

# **The Impact of Sample Bias on Consumer Credit Scoring Performance and Profitability**

Geert Verstraeten, Dirk Van den Poel

Corresponding author: Dirk Van den Poel

*Department of Marketing, Ghent University, Hoveniersberg 24, 9000 Ghent, Belgium  
{geert.verstraeten,dirk.vandenpoel}@ugent.be*

# The Impact of Sample Bias on Consumer Credit Scoring Performance and Profitability

## Abstract

This article seeks to gain insight into the influence of sample bias in a consumer credit scoring model. Considering the vital implications on revenues and costs concerned with the issuing and repayment of commercial credit, predictive performance of the model is crucial, and sample bias has been suggested to pose a sizeable threat to profitability due to its implications on either population drainage or biased estimates. Whereas in previous research, different techniques of reducing sample bias have been proposed and deployed, the debate around the impact of sample bias itself has predominantly been held on a theoretical level. The dataset that was used in this study, however, provides the opportunity to investigate the issue in an empirical setting. Based on the data of a mail-order company offering short term consumer credit to their consumers, we show that (i) given a certain sample size, sample bias has a significant effect on consumer credit-scoring performance and profitability, (ii) its effect is composed of the inclusion of rejected orders in the scoring model, and the inclusion of these orders into the variable-selection process, (iii) the impact of the effect of sample bias on consumer credit scoring performance and profitability is limited and (iv) in consumer credit scoring, by merely increasing the sample size of the biased sample, the impact of sample bias can likely be reduced. Hence, we conclude that the possible impact of any reduction of sample bias is modest in a consumer credit scoring model, and that attention might optimally be focused on other factors leading to improved consumer credit scoring.

*Keywords:* Consumer Credit Scoring; Sample Bias; Reject Inference.

## 1. Introduction

In this paper, the term ‘credit scoring’ is used as a common denominator for the statistical methods used for classifying applicants for credit into ‘good’ and ‘bad’ risk classes. Using various predictive variables from application forms, external data suppliers and own company records, statistical models, in the industry often termed scorecards, are used to yield estimates of the probability of defaulting. Typically, an accept or reject decision is then taken by comparing the estimated probability of defaulting with a suitable threshold (see e.g. Hand and Henley, 1997).

Since the very beginning of credit-scoring techniques, the issue of sample bias has rapidly grown to become an intriguing topic in the credit-scoring domain (see e.g. Chandler and Coffman, 1977). The challenge lies in estimating the default probabilities for all future credit applicants using a model trained on a skewed sample of previously accepted applicants only. Indeed, for the historically rejected applications, we are unable to observe the outcome, being whether or not the applicant was able to refund his debt. ‘Reject inference’ (see e.g. Hand and Henley, 1994) comprises the set of procedures determined to decrease the bias that arises by building scoring models on accepted applicants only, e.g. by imputing the target variable for rejected cases. While the existing literature in the domain has mainly focused on describing and (to a lesser extent) testing different procedures of reject inference, in this paper, due to the special features of the data set used, we focus on the results that could be reached when perfect reject inference would occur. In this way, we hope to shed some new light upon the relevance of reject-inference procedures in a consumer credit scoring setting.

The remainder of this paper is structured as follows. Section 2 discusses the issue of sample bias and reject inference in the credit-scoring literature. Section 3 covers the methodology used in the empirical part of this paper to research the impact of sample bias on credit-scoring performance and profitability. Section 4 handles the data description of the data set used, and the sample composition needed for the empirical study. Section 5 reports the findings of the different research questions that are defined throughout the paper. Finally, conclusions and limitations and issues for further research are given in Sections 6 and 7 respectively.

## 2. Sample bias in the credit-scoring literature

Considering the widespread use of the statistical scoring techniques in the consumer credit industry, and considering the longevity of consumer credit scoring research (see, e.g. Myers and Forgy, 1963 for an early application<sup>1</sup>), the literature surrounding customer credit scoring has been growing steadily (for an overview, see, e.g. Thomas, 2000). In this study, we focus on application scoring (see, e.g. Hand, 2001 for an overview), i.e., we consider the decision whether or not to grant credit to potential lenders upon application, in contrast to the more recently introduced behavioral scoring (see, e.g. Thomas et al, 2001) where the performance of the customer is assessed for decision-making purposes during the lifetime of the relevant credit (e.g. whether the credit limit of a current borrower should be increased). Hence, we focus on the core application within the domain.

An important and fascinating topic that has provided much debate in credit scoring concerns the issue of sample bias: if new credit-scoring models are to be built on previous company records, only the previously accepted orders can be used for building the new credit score. Hence, sample bias arises when the sample of orders used for model building is not representative of the ‘through-the-door’ applicant population (Chandler and Coffman, 1977). Indeed, as mentioned by Hand and Henley (1994), the reject region has been so designated precisely because it differs in a non-trivial way from the accept region. Historically, sample bias has been accused of introducing at least one of two major shortcomings into the models, namely population drainage or biased estimates. Technically, the new score should only be applied to the customers that were accepted in the past (see, e.g. Joanes, 1994), and those rejected should remain rejected, leading inevitably to a decreasing customer base. Lacking an appropriate term for this in credit scoring, in this paper, we use the term ‘population drainage’ to cover this phenomenon. Alternatively, should the score be applied to all future orders, biased estimates would result for those credit applicants that would be rejected by the previous credit score (see, e.g. Hand and Henley, 1994). Understandably, both consequences have been proposed to have a negative influence on a company’s profitability.

Considering firstly the intriguing problem at hand, and secondly initial reports of the sizeable influence of sample bias when using discriminant analysis (Avery, 1981), previous research has historically focused on imputation techniques whereby one attempts to predict

---

<sup>1</sup> In this study, previous studies dealing with the development of a ‘numerical rating system’ are cited that date back to the 1940s.

the (unobserved) outcome of the previously rejected orders, henceforth called 'reject inference'. In fact, research on possible methods of reject inference date back almost as far as the beginning of credit scoring itself (see, e.g. Hsia, 1978).

While overviews of such methods are widely available (see, e.g. Joanes, 1994 and Hand and Henley, 1997), it might be useful to cover briefly the ideas behind two of the most renown reject inference techniques here. They include (i) augmentation, where the accepted orders are weighed inversely proportional to the probability with which the orders were accepted, in order to increase the impact of orders that are comparable to those orders that were rejected, and (ii) iterative reclassification, where rejected orders are scored and discretized using the classification rule derived from accepted orders, and where the model is re-estimated. After dividing the data again into samples of the same size as the original accept and reject regions, this procedure is iteratively repeated until convergence occurs (see e.g. Joanes, 1994). The results of these various attempts of reject inference, however, seem risky at best, leading some authors to conclude that reliable reject inference is impossible (Hand and Henley, 1994).

In contrast to studies on reject inference, the possible impact of sample bias itself on customer credit scoring has been covered to a much lesser extent. Additionally, it has been argued recently that the use of discriminant analysis itself introduces bias (*infra*), whereby the previous findings concerning the impact of sample bias (in e.g. Avery, 1981) should be reconsidered. While the shortage of a random sample of rejected orders is mentioned frequently, the costs involved with gaining such data are very often mentioned in the same breath (see, e.g. Hand and Henley, 1994 and 1997). Due to the fact that the data set used in this study contains real outcome values for a sizeable group of orders rejected by the statistical scoring process (*infra*), we are able to assess the importance of sample bias itself, as an upper limit of the benefits that could result from using models of reject inference. Hence, it lies not within the ambition of this study to test the performance of each of the proposed reject inference methods, but given the fact that reject inference deals with attempting to infer the true creditworthiness status of rejected applicants (Hand and Henley, 1994), it would seem beneficial to estimate the maximal improvement that could be reached when the imputation method is 100 % correct.

### 3. Methodology

#### 3.1 General outlay of this study

As mentioned earlier, a special feature of the data set used, exists in the fact that we have available the real outcome for a sizeable set of customers that were rejected by the scoring process. This sample of borrowers is henceforth referred to as the 'calibration' set. While the details of the data set will form the main topic of Section 4 of this paper, in this section we will describe how this calibration set will be used in the current study. First and foremost, it should be clear that the advantages of having such a set are twofold: while it allows the researcher to include the orders into the model-building process, it is equally valuable that these orders can be used in constructing the holdout sample. In this way, it will be possible to mimic the behavior of the score on a set of orders that is proportional to and hence more representative for the 'through-the-door' applicant population discussed earlier.

This study will roughly focus on three parts. Using the calibration set, we will first attempt to acknowledge the problems resulting from sample bias. Hence, in our first research question (henceforth called Q1), we will investigate whether sample bias occurs by (i) testing the performance of a classifier built on the orders that were accepted by the score yet applied on the calibration sample only, and (ii), using an extensive variable-selection procedure proposed by Furnival and Wilson (1974), we will investigate whether different characteristics would prevail when the calibration sample is included into the variable-selection process. Indeed, in his study, Joanes (1994) indicated the common practice of using a variable-selection procedure for detecting a small - yet effective - subset of the total list of potential variables, and uttered that a model derived from previously accepted applicants only may fail to take into account all the relevant risk characteristics.

More crucial to this study, in our second research question (Q2), we will attempt to estimate the gain in performance if the outcome of the rejected orders would be available. In this step, we will compare the performance of a model with sample size  $n$ , only containing previously accepted orders, versus a model with an equal sample size  $n$  containing a sample of accepted and rejected orders that is proportional to the ratio of accepted and rejected orders in the applicant population - henceforth referred to as 'proportionality'. In this effort, we will clearly distinguish between the benefits of proportionality through the use of different orders in the training data set, and the benefit arising through the application of the variable-selection procedure on a proportional sample. It should be clear that all models will be tested on a proportional holdout sample. To conclude, in this subsection we will test

whether mimicking proportionality increases predictive performance and profitability, given a certain sample size.

While sample size was held constant in the analysis described above, in our last research question (Q3), we will evaluate the impact of a variation in sample size on credit scoring performance and profitability. In a consumer credit scoring setting, it is likely that the sample size of the accepted orders is larger than the sample size of the proportional sampling. Indeed, if one considers the acceptance rate as  $\alpha$ , and the percentage of rejected orders for which the outcome is available as  $\sigma$ , then the set of accepted orders is larger whenever  $\alpha > \alpha.\sigma + (1 - \alpha).\sigma$ , hence when  $\alpha > \sigma$ . Since the acceptance rate is often rather high in consumer credit scoring (Hand and Henley, 1997), this situation is considered as likely. Hence, since an increase in sample size can be expected to have a non-negative influence on a model's performance, it is plausible that the possible improvement stemming from proportionality can be reduced whenever the sample size of the set of accepted orders can be augmented. To conclude, here we will make the trade-off between proportionality and sample size.

While the general purpose of this study was described here, some methodological decisions were made, resulting in the choice of an iterative resampling procedure using logistic regression analysis, monitored by three different performance indicators. Additionally, we will perform a sensitivity analysis to test the robustness of our findings. The reasons behind these decisions are presented below.

### 3.2 Credit-scoring technique

Recent research in credit scoring has been focused on comparing the performance of different credit-scoring techniques, such as neural networks, decision trees, k-nearest neighbour, support vector machines, discriminant analysis, survival analysis and logistic regression (see, e.g. Baesens et al, 2003, Stepanova and Thomas, 2002, Desai et al, 1996, Davis et al, 1992) . The main conclusions from these efforts are that the different techniques often reach comparable performance levels, whereby traditional statistical methods, such logistic regression perform very well for credit scoring. Hence, in this paper, we will use the latter method for modeling credit risk. Two other reasons confirm this choice: firstly, several authors consider logistic regression to be one of the main stalwarts of today's scorecard builders (see, e.g. Thomas, 2000, Hand and Henley, 1997), and secondly, discriminant analysis, being another technique that has extensively been used in credit analysis

(Rosenberg and Gleit, 1994), has been proven to introduce bias when used for extrapolation beyond the accept region (Hand and Henley, 1994, Feelders, 2000, Eisenbeis, 1977).

Technically, we can represent logistic regression analysis (also called logit analysis) as a regression technique where the dependent variable is a latent variable, and only a dummy variable  $y_i$  can be observed (Maddala, 1992):

$$y_i = \begin{cases} 1 & \text{if the borrower defaults} \\ 0 & \text{if the borrower is able to refund his debt} \end{cases}$$

The parameters ( $b_0 \dots b_k$ ) of the  $k$  predictive characteristics used, are then typically estimated using the maximum-likelihood procedure, and the default probability can be expressed as follows:

$$P(y=1 | X) = P_1(X) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + \dots + b_k X_k)}}$$

### 3.3 Performance Measurement

In agreement with recent studies where performance measures for classification are crucial (see, e.g. Baesens et al, 2003), we will not rely on a single performance indicator in reporting the results of our research. The performance measures used are: (i) classification accuracy, or the percentage of cases correctly classified (PCC), (ii) area under the receiver operating characteristics curve (AUC), and (iii) profitability of the new score. While both the first and the last measure reports a result based on a fixed threshold, the receiver operating characteristic curve illustrates the behavior of a classifier without regard to one specific threshold, so it effectively decouples classification performance from this factor (see e.g. Egan, 1975 for more details, or Hand and Henley, 1997 for an overview of related performance assessment tools in the credit-scoring domain). An intuitive interpretation of the AUC is that it provides an estimate of the probability that a randomly chosen defaulter is correctly rated (i.e. ranked) higher than a randomly selected non-defaulter. Thus, the performance measure is calculated on the total ranking instead of a discrete version of it, so it is clearly independent of any threshold applied ex-post. Note that this probability equals 0.5 when a random ranking is used. Both PCC and AUC have proven their value in related domains for binary classification, such as e.g. direct-mail targeting (see, e.g. Baesens et al, 2002). Additionally, since it was feasible to trace back all revenues and costs to the individual



orders, and since profitability is by definition the critical performance measure in a business context, we included it as the third credit-scoring performance measure used in this study.

### 3.4 Resampling procedure

Throughout the different empirical analyses of this study, a resampling procedure was used to assess the variance of the performance indicators. Considering a low proportion of defaulters in the data set used, in this study we will draw samples of  $n$  points without replacement from the  $n$  points in the original data set, allocating an equal amount of defaulters to training as to holdout samples. Hence, we will use a stratified resampling procedure. This repartitioning of the data will be performed 100 times, and the differences between the different models will be computed within every iteration (i.e. paired comparisons).

### 3.5 Sensitivity analysis

One could argue that the value of reject inference is driven by the extent of truncation, being the relative size of the reject region. In order to ensure the validity of our results, using a sensitivity analysis, we will treat previously marginally accepted orders (the orders having a probability of defaulting very close to, but still below the threshold) as rejected orders, to the extent that 70 % of the total applicants for credit are considered as historically accepted, considering Hand and Henley's (1997, p.526) expertise of this being a normal acceptance rate in mail-order consumer credit. Hence, given historical scores of orders, we can pretend that some orders were rejected, while they were actually accepted, so we are able to include them in our calibration sample.

## **4. Data Description and Sample Composition**

### 4.1 Data description

For our research, we used data of a large Belgian direct mail order company offering consumer credit to its customers. Its catalogue offers articles in categories as diverse as furniture, electro, gardening and DIY equipment and jewelry. We performed the modeling at a moment when the former credit score - constructed by an international company specialized in consumer credit scoring - was about to be updated since it had been in use for 6 years.

For modeling purposes, we will use data of all short-term credit orders placed between July 1<sup>st</sup> 2000 and February 1<sup>st</sup> 2002, and their credit repayment information until

February 1<sup>st</sup> 2003. Within this period, all the credits observed had to be refunded, so it was possible to indicate good versus bad credit repayment within 12 months of follow-up.

In the remainder of this section, we try to clarify in which way the data that was used in this study contains several advantages compared to data used in previous research. In order to do this, however, it is crucial to describe the company's order-handling process in detail. We attempt to do so below.

The ordering process at the focal company is bipartite. Firstly, orders are always scored by an automatic scoring procedure, previously called the former credit score. However, in addition to this procedure, an independent manual selection procedure (also called a 'judgmental' procedure) is used for orders with specific characteristics, hence selecting a rather large set of orders that were handled manually, regardless of their score. Therefore, since the scores of all orders were tracked, it is possible to assign each order exclusively to one of the six possible order routes, as given in Table 1. Manual acceptance overrules, yet is not always applied. Hence, we can ex-post define six possible order flows for the orders that were handled.

		Judgmental Method		
		Not handled	Accepted	Rejected
Scoring Method	Accepted	<b>A1</b> 32503 obs 528 defaults (1.62 %)	<b>A2</b> 3536 obs 101 defaults (2.86 %)	<b>R3</b> 2844 obs
	Rejected	<b>R1</b> 234 obs	<b>A3</b> 2009 obs 107 defaults (5.33 %)	<b>R2</b> 3228 obs

Table 1: Order flow frequencies

In order to give a clear overview of the relevance of each of the groups for this study, in Figure 1 we have indicated the positioning of the different groups according to the existing credit score. Note that we have bracketed the traditional terminology concerning overrides, because 'high' and 'low' are conditional on the coding of the dependent variable. In this study, as noted before, we have coded a defaulter as 1 and a non-defaulter as 0, while often in literature reverse coding is used<sup>2</sup>. Additionally, the term 'override' here does not

<sup>2</sup> Note that, following most published credit scoring applications, we shall not consider different kinds of defaulters here, yet we will merely distinguish between 'goods' and 'bads'.

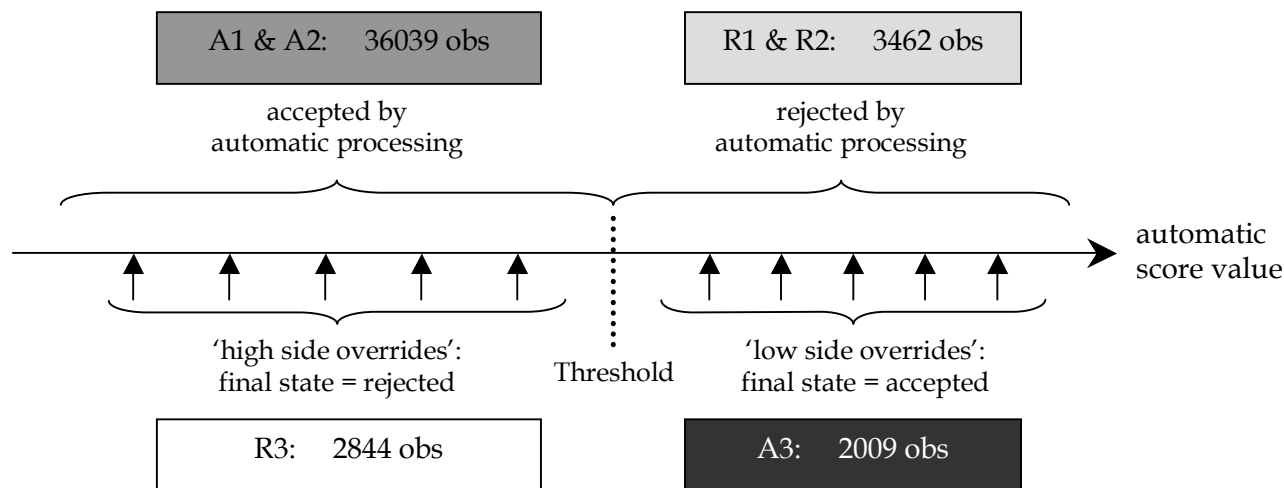


Figure 1: Order flow visualization

really correspond to its traditional connotations, since the decision to treat an order manually is an autonomous decision, and is by no means based on the scoring of the automatic scoring procedure, yet on a set of disjunctive decision rules: once one of the criteria is met, the order will be processed manually.

Several consequences can be drawn from the knowledge of the order handling process at the company in case, and are crucial to this study. First, it should be noted that the group of orders that were accepted by the score yet rejected by manual decision (i.e. the orders termed 'R3'), are of little importance to this study: they were mainly rejected for strategic and/or legal reasons (e.g. aged under 18) and thus cannot serve in the modeling process. Indeed, following Hand and Henley (1997), high side overrides will not lead to biased samples if the relevant application population is defined exclusively of those eliminated by a high side override. Second, the score accepted 36039 relevant orders, yet rejected a total of 5471 orders, resulting in an acceptance rate of 86,8 %. Third, of the latter set, 2009 orders were overridden (the orders termed 'A3'), and hence these cases provide a sizeable sample of the rejected cases containing default information. Indeed, while there is no objective reason why these orders should be accepted (they have a high probability of defaulting), according to the automatic scoring procedure they were still accepted regardless of any decision rule. It should be mentioned that in more than 95% of the orders rejected by the score, the orders were handled by the judgmental process (cf. cell 'R1' in Table 1 comprises only 234 orders). Hence, the decision to accept an order that was rejected by the score, was to a very large extent driven by manual decision making, where different company employees rely on different beliefs about credit risk. This sample of orders permits

the analyses that are proposed in this study, and has previously been called the calibration sample.

#### 4.2. Sample composition

For our research, each analysis proposed in the methodological section requires the use of different samples. Considering the importance of this sampling for the study, below, we will describe the data used in detail.

##### 4.2.1. Q1: Detection of sample bias

###### a) Underperformance of the score on rejected orders

In order to detect whether a model trained on the orders accepted by the score is better able to predict similar orders than orders from the calibration set, 50 % of the orders accepted by the score was used for training the model. As holdout samples, the remaining 50 % of the accepted orders was compared with the total calibration sample. We remind the reader that this sampling procedure was repeated 100 times, and that a stratified sampling procedure was used.

###### b) Variable-selection process

In order to detect whether proportionality influences the variable-selection process, we compared the variables selected on a (training) sample of 50 % of the accepted orders with a proportional sample constituted by 50 % of the calibration sample (i.e. 1005 rejected orders), and a sample of accepted orders (6617 accepted orders), ensuring the proportionality of the 'trough-the-door' applicant population (i.e. 36039 accepted versus 5471 rejected orders).

##### 4.2.2. Q2: Influence of sample bias

In this research question, we test the effect of proportionality given a certain sample size. Hence, the (proportional) holdout sample that will be used will consist of 50 % of the orders in the calibration sample (i.e. 1004 rejected orders), and a sample of accepted orders (6617 accepted orders), ensuring the proportionality of the 'trough-the-door' applicant population. In terms of the training data, we will compare a model built on the remaining 1005 rejected orders and a proportional sample of accepted orders (6617 accepted orders) with a sample of 7622 accepted orders only. A visualization of this sampling procedure can be found in Appendix 1.

Considering the possible impact of the ratio of accepted versus rejected orders on the results, following Hand and Henley's suggestion (1997, p 526), a sensitivity analysis will be performed whereby only 70 % of the orders will be treated as historically accepted orders. While the real threshold was defined on a probability of defaulting of 0.04, in this analysis, we have reconstructed the situation where orders would have been rejected if they had a default probability larger than 0.01743. Hence, a sample of 6833 orders was appended to the calibration sample, ensuring that half of this sample was used for model building, while the other half was used for model testing. In this case, the holdout sample consisted of 50% of the orders in the calibration sample (4421 'rejected' orders), and a sample of accepted orders (10494 accepted orders) , ensuring the 70 % proportionality. In order to check the impact of sample bias on credit-scoring performance, a model built on a proportional sample of 4421 rejected and 10494 accepted orders was compared with a model built on a sample of 14915 orders only. Again, we remind the reader that this sampling procedure was repeated 100 times, and that a stratified sampling procedure was used to ensure an equal ratio of defaulters in training versus holdout samples.

#### 4.2.3. Q3: Proportionality versus sample size

In this research question, we will evaluate the impact of an increase in the sample size of the set of orders accepted by the scoring procedure, and hence compare the effect of proportionality given unequal sample sizes. In order to evaluate this effect, we used the same setup as in the previous research question. However, the set of 21800 orders that was not used in the previous analysis was split up into 50 samples of 436 orders each, where each sample contained on average 7.6 defaulters. The samples were added successively, and predictive performance was compared on a holdout sample proportional to the applicant population. Again, in order to test for significant differences, this process was repeated 100 times.

#### 4.3. Variable creation

Both in the conception of the dependent (classification) variable as in terms of the independent variables (characteristics), the authors have relied heavily on the experience by the managers of the company as well as on findings in previous research. In terms of dependent variable, we have used the fact whether a third reminder was sent to the customer, because (i) this is the moment that the customer will be charged for his delay in

refunding, (ii) this reminder really urges the customer to repay and (iii) this variable had historically been used by the company to investigate defaulting behavior<sup>3</sup>.

The customer characteristics that were used in the study belong to the traditional set of characteristics used for credit-scoring purposes. However, due to the fact that this is a consumer credit setting, it is a strategic decision of the company to limit the information inquired upon application and to rely more on own-company credit records. Nevertheless, we were also able to include characteristics covering demographic information and occupation, financial information (e.g. number of credits still open) and default information (e.g. ratings by credit bureau reports, own company default information). The list and description of the 45 characteristics used in this study can be found in Appendix 2.

## 5. Results

### 5.1. Q1: Detection of sample bias

#### 5.1.1. Q1a: Underperformance of the score on rejected orders

When considering sample bias, a basic assumption is that a scoring model built on accepted orders only, performs significantly worse on orders that would have been rejected by the previous score, than on orders that would have been accepted by the previous score. The results of this first step confirm the hypothesis. A model trained on orders accepted by the score was tested on similar orders, and compared with a model tested on orders rejected by the score. After repeating a resampling procedure 100 times, the AUC of the latter set was on average 8.12 percentage points lower than the AUC of the first set, a difference that was significant at  $p < 0.0001$  (t Value of 48.02). Hence, we can indeed conclude that the extrapolation of the score towards a range of customers that was not used for training does prove to be more erroneous than applying the score towards similar orders as those in the training set. The degree to which this poses a problem for the scoring performance as a whole, will be discussed at length after the following section.

#### 5.1.2. Q1b: Variable-selection process

In terms of the variable selection procedure, we have used the leap-and-bound algorithm of Furnival and Wilson (1974) to detect the best model for all possible model sizes

---

<sup>3</sup> While information was also available about whether the individual order was eventually profitable to the company, the use of the third reminder was preferred over the profit information, as the number of unprofitable orders was low, which degraded the performance of all models severely.

(number of variables), requiring a minimum of arithmetic, and the possibility for finding the best subsets without examining all possible subsets<sup>4</sup>.

However, since this procedure only provides a likelihood score (chi-square) statistic without significance tests, we have used the algorithm proposed by De Long et al. (1988) in order to investigate whether the AUC of a model with a given model size differs significantly from the AUC of the single model containing all of the explanatory effects. Starting from the full model containing 45 characteristics and reducing model size, we have selected the last model that does not differ significantly from the model using all characteristics at a 0.05 significance level. Coincidentally, both in a model based on accepted orders only, as in a model built on a proportional sample of accepted and rejected orders, this resulted in a selection of 31 characteristics. Of those, 24 characteristics were selected by both models, while 7 characteristics (i.e. about 23 %) were different. Hence, it is likely that (at least some of) the characteristics needed for the prediction of creditworthiness of previously rejected orders are different from the characteristics needed to evaluate accepted orders. The degree to which this difference has an influence on credit-scoring performance and profitability is an important topic in the following section.

The graphical output of the variable selection process, in terms of the different tests, is given in Appendix 3, where the white line represents the performance of a model using all characteristics, and the black line represents the performance given a specific model size. The number of variables we finally withheld is represented by a dashed vertical line.

## 5.2. Q2: Influence of sample bias

In this section, we will give an overview of the results of our main research topic, namely the impact of sample bias on consumer credit-scoring performance and profitability for a given sample size. In this set of analyses, we have attempted to assess the confidence interval around each performance indicator by resampling the data 100 times, where the reported differences in (test set) performance between different models were always computed within one iteration of the resampling procedure. Hence the degrees of freedom used for the tests were 99. Following Hand and Henley (1997), who state that, in mail-order purchasing, a figure of 70 % accepted orders is quite usual, all analyses that were performed here are validated in a sensitivity analysis, where we reuse the data to reconstruct the case if

---

<sup>4</sup> Note that, if all possible subsets are to be evaluated, a full model containing  $n$  variables would require the computation of  $\sum_{i=1}^n \frac{n!}{(n-i)! i!}$  models containing  $1 \dots n$  variables.

only 70 % of all applicants had been accepted. To enhance comparability, the results of the real setting (acceptance rate of 86.8 %) and the sensitivity analysis (70 %) are always presented side by side. To enable the reader to compare the results with the performance of the previous credit score, we included its performance and labeled it “old” performance. Model 1 represents a model where only orders were used that were accepted by the scoring procedure, both in terms of inclusion of the orders in the training sample, as in terms of inclusion of the orders in the variable selection procedure (labeled ‘AA’). Model 2 uses a sample of orders accepted by the score and/or accepted by the manual selection process, in such a way that the proportionality of accepted versus rejected orders is respected. However, in model 2, the variable-selection procedure is still performed on a sample of accepted orders only (‘PA’). Finally, in model 3, the same orders were used as model 2, but the variables of this model were selected on a proportional sample of accepted and rejected orders (‘PP’). We first start by covering credit-scoring performance, and then discuss profitability issues.

### 5.2.1. Predictive Performance

The results in terms of PCC performance can be found in Table 2. In order to compute these results, the probability of defaulting was discretized into two classes, new accepts and new rejects, in a way that the proportionality of the new model was ensured. Hence, we report the performance if the same acceptance rate would be applied for the use of the new model than during the use of the previous model.

	Actual setting (86.8 % accepted)			Sensitivity analysis (70 % accepted)		
<b>Morrison</b>	0.8520			0.6950		
	<b>Mean</b>	<b>Std Dev</b>		<b>Mean</b>	<b>Std Dev</b>	
<b>PCC old</b>	0.8601	0		0.7069	0	
<b>PCC model 1: ‘AA’</b>	0.8643	0.0016		0.7114	0.0011	
<b>PCC model 2: ‘PA’</b>	0.8646	0.0018		0.7117	0.0009	
<b>PCC model 3: ‘PP’</b>	0.8643	0.0018		0.7112	0.0010	
	<b>Mean</b>	<b>t Value</b>	<b>P &gt;  t </b>	<b>Mean</b>	<b>T Value</b>	<b>P &gt;  t </b>
<b>PCC 2 - PCC 1</b>	0.0003	2.00	0.0482	0.0003	3.99	0.0001
<b>PCC 3 - PCC 1</b>	-26E-6	-0.14	0.8872	-16E-5	-1.72	0.0890
<b>PCC 3 - PCC 2</b>	-38E-5	-2.60	0.0106	-5E-4	-8.23	<.0001

Table 2 : PCC performance when reducing sample bias at a given sample size



In terms of testing the classification accuracy versus a random model, we have used the formula proposed by Morrison (1969) which states that the accuracy of a random model is defined by:

$$pa + (1 - p) (1 - a),$$

where  $p$  represents the true proportion of refunded orders and  $a$  represents the proportion of applicants that will be accepted for credit (Morrison, 1969, p. 158). All models performed significantly better than this random-model benchmark ( $p < .0001$ ). Hence, we are confident that all models (also the previous credit score) perform reasonably well. Additionally, the PCC of all new models was significantly higher than the PCC of the previous model ( $p < .0001$ ), while the mean differences ranged between 0.2 and 2 %, indicating a clear improvement by the new model.

The impact of sample bias on PCC performance is illustrated by the three lower rows of Table 2, where the differences between the new models are tested. From these tests, it is clear that (i) the second model - containing a proportional sample of accepted and rejected orders - performs significantly better than the first model, built on accepted orders only. Henceforth, we consider significance on a 0.05 confidence level. Furthermore, (ii) the third model - containing the variables selected on a proportional sample of rejected and accepted orders - does not perform better than the first model. Additionally, (iii) the third model performs significantly worse than the second model. Consequently, the second model performs best in terms of PCC, both in the real setting as in the sensitivity analysis, but the impact on PCC that can be reached by including the calibration sample in a proportional way seems to be low (0.0003), especially when compared to the difference resulting from the update of the model (0.0042).

A main drawback of PCC performance is that it requires the user to discretize the probability of defaulting, such that the model will only be evaluated for a given threshold. This, however, does not give the user any indication of how the model performs if other threshold levels were to be used. Since AUC does give an evaluation of a score across the total range of default probabilities, we report the AUC performance in Table 3.

	Actual setting (86.8 % accepted)			Sensitivity analysis (70 % accepted)		
	Mean	Std Dev		Mean	Std Dev	
AUC old	0.7131	0.0138		0.7041	0.0054	
AUC model 1: 'AA'	0.7487	0.0198		0.7534	0.0143	
AUC model 2: 'PA'	0.7550	0.0182		0.7584	0.0117	
AUC model 3: 'PP'	0.7472	0.0191		0.7514	0.0123	
	Mean	t Value	P >  t	Mean	t Value	P >  t
AUC 2 - AUC 1	0.0063	5.31	<.0001	0.005	4.49	<.0001
AUC 3 - AUC 1	-0.002	-1.09	0.2799	-0.002	-1.90	0.0605
AUC 3 - AUC 2	-0.008	-8.45	<.0001	-0.007	-13.58	<.0001

Table 3 : AUC performance when reducing sample bias at a given sample size

The results of this analysis are completely analogous to the PCC results. Hence, (i) all new models perform significantly better than the previous model ( $p < 0.0001$ ), (ii) model 2 performs significantly better than models 1 and 3, (iii) there is no statistical difference between 1 and 3, and (iv) the improvement of performance between model 2 and model 1 seems relatively small compared to the difference resulting from the update of the model.

### 5.2. 2. Profitability

Since order profitability could be computed at the individual order level in the database, in this section, we will review the impact of sample bias on consumer credit scoring profitability, given a certain sample size. Considering the confidentiality of the data, the authors were unable to reveal absolute profit information. Therefore, Table 4 only represents the relative profit changes that could be reached by introducing the information

	Actual setting (86.8 % accepted)			Sensitivity analysis (70 % accepted)		
	Mean	t Value	P >  t	Mean	t Value	P >  t
Profit Difference 2 vs. 1	0.0039	5.31	<.0001	0.0063	3.93	0.0002
Profit Difference 3 vs. 1	0.010	9.33	<.0001	0.0309	12.98	<.0001
Profit Difference 3 vs. 2	0.0043	6.11	<.0001	0.0182	8.40	<.0001

Table 4 : Profit implications when reducing sample bias at a given sample size

stemming from rejected orders. This difference is again computed per resampling iteration, and the average of the 100 resamples is represented and tested against the null hypothesis that this difference is zero.

In contrast to the results for both classification-performance measures, the effect of the bias in terms of variable selection does play when considering credit-scoring profitability. Indeed, while model 2 again performs significantly better than model 1, model 3 now performs significantly better than model 2. Hence, the model that uses the calibration sample to reduce sample bias by ensuring the proportionality of accepted versus rejected orders and by improving the variable-selection process, clearly delivers the optimal solution. Thus, it is confirmed that insights gained from credit-scoring profitability possibly differ from classification-performance results. Additionally, in the company's perspective, profitability is the most relevant indicator of model quality.

While the difference is significant, it seems to be low. For example, in the current setting of 86.8 % accepted orders, the maximum profit gain that could be reached by gaining the knowledge of the outcome for all rejected orders (assuming the price for gaining this knowledge to be zero), would be 1 %. Hence, any procedure of reject inference that results in perfect imputations of the defaulting behavior of rejected orders would maximally reach this improvement. In conclusion, the profit that can be realized by introducing information from the rejected orders into the model seems modest, especially when it should be able to cover the defaulting cost of including a random sample or the time cost involved in applying any reject-inference procedure.

To conclude this section, we detect from the comparison of the profit implications of the actual setting with the sensitivity analysis that the profitability from including the calibration sample rises as the proportion of rejected orders grows larger. While this effect was not tested statistically, it seems only logical that the impact of the reduction of sample bias rises when sample bias itself grows in size.

### 5.2.3. Q3: Proportionality versus sample size

The results of the previous analysis indicate the improvement that could be realized whenever the outcome of all rejected orders would be available. However, in practice (e.g. also in the current application), it is likely that the outcome of only a sample of the rejected orders is available. We have proven earlier that, whenever the acceptance rate is larger than the proportion of rejected orders for which the outcome is available, the sample size of the set of accepted orders will be larger than the sample size of the proportional order set.

Hence, it is plausible that this asymmetry in sample sizes can be used to reduce the impact of sample bias. The results of this research question confirm the hypothesis. Both in terms of predictive performance and profitability, the surplus performance stemming from including a proportional sample decreases (or even vanishes) as the sample size of the set of accepted order increases. In this application, considering predictive performance, a model containing all available orders accepted by the score performs better than the model built on a proportional sample of rejected versus accepted orders. However, in terms of profitability, while the profitability of the model built on accepted orders rises as its sample size increases, the profitability of the latter models never outperforms the profitability of the proportional model. Nevertheless, in general, all profitability differences seem modest at best. Again, the usefulness of using different performance indicators is shown here.

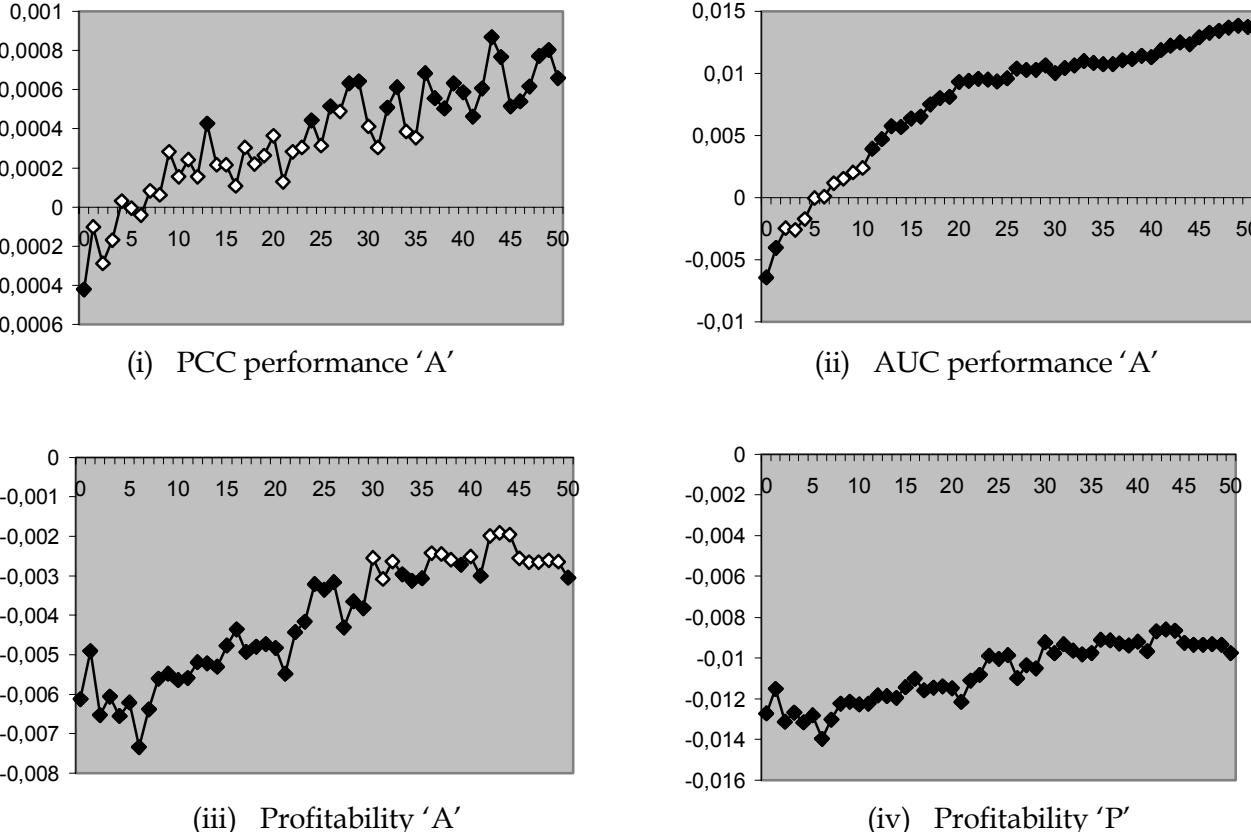


Figure 2: Credit scoring performance and profitability at an increasing sample size

The output of these analyses can be found in Figure 2, where the difference in performance between a model using accepted orders versus a model built on a proportional sample is plotted for 50 increases in sample size. A white mark is used whenever the difference is not significant, whereas black marks indicate significant differences at an alpha

level of 0.05. The graphs are labeled 'A' whenever the comparison is made by means of a variable-selection procedure using accepted orders only, and 'P' in the case of a variable selection procedure using a proportional set of accepted versus rejected orders. The two missing graphs can be found in Appendix 4, since they added no further insights into the research question.

## **6. Conclusions**

In this study, the authors attempted to indicate and quantify the impact of sample bias on consumer credit scoring performance and profitability. Historically, sample bias has been suggested to pose a sizeable threat to profitability due to its implications on either population drainage or biased estimates. While previous research has mainly been focused on offering various attempts to reduce this bias, mainly due to lack of appropriate credit scoring data sets, the impact of sample bias itself has been largely unexplored. By means of the properties of the data available for this study, however, the authors were able to assess the existence and the impact of sample bias in an empirical setting. In the remainder of this section, we summarize and discuss the results of the analyses performed.

The results of the study indicate that sample bias does appear to have a negative influence on credit-scoring performance and profitability. Indeed, firstly, a model trained on accepted orders only reveals an important and significant underperformance on the rejected orders compared to its performance on other accepted orders. Secondly, the variable selection procedure used in this paper, showed that (at least some) other characteristics may be relevant for predicting the creditworthiness of previously rejected or previously accepted orders. Thirdly, and most importantly, in order to quantify the effect of sample bias on credit scoring performance and profitability, we compared model performance of a model built on accepted orders only, with an equally-sized sample of accepted and rejected orders proportional to the applicant population. The results from this analysis indicate that sample bias does prove to have a significant, albeit modest effect on consumer credit scoring performance and profitability. Additionally, in terms of predictive performance, the negative impact only occurs through the inclusion of rejected orders in the training data set, while in terms of profitability, both the inclusion of the rejected orders in the training data set, as in the variable-selection procedure proved to have a significant impact. It should be mentioned, however, that both in the actual setting as during a sensitivity analysis, the impact of knowing the outcome of all the orders rejected by the score is limited, especially when the costs of gaining this knowledge must be accounted for. Fourthly, and finally, we have shown

that comparable performance levels can be reached by increasing the size of the homogeneous sample of accepted orders only, instead of introducing a proportional sample of rejected orders into the model-building process. To conclude, on a theoretical level, the effect of proportionality prevails, and enhancing proportionality can result in improvements in classification accuracy and profitability. However, it should be clear that, at least in this consumer credit setting, the resulting benefits from determining the true outcome values of the rejected cases are low, and company resources could be spent more efficiently by handling other topics relevant to consumer credit scoring.

## **7. Limitations and Issues for Further Research**

While the specificity of the data was attempted to be minimized through a sensitivity analysis and the use of different performance measures, this study was executed on the data of a direct-mail company. Unfortunately, the results cannot be extrapolated without reflection towards non-consumer credit scoring, considering the specific properties of the data set used, being (i) a rather large percentage of accepted orders (86.8 %), (ii) a rather low percentage of defaulters (1.94 %), (iii) a rather balanced structure of the misclassification costs (the cost involved with a defaulter was only 2.58 times higher than the profit gained from a non-defaulter). Consequently, it would be useful to replicate the analyses performed here on the data of other credit-offering institutions. Nevertheless, in this paper, we have offered a workable methodology towards analyzing the impact of sample bias in any credit scoring environment. More specifically, during the sensitivity analysis, we offered a procedure that can be implemented to investigate the impact of sample bias whenever historical score values were recorded. Further research largely depends on the availability of other credit scoring datasets.

## **8. Acknowledgements**

The authors wish to thank Jonathan Burez, Joanna Halajko and Bernd Vindevogel, graduates from the Master in Marketing Analysis at Ghent University, for their useful assistance in terms of data preparation.

## References

- Avery R.B. (1981), "Credit scoring models with discriminant analysis and truncated samples", *Research Papers in Banking and Financial Economics*, 54.
- Baesens B., Van Gestel T., Viaene S., Stepanova M., Suykens J., and J. Vanthienen (2003), "Benchmarking state of the art classification algorithms for credit scoring", *Journal of the Operational Research Society*, forthcoming.
- Baesens B., Viaene S., Van den Poel D., Vanthienen J. and G. Dedene (2002), "Using bayesian neural networks for repeat purchase modelling in direct marketing", *European Journal of Operational Research*, 138 (1), pp. 191-211.
- Chandler G.G. and J.Y. Coffman (1977), "Using credit scoring to improve the quality of consumer receivables: legal and statistical implications", Paper presented at the Financial Management Association meetings, Seattle, Washington.
- De Long E.R., De Long D.M. and D.L. Clarke-Pearson (1988), "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach", *Biometrics*, 44, pp. 837-845.
- Desai V.S., Crook J.N. and G.A. Overstreet Jr. (1996), "A comparison of neural networks and linear scoring models in the credit union environment," *European Journal of Operational Research*, 95, pp. 24-37.
- Davis R.H., Edelman D.B. and A.J. Gammerman (1992), "Machine-learning algorithms for credit-card applications", *IMA Journal of Mathematics Applied in Business and Industry*, 4, pp. 43-51.
- Egan J.P. (1975), "*Signal detection theory and ROC analysis*", Academic Press, New York.
- Eisenbeis R.A. (1977), "Pitfalls in the application of discriminant analysis in business, finance and economics", *Journal of Finance*, 32, pp. 875-900.
- Feelders A.J. (2000), "Credit scoring and reject inference with mixture models", *International Journal of Intelligent systems in Accounting Finance and Management*, 8 (4), pp. 271-279.
- Furnival G.M. and R.W. Wilson (1974), "Regressions by leaps and bounds," *Technometrics*, 16, pp. 499-511.
- Hand D.J. (2001), "Modelling consumer credit risk", *IMA Journal of Management Mathematics*, 12 (2), pp. 139-155.
- Hand D.J. and W.E. Henley (1994), "Can reject inference ever work?", *IMA Journal of Mathematics Applied in Business and Industry*, 5 (1), 45-55.
- Hand, D.J. and W.E. Henley (1997), "Statistical classification methods in consumer credit scoring: a review.", *Journal of the Royal Statistical Society A*, 160, pp. 523-541.

- Hsia D.C. (1978) "Credit scoring and the equal credit opportunity act." *Hastings Law Journal*, 30 (2), pp. 371-448.
- Joanes D.N. (1994), "Reject inference applied to logistic regression for credit scoring", *IMA Journal of Mathematics Applied in Business and Industry*, 5 (1), pp. 35-43.
- Maddala G.S. (1992), "*Introduction to Econometrics*", Maxwell MacMillan Int. Editions, NY.
- Myers J.H. and E.W. Forgy (1963), "The development of numerical credit evaluation systems", *Journal of the American Statistical Association*, 58 (303), pp. 799-806.
- Morrison D.G. (1969), "On the interpretation of discriminant analysis", *Journal of Marketing Research*, 6, pp. 156-163.
- Rosenberg E. and A. Gleit (1994), "Quantitative methods in credit management: a survey", *Operations Research*, 42 (4), pp. 589-613.
- Stepanova M. and L. Thomas (2002); "Survival analysis methods for personal loan data", *Operations Research*, 50 (2), pp. 277-289.
- Thomas L.C. (2000), "A survey of credit and behavioral scoring: forecasting financial risk of lending to customers", *International Journal of Forecasting*, 16 (2), pp. 149-172.
- Thomas L.C., Ho J. and W.T. Scherer (2001), "Time will tell: behavioural scoring and the dynamics of consumer credit assessment", *IMA Journal of Management Mathematics*, 12, pp. 89-103.



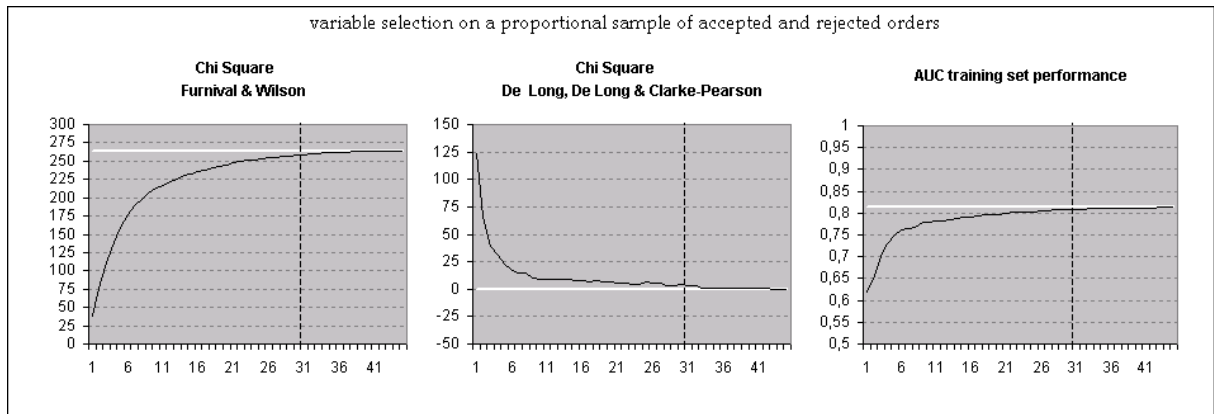
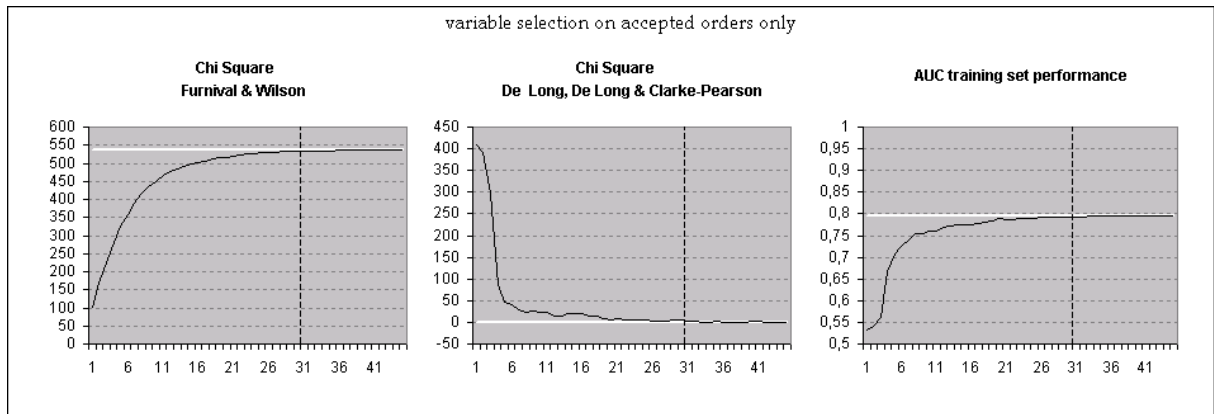


## APPENDIX 2: List of variables used

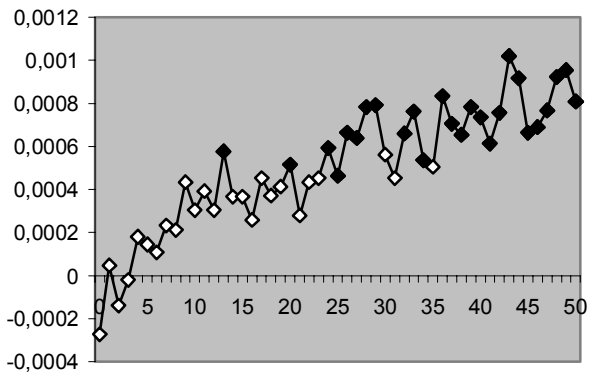
	Name	Type	Description	A*	P*
1.	Home_Own1	Binary	Home ownership: 1=owner, 0=renter	x	x
2.	Home_Own2	Binary	Home ownership: 1=missing, 0=non missing	x	x
3.	Occup1	Binary	1=full-time, 0=other cases		
4.	Occup2	Binary	1=retired, 0=other cases	x	x
5.	Occup3	Binary	1=housewife, 0=other cases		x
6.	Occup4	Binary	1=living on social welfare, 0=other cases	x	x
7.	Occup5	Binary	1=student, 0=other cases	x	x
8.	Occup6	Binary	1=without profession, 0 = other cases	x	x
9.	Occup7	Binary	1=missing, 0=other cases	x	x
10	Bank_Acc	Continuous	Number of bank accounts	x	x
11	Debt1	Binary	1=current debt <10.000, 0=other cases	x	x
12	Debt2	Binary	1=(10.000<=current debt<25.000), 0=other cases	x	x
13	Debt3	Binary	1=(25.000<=current debt <60.000), 0=other cases	x	x
14	Debt4	Binary	1=(current debt >=60.000), 0=other cases		x
15	Open_credit	Continuous	Number of credits that are still due	x	x
16	Amount_open	Continuous	Amount that is still open for those credits	x	
17	Open_credit_old	Continuous	Number of credits that are older than 120 days		x
18	Amount_open_old	Continuous	Amount open for those credits	x	
19	Amount_all	Continuous	Amount of all the credits of a customer	x	
20	Amount_paid	Continuous	Amount of the credits that were refunded	x	
21	Ratio_paid	Continuous	Amount_paid / Amount_all	x	x
22	Blacklist1	Binary	1=listed on the 'black list' VKC, 0=other cases	x	
23	Blacklist2	Binary	1=missing value 'black list' VKC, 0=other cases	x	
24	Blacklist3	Binary	1=listed on the 'black list' UPC, 0=other cases	x	x
25	Blacklist4	Binary	1=missing value 'black list' UPC, 0=other cases		
26	Remind1	Binary	1=reminder history not known, 0=other cases		x
27	Remind2	Binary	1=received 2 <sup>nd</sup> reminder, 0=other cases	x	x
28	Remind3	Binary	1=received 3 <sup>rd</sup> reminder, 0=other cases		
29	Remind4	Continuous	Number of 1 <sup>st</sup> reminders over customer relationship		
30	Remind5	Continuous	Number of 2 <sup>nd</sup> reminders over customer relationship		x
31	Remind6	Continuous	Number of 3 <sup>rd</sup> reminders over customer relationship	x	x
32	Remind_install4	Continuous	Number of 1 <sup>st</sup> reminders per installment		
33	Remind_install5	Continuous	Number of 2 <sup>nd</sup> reminders per installment	x	x
34	Remind_install6	Continuous	Number of 3 <sup>rd</sup> reminders per installment		
35	Remind7	Continuous	Number of 1 <sup>st</sup> reminders on short term consumer credit	x	
36	Remind8	Continuous	Number of 2 <sup>nd</sup> reminders on short term consumer credit		
37	Remind9	Continuous	Number of 3 <sup>rd</sup> reminders on short term consumer credit		x
38	Remind_install7	Continuous	Number of 1 <sup>st</sup> reminders on short term credit per installment	x	x
39	Remind_install8	Continuous	Number of 2 <sup>nd</sup> reminders on short term credit per installment	x	x
40	Remind_install9	Continuous	Number of 3 <sup>rd</sup> reminders on short term credit per installment		x
41	Default1	Binary	Client has defaulted on his credit during the last two years	x	x
42	Default2	Binary	Client has defaulted on his credit during the last 15 years	x	x
43	Remind_last	Continuous	Summary score for the reminders on the last order	x	x
44	Remind_1butlast	Continuous	The same summary score for the one but last order	x	x
45	Increase_remind	Continuous	Increase/decrease in the summary score for reminders	x	x

\* In this table, 'A' represents the variables selected on a sample of accepted orders only, while 'P' represents the variables selected on a proportional sample of accepted versus rejected orders.

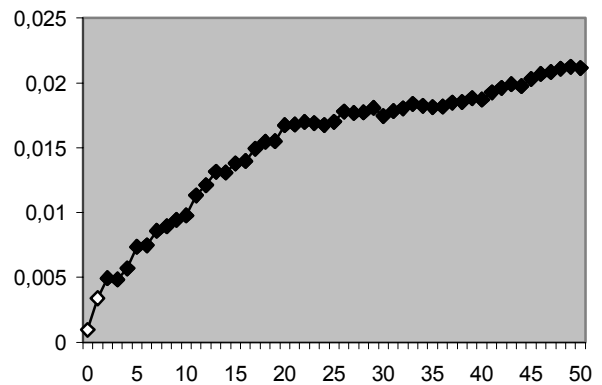
### APPENDIX 3: Variable selection process



### APPENDIX 4: Credit scoring performance at increasing sample size



(i) PCC performance 'P'



(ii) AUC performance 'P'