered into the data base. Secondly, one could consider to change the defaulter criterion. The current criterion, an arrear of 60 days or more, is determined by the legal obligation to report such arrears. As such, it is rather arbitrary, and from a financial viewpoint it is not the most appropriate criterion.

We already indicated that this research, due to the nature of the data available, has only a limited scope. We don't know whether a loan that was rejected by the credit scoring system would have been a good or bad loan. There are two ways in which this problem can be approached. Firstly, one can try to gather information about people that have been rejected, and try to make an "informed guess" whether they would have become defaulters on the basis of this information. For example, if the person rejected did get a loan at another financial institution and became a defaulter, then this information will be available at the Credit Registration Bureau. In that case we may classify the client as a defaulter. We think it's not advisable to take this approach; the Credit Registration Bureau is not allowed to provide such information in large amounts, and the information you might obtain on rejected loans is incomplete.

Secondly, one could randomly accept a certain percentage of the loans that are rejected by the credit scoring system. The chance that a "rejected" loan enters our data base is then known, and statisitical inference becomes possible again. This approach is undoubtedly more expensive, but yields far more reliable information.

Both approaches may be on the border of the definition of *data mining*. One purposefully gathers data to answer a specific question. In as far as data mining is restricted to analysing data that are generated as a product of the "normal" course of business, these approaches would be outside that scope.

References

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. T. 1984. *Classification and Regression Trees.* Belmont, California: Wadsworth.

Cooper, G. 1995. Causal discovery from data in the presence of selection bias. In Fisher, D., and Lenz, H., eds., *Proceedings of fifth international workshop on AI and statistics*, 140–150.

Feelders, A., and Verkooijen, W. 1995. Which method learns most from the data? In Fisher, D., and Lenz, H., eds., *Proceedings of fifth international workshop on AI and statistics*, 219–225.

Fleiss, J. L. 1973. *Statistical methods for rates and proportions.* New York: John Wiley & Sons.

McLachlan, G. J. 1992. *Discriminant analysis and statistical pattern recognition.* New York: Wiley.

Quinlan, J. 1983. Learning efficient classification procedures. In Michalski, R.; Carbonell, J.; and Mitchell, T., eds., *Machine learning: an artificial intelligence approach.* Palo Alto, CA: Tioga Press.

Quinlan, J. R. 1993. *C4.5 Programs for Machine Learning.* San Mateo, California: Morgan Kaufmann.

Ripley, B. D. 1994. Flexible non-linear approaches to classification. In Cherkassky, V.; Friedman, J.; and Wechsler, H., eds., *From Statistics to Neural Networks. Theory and Pattern Recognition Applications.* Springer-Verlag.