The use of knowledge discovery techniques for behavioural scoring

N. Meeus[1], J. Huysmans[1], B. Baesens[2], J. Vanthienen[1]
& M. Vandebroek[1]
[1]*Department of Applied Economic Sciences, K.U.Leuven, Belgium*
[2]*School of Management, University of Southampton, UK*

Abstract

This paper discusses the use of knowledge discovery techniques for a recent development in the field of scoring: behavioural scoring. The goal of behavioural scoring is to develop a model that predicts the creditworthiness of existing customers on the basis of their behaviour in the past. This paper explains briefly the Knowledge Discovery in Data process and applies the technique of logistic regression to real life datasets of a Belgian financial institution. It describes the development of scoring models for a cheque account, a credit account and the customer level and compares the model results for different pre-processing values and selection methods by means of the ROC curve, p-values and misclassification rates.
Keywords: behavioural scoring, ROC-analysis, credit scoring, KDD.

1 Introduction

Credit companies have always aimed at predicting as accurate as possible the creditworthiness of their (future) customers. In the past this was based on the personal judgement of the lender. Nowadays however credit-granting decisions are based on statistical or operational research methods.

The continuous search for better techniques led to the idea of applying Knowledge Discovery in Data (KDD) techniques for scoring. A financial institution has access to lots of customer data in which one can possibly find interesting patterns and correlations that can contribute to a better credit granting.

To indicate the prediction of customer creditworthiness one often uses the term ‘scoring’. Through the help of a scoring technique the customer is given a

score which indicates the customer's creditworthiness. We can distinguish three forms of scoring. The most commonly known type is **credit scoring**. This type of scoring deals with decisions concerning credit applications from new customers. This decision has to be taken at the moment of the application and the credit scoring will be mainly based on input data from the application form and the credit bureau. **Behavioural scoring** on the other hand deals with decisions concerning existing customers. What if an existing customer wants to increase his credit limit? What marketing if any should the firm aim at that customer? If the customer starts to fall behind in his repayments what actions should the firm take? Apart from the application form data, behavioural scoring uses data concerning the repayment and performance behaviour of the customer in the past. A third form of scoring no longer deals with minimising the default risk, as the previous two methods did, but focuses on maximising the total profit provided by the customer. This approach is called **profit scoring**.

2 Knowledge Discovery in Data

Fayyad et al. [1] give the following definition of KDD: "Knowledge Discovery in Data (KDD) is the non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data." KDD is an iterative process that can generally be split up in three phases: "Sampling and data pre-processing", "Data mining" and "Development of decision support systems". In the rest of this section, we will briefly discuss these three phases for a scoring application.

2.1 Sampling and data pre-processing

This first phase of the KDD-process consists in selecting the relevant data from the available datasets and cleaning this data to make it useful for data mining. This is a very important phase, as a good data pre-processing is crucial for the next steps in the KDD process; it forms the foundation on which the rest of the process is built.

First of all, there are several points which deserve careful consideration when selecting the data. The data that is being used for the scoring process should be as recent as possible in order to obtain data that is representative for the future goal customers. In the mean time, the time period should also be sufficiently long, to allow a clear distinction between the good and bad repayment behaviour; in other words it has to contain sufficient characteristics indicating the customer behaviour in the population. Furthermore, we have to decide about the amount of goods and bads in the sample. Is there a need for an equal proportion of goods and bads in the sample or should the sample represent the good/bad proportion of the total population? The difficulty with the second option is that not enough bad observations might be present to identify their specific characteristics.

Concerning the size of the sample suggests Lewis [2] that 1.500 goods and 1.500 bads are probably enough. Nevertheless in practice much larger samples are being used. Makuch [3] remarks that once you have 100.000 goods, it is not really necessary to add more information concerning the goods.

Selection of the input variables is also of crucial importance for the performance of the final scoring application. We can distinguish three types of input variables: those from the credit bureau, those of the application form and most important for behavioural scoring, those who describe the transaction history of the customer, such as there are: the mean, minimum and maximum balance of the account, the total value of the credit and debit transactions, the amount of times the credit card limit was exceeded the past month, 6 months or 12 months...

Secondly, we have to decide about the definition of good and bad. A commonly used definition of bad is the case in which a customer is three instalments in arrears. There are however several definitions possible for bad but Thomas et al. [4] indicate that even though the definitions for bad are quite different, the results might be very similar.

Thirdly, missing values for certain variables need to be dealt with. The larger the dataset, the higher the chance of missing values. Missing values can be a result of mistakes in gathering the data, incomplete answers from the customer, system and measurement mistakes… Depending on the data mining model being used the effects of these imperfect data can be trivial or dramatic. If we would develop a model based on decision trees, the missing values would not cause such a problem. The regression and neural network models in our analysis tool on the other hand, ignore observations containing missing values, which can lead to failure of discovering valuable information or to a biased sample. Various imputation procedures have been proposed to resolve the missing value problem.

Finally, we need to take into account that the data can contain extreme values. There exist two types of extreme values: correct extreme observations, like for instance an exceptional high salary of a top manager and incorrect extreme observations, like age=111111. It is important to make sure that the correct extreme observations are not removed from the dataset, because precisely the extremeness of the observation can form an important indication of the event that we are trying to predict.

2.2 Data mining

After the data pre-processing we can start with the data-mining phase. In this phase we try to extract knowledge from the data with the help of a learning algorithm. Depending on the problem, Baesens et al. [5] distinguish two forms of data mining: predictive and descriptive data mining.

There exists a wide range of algorithms to conduct data mining. In choosing an algorithm we should take into account several characteristics. The ideal model should be apart from accurate, also comprehensible and as simple as possible.

In describing data mining techniques for behavioural scoring, Thomas [6] splits the models into two approaches: those which seek to use credit scoring methods, but with the extra variables concerning the customer behaviour added, and those that build probability models of customer behaviour.

In section 3, the credit scoring approach is applied to three data sets obtained from a major Belgian financial institution and logistic regression is used to perform the scoring. In the rest of this section, we will briefly describe this

model. The logistic regression approach says that p, the probability of default, for a certain applicant i is related to the characteristics $x_1, x_2, ...x_m$ by

$$\log(\frac{p_i}{1-p_i}) = w_0 + w_1 x_{1i} + w_2 x_{2i} + ... + w_m x_{mi}$$

As $\frac{p_i}{1-p_i}$ takes values between 0 and ∞, $\log(\frac{p_i}{1-p_i})$ takes values between $-\infty$ and $+\infty$. Taking the exponential of each side, we obtain the following equation:

$$p_i = \frac{e^{(w_0 + w_1 x_{1i} + w_2 x_{2i} + ... + w_m x_{mi})}}{1 + e^{(w_0 + w_1 x_{1i} + w_2 x_{2i} + ... + w_m x_{mi})}} = \frac{1}{1 + e^{-(w_0 + w_1 x_{1i} + w_2 x_{2i} + ... + w_m x_{mi})}}$$

The parameters $w_0, w_1, ...w_m$ are typically determined by maximising the following likelihood function:

$$L = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

$$L = \prod_{i=1}^{n} \left(\frac{e^{(w_0 + w_1 x_{1i} + w_2 x_{2i} + ... + w_m x_{mi})}}{1 + e^{(w_0 + w_1 x_{1i} + w_2 x_{2i} + ... + w_m x_{mi})}} \right)^{y_i} \left(\frac{1}{1 + e^{(w_0 + w_1 x_{1i} + w_2 x_{2i} + ... + w_m x_{mi})}} \right)^{1-y_i}$$

Apart from logistic regression several other methods for behavioural scoring have been described in the literature. An overview of these approaches can be found in Baesens et al. [7].

2.3 Development of decision support systems

When the knowledge is obtained from the data, the final phase of the KDD process can begin. The final goal of the KDD process is to build a complete and easily implementable decision support system as to automate the scoring process. In this final phase, it is important to understand the system's motivation behind its classification decisions, as customers will request more information about denied applications.

3 Behavioural scoring: an application

3.1 Introduction

This section will describe the behavioural scoring analysis of three datasets obtained from a Belgian financial institution. As we have stated earlier in this

paper we will use Knowledge Discovery in Data techniques to conduct this analysis. As discussed in section 2, the KDD process consists of three main phases: pre-processing, data mining and the development of decision support systems. In this section we first briefly describe the analyzed datasets and then we will discuss in detail the application of the first two phases of the KDD process to the datasets. For our analysis we use the SAS Enterprise Miner tool based on the SEMMA KDD process model.

The available data consists of three datasets. The **Cheque Account** dataset contains data concerning the current account of the customer: the balance of the account, the number of days the limit of the account is exceeded, and the expenditures of the credit cards… The **Non Cheque Account** dataset deals with the credit level of the customer: the type of credit product, the way of payment, the outstanding balance, and the amount of debts in arrears… In the **Customer** dataset every customer entity from the Cheque and Non Cheque Account datasets is taken in. Apart from behavioural scoring data, like '*amount of months since the request for a credit product*', this dataset also contains credit-scoring data concerning the general features of the customer: his professional activity, expected income… We have chosen to analyse these three datasets separately, because this allows us to draw conclusions at each separate level: current account level, credit level and customer level.

To obtain an acceptable amount of bads in the sample there was chosen to take two months as an observation point: January and May 2002. The performance point was chosen 12 months later: January and May 2003. Any overlap between these two months was corrected by specific rules.

To be selected, the data had to meet the conditions shown in Table 1 at the observation point.

The variable Badindicator will obtain the value bad if the conditions in Table 2 are fulfilled.

Table 1: Conditions at the observation point.

Cheque Account	Non Cheque Account
• Performance Indicator <> 9 • Number of days exceeding the account in the previous month 2002 $\geq$ 15 • Exceeding amount > 125 euro • Number of days exceeding the account in the last 3 months < 62 • Excluding legal persons • Excluding CBC	• Performance Indicator <> 9 • One arrear in the previous month 2002 • Excluding legal persons • Excluding CBC

The Customer dataset takes over the score entities selected for the Cheque and Non Cheque Account datasets. An entity will be indicated as bad if is bad in the Cheque Account dataset or in the Non Cheque Account dataset or in both.

The pre-processing phase starts with the exploration of the available data. We identified for each dataset the observations and variables which had to be

removed from the dataset, because they were irrelevant for our analysis. Their irrelevancy was indicated during meetings with the financial institution or it was determined that the variable provides exactly the same information as other variables.

Table 2: Conditions badindicator.

Cheque Account	Non Cheque Account
• Satisfy the conditions at the observation point: see table 1 • At the performance point: number of days exceeding the account the last 3 months $\geq$ 89 days	• Satisfy the conditions at the observation point: see table 1 • At the performance point: number of arrears $\geq$ 3

Furthermore, we identified the variables which could not enter the analysis in their present form. These variables need additional processing. A common problem was the presence in the data of values like 99999, 99998 and 0 which represented special values. To be able to analyze this kind of data it is necessary to 'separate' these special values from the rest of the data. A transformation of numerical to categorical variables offers a solution to this problem. A separate category is created for the special values. For example, if a variable represents '*the number of months passed since the last limit modification of the account'* and a value of 99999 indicates that the account has no limit, we can divide the normal values into different categories and create an additional category for the special values.

We split up our data into a training and test dataset. The training dataset was used for the analysis and the development of our model. The test dataset allowed us to measure the performance of our model on data not used for the model development. We chose a simple random partition and split up the data into 75% training data and 25% test data. It was not necessary to create a validation dataset, as the logistic regression modelling is not sensitive to the problem of overfitting.

Table 3 provides an overview of the datasets after this first part of the pre-processing.

Table 3: Datasets overview after first part of pre-processing.

	Number of variables	Number of observations	Training set	Test set	Number of bad	% bad
Cheque Account	54	13.048	9.786	3.262	279	2%
Non Cheque Account	24	3.157	2.368	789	359	11%
Customer	65	16.077	12.058	4.019	641	4%

To eliminate the extreme values in the data, we applied a filter that eliminates the top and bottom X% of the observations. To prevent the elimination of extreme values which are correct observations, a detailed study in close cooperation with the members of the scoring team of the financial institution is needed.

To get an idea of the influence of the extreme value filtering on the performance of our model, we compared the ROC curves and misclassification rates for three filter levels: no filter, 0,5% and 3% top/bottom percentiles filter.

3.2 Data mining

Due to the large number of variables in the datasets; we applied selection methods for deciding which variables should be entered into the models. The output for the following selection methods was compared: backward, forward and stepwise selection.

In assessing the output of the developed models we compared the performance of different models. The models differed according to the applied selection method, the application or not of data transformation and the percentage of the extreme value filter. We compared the models on the basis of three performance criteria: the ROC curve, the misclassification rate and the p-values. We paid a great deal of attention to the ROC curve, as this is a very important performance criterion for our analysis and for the scoring team of the financial institution. We will describe for each dataset the main results.

The dataset **Cheque Account** allowed us to investigate the influence of the current account behaviour of the customer, on the creditworthiness of the customer in the future. The ROC curve taught us that the generated model classifies the score entities only moderate. On the basis of the p-values, misclassification rate and ROC curve, the model with data transformation, backward selection and 3% extreme value filter was selected as the best performing model.

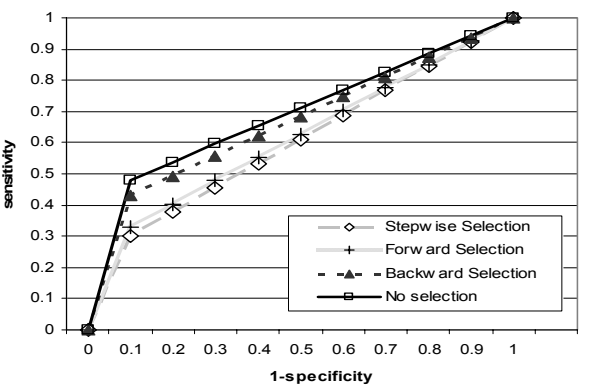

Figure 1: ROC Cheque Account: data transformation: 3% filter.

The dataset **Non Cheque Account** made it possible to investigate the influence of the credit account behaviour of the customer, on the creditworthiness of the customer in the future. The ROC curve showed us that the generated model classifies the score entities quite well. Observing the p-values, the misclassification rates and the ROC curves, we selected the model without data transformation, forward selection and 0.5% extreme value filtering as the best performing model. Applying a higher filter (like 3%) led to a clear deterioration of the ROC curves. It can be concluded that we obtained a good classification model which can contribute to delivering value information to the financial institution concerning the creditworthiness of their customers in the future.

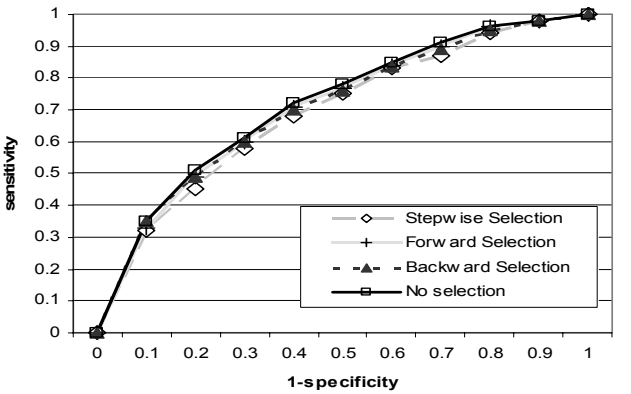

Figure 2: ROC Non Cheque Account: no data transformation: 0,5% filter.

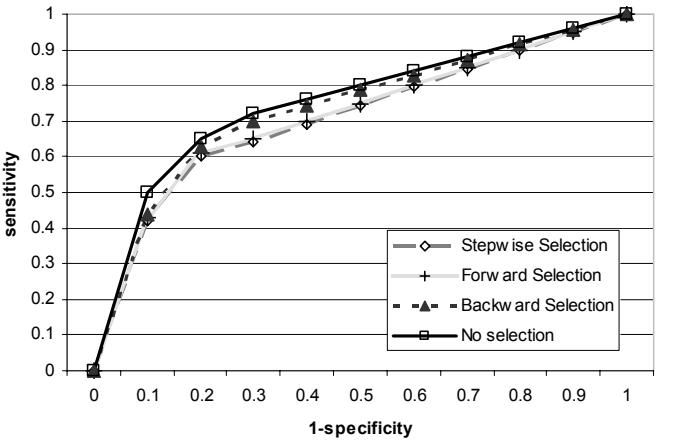

Figure 3: ROC Customer: no data transformation: 0,5% filter.

We investigated in the dataset **Customer** the influence of as well behavioural as credit scoring data concerning the characteristics of each score entity on the

creditworthiness of this entity in the future. The ROC curve showed us that the generated model classifies the score entities rather good. Based on the p-values, misclassification rates and ROC curves, we selected the model without data transformation, backward selection and 0,5% extreme value filter, as the best performing model.

We can conclude that this model is able to provide the financial institution with useful information concerning the creditworthiness of its customers in the future.

As stated before, we aim at obtaining a model which is not a black box model, but a model that is comprehensible by the members of the financial institution. A motivation for this is that such a model will be accepted faster and trusted more by the people who are going to apply the results.

In the same line of thinking we observed after examining the output results of the Cheque Account and Customer model, that the estimated coefficients often show a sign opposite to what one should expect on the basis of logic reasoning. One of the causes of this phenomenon appears to be after further research: multicollinearity. With the aim of comprehensibility in mind we can try to develop methods to reduce the multicollinearity, such as there are: the use of stepwise selection and the removal of correlated variables.

4 Conclusion

The search of financial institutions for methods to improve the ability of predicting the creditworthiness of their customers, the idea came forward to use KDD techniques on behavioural data of the existing customers. The purpose of the behavioural scoring was to discover patterns in the behaviour of existing customers which could give us an indication of the creditworthiness of these customers in the future.

During our analysis it was clear that the larger part of the time was occupied by the pre-processing phase. A good preparation and cleaning of the data before the analysis is a fundamental step which forms the basis of a successful model. We can conclude that in using logistic regression as a modelling technique, we reach acceptable predictive models for both the credit level as the customer level.

In the future it would be useful to conduct a thorough study of the extreme values in close cooperation with the financial institution. Furthermore it would be interesting to instead of logistic regression, use other classification methods, like neural networks and decision trees and compare the performance of these other models with that of logistic regression.

In the end, one could proceed to the final step of the KDD process. The ultimate goal of this process is to develop a decision support system which will allow the financial institution to use the obtained results in practice.

References

[1] Fayyad U.M., Piatetsky-Shapiro G., Smyth P. & Uthurusamy R. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press. 1996.

[2] Lewis E.M. An introduction to Credit Scoring. Athena Press, San Rafael, CA. 1992.
[3] Makuch W.M. The basics of a better application score. *Credit Risk Modeling, Design and Applications*. Fitzroy Dearborn Publishers. pp 59-80. 1998.
[4] Thomas L.C., Edelman D.B. & Crook J.N. *Credit Scoring and it's Applications.* SIAM Monographs on Mathematical Modeling and Computation, SIAM: Philadelphia. pp 1-117; 219-232. 2002.
[5] Baesens B., Mues C. & Vanthienen J. Knowledge Discovery in Data: van academische denkoefening naar bedrijfsrelevante praktijk. *Informatie,* pp. 30-35. 2003.
[6] Thomas L.C. A Survey of Credit and Behavioural Scoring; Forecasting financial risk of lending to consumers. Edinburgh. 2000.
[7] Baesens B., Van Gestel T., Viaene S., Stepanova M., Suykens J., Vanthienen J., Benchmarking State of the Art Classification Algorithms for Credit Scoring, Journal of the Operational Research Society, Volume 54, Number 6, pp. 627-635, 2003.