An Analysis of the Applicability of Credit Scoring for Microfinance

Joris Van Gool[1], Bart Baesens[1,2], Piet Sercu[1], Wouter Verbeke[1]

[1] Faculty of Business and Economics
Katholieke Universiteit Leuven
Naamsestraat 69, B-3000 Leuven, Belgium

[2] School of Management
University of Southampton
Southampton SO17 1BJ, United Kingdom

Abstract

Credit scoring has been succesfully applied in domains as mortgage loans and credit cards. This paper analyzes whether credit scoring should be adopted in microfinance institutions. Previous research on credit scoring in the microfinance field was mostly situated in Latin America and Southern Africa. To the best of our knowledge, no studies have been published before on the Eastern Europe-Central Asia and Middle East-North Africa regions. Furthermore, opinions on the applicability of the concept are widely diverging and statistical concepts as weight of evidence coding and measures as AUC are not commonly used in credit scoring studies for microfinance. This study provides evidence from a mid-sized Bosnian microlender, includes relevant statistical concepts and measures, and formulates general conclusions for the field. A binary logistic model with dummy coding and a binary logistic model with weight of evidence coding are developed. Based on the stability, readability and discriminatory power results of the models, it is shown that credit scoring is not able to fully replace the traditional credit process for microfinance. Credit scoring can however be introduced as a refinement tool in the credit process, to combine both statistical and human best practices.

Keywords: Microfinance, Credit Scoring, Logistic Regression, Credit Risk, Bosnia-Herzegovina

Corresponding Author is Joris Van Gool (joris.vangool@student.kuleuven.be). The authors would like to thank Nejira Nelic, CEO of Mi-Bospo for her willingness to share the organization's database and staff expertise during the study-visit in January 2009. We are also thankful for the support from J. Andrew Spindler (FSVC), Wim Vermeir (Dexia), Renée Spierings (Dexia) and Hans Salzmann (FMO)

1 Introduction

Robinson (2001) pioneered the idea of the "win-win" proposition in microfinance: social impact could go hand in hand with financial sustainability or even profit-making. Powerful actors such as the Consultative Group to Assist the Poorest (CGAP; an independent policy and research center housed at the World Bank) and United States Agency for International Development (USAID; US development agency) promoted this proposition through publications, workshops etc . Even though some scholars as Morduch (2000) saw limits to the "win-win" idea, the general opinion was inclined to believe in both the social and sustainable powers of microfinance. Furthermore, due to increasing competition, over-indebtedness and economic crisis in several microfinance regions, microfinance institutions (MFIs) had to pursue their social and financial objectives in increasingly constrained environments. Using the right tools to manage risk became more than ever a key competence to survive. It is in this context that established techniques from traditional financial organizations were introduced into the microfinance industry, with the aim to improve both social outreach and financial sustainability. One of these techniques was credit scoring, which analyzes historical client data and derives a model which links repayment behavior with characteristics of the loan, lender and borrower.

Vigano (1993) pioneered the application of credit scoring models for microfinance. Since then, most studies focused on Latin America and Southern Africa. Empirical evidence on credit scoring for developing countries in general is rather limited (Vogelgesang, 2003; Kleimeier and Dinh, 2007). No studies seem to have been published for the Eastern Europe-Central Asia and Middle East-North Africa microfinance sectors. Furthermore, opinions on the applicability of the concept for microfinance are diverging and mostly qualitatively motivated. This paper aims to extend the research on credit scoring by specifically investigating why credit scoring should be adopted for microfinance, based on quantitative results from a credit scoring application in Bosnia-Herzegovina. In a first step, two scoring models are developed and interpreted. Next, based on an assessment of the stability, readability and discriminatory power of the models, a recommendation with regards to the applicability of credit scoring for microfinance is made.

Section 2 presents a summary table on prior credit scoring studies for microfinance and analyzes evaluations on the applicability of the concept. Section 3 discusses the building blocks necessary to construct a statistical credit scoring system for microfinance, leading to the development of two competing logistic regression-based models. The data set and the data features are presented in section 4. Furthermore, also the two models developed for this study are introduced. Section 5 presents the results, section 6 discusses whether credit scoring should be adopted in microfinance institutions, based on the results for the Bosnian microlender. Finally, section 7 concludes with a summary of the key findings.

2 Credit Scoring for Microfinance: Context and Evaluation

Basel II, as developed by the Basel Committee on Banking Supervision (2006), provides a standard framework for measuring capital adequacy and assessing the underlying risk management of internationally active banks. Even though microlending is not explicitly discussed, Basel II allows to situate the content of this paper in a wider context as several Basel II concepts were found to be applicable for microfinance (Navarrete and Navajas, 2006).

One of the main risks lined out in the Basel II framework is credit risk, defined as the risk of default by the borrower. Two methodologies are proposed in Basel II to calculate the capital requirements for credit risk: the standardized and the internal rating based approach. The former approach measures credit risk in a standardized manner, executed by external credit rating agencies. The latter approach allows and encourages banking institutions to develop their own internal measures for the assessment of credit risk capital. Basel II explicitly mentions credit scoring as a possible technique to determine drivers of credit risk.

Credit scoring can take different forms. Three main types of credit scoring approaches can be distinguished (Thomas, 2000): (i) judgmental, (ii) statistical and (iii) non-statistical, non judgmental. The judgmental approach is still in place at most microlenders (Schreiner, 2004) and assesses risk based on the experience and opinion of the loan officer itself. In contrast, statistical approaches are based on historical data and include discriminant analysis and logistic regression. The statistical approach forms the focus of this paper and more information is provided in section 3.2. Non-statistical, non judgmental methodologies include a variety of operational research methods, neural networks and genetic algorithms. As Baesens et al. (2003) report, results on the performance of different credit scoring approaches are often conflicting. For example Desai et al. (1996) report that neural networks performed significantly better than linear discrimant analysis while Yobas et al. (2000) reported the inverse results. In this context, Thomas (2000) states that credit scoring typically employs a pragmatic approach – 'if it works, use it'. In particular, combinations of different approaches are often used as they might generate the best results in particular circumstances.

Table 1 gives an overview on published statistical credit scoring models for developing countries. Three main remarks can be made. First, this overview confirms the reports of Vogelgesang (2003) and Kleimeier (2007) that most published credit scoring studies for microfinance have focused on Latin America and Southern Africa and that empirical evidence on credit scoring for developing countries in general is very limited. Furthermore, as Schreiner (2004) also remarks, the predictive power of the African or Asian models of Reinke (1998), Zeller (1998) and Sharma and Zeller (1997) is not statistically validated.

Third, the models of Zeller (1998) and Sharma and Zeller (1997) are group lending models, while Schreiner (2003) remarks that scoring will probably not work for group loans.

Published Credit Scoring Models for Developing Countries

Author (Date, Country)	Institution Type	Sample Size	Number of (Included) Inputs	Technique(s)	Performance Metrics
Vigano (1993, Burkina Faso)	Microfinance	100	53(13)	Discriminant Analysis	PCC, R^2
Sharma and Zeller (1997, Bangladesh)	Microfinance	868	18(5)	TOBIT Maximum Likelihood Estimation	N/A
Zeller (1998, Madagascar)	Microfinance	168	19(7)	TOBIT Maximum Likelihood Estimation	N/A
Reinke (1998, South Africa)	Microfinance	1641	8(8)	Probit Regression	N/A
Schreiner (1999, Bolivia)	Microfinance	39,956	9(9)	Logistic Regression	PCC
Vogelgesang (2001, Bolivia)	Microfinance	8,002	28(12)	Random Utility Model	PCC, Pseudo-R^2
Vogelgesang (2001, Bolivia)	Microfinance	5,956	30(13)	Random Utility Model	PCC, Pseudo-R^2
Diallo (2006, Mali)	Microfinance	269	17(5)	Logistic Regression, Discriminant Analysis	PCC, R^2
Kleimeier et al. (2006, Vietnam)	Retail Bank	56,037	22 (17)	Logistic Regression	PCC, SENS, SPEC

Table 1: Overview of Published Credit Scoring Models for Developing Countries. Sample Size is total number of observations used, combining training and test sets. Number of Inputs is the total number of inputs available, Number of Included Variables is the number of selected inputs in the final model. If known, a 5 percent significance level is employed as selection criterium. Dummy variables or transformations belonging to one (categorical) variable are counted as one variable. PCC stands for Percentage Correctly Classified, SENS for sensitivity and SPEC for specificity (see section 3.4). Vogelgesang (2001) published multiple models in her study, the two models reviewed in this table are illustrative for the other models.

While several authors have evaluated the usefulness of credit scoring, only few have focused on the applicability of the concept for microfinance. Diverging opinions on the usefulness of credit scoring for microfinance exist. Some authors point out advantages as smaller default losses, potential for marketing to different segments and decreasing loan officer time spent to individual clients (Dennis, 1995; Schreiner, 2003; Kulkosky, 1996). Others note that credit scoring is vulnerable to several statistical limits and that it cannot model risks such as unwillingness to repay and inability due to natural catastrophes, even though those risks often significantly influence default (Capon, 1982; Schreiner, 2003; Freytag, 2008). It should be noted that most of the authors do not base their opinions on quantitative grounds.

3 Building a Scoring Model for Microfinance: Theory

The goal of credit scoring for microfinance and for other financial purposes is to optimally discriminate between good and bad loans. Best practices employed in other domains such as weight of evidence coding and the AUC performance measure, which have not been commonly used before in credit scoring for microfinance, are included in this study. Database construction and data preprocessing issues are discussed in section 3.1. Section 3.2 assesses different statistical methodologies and explains the choice for logistic regression. The treatment and selection of explanatory variables is discussed in section 3.3. Finally, section 3.4 explains the validation of credit scoring models.

3.1 Database Construction and Data Preprocessing

The data preparation step deals with the choice and creation of the desired variables and the treatment of missing values and outliers (Van Gestel et al., 2006).

As Kleimeier (2007) remarks, there is no universally accepted approach to choose the candidate explanatory variables for a credit scoring model. A literature review indicates that most authors refer to expert advice and (or) prior studies to explain their choices (Kleimeier and Dinh, 2007; Schreiner, 2004). The candidate explanatory variable choice for this paper was based on expert advice from the Bosnian microlender.

For several credit scoring models, new explanatory variables are needed, in particular to obtain averages or to create proxies for non-measurable data (Vogelgesang, 2003; Van Gestel et al., 2006). Also for the dependent variable in credit scoring, authors often need to create their own variables when the required data is not directly available or when the purposes of the rating model require a specific variable. In this paper, the dependent variable was created based on expert advice from the Bosnian microlender. For binary logit models 1 and 2 (see section 4.2), the dependent variable for each i^{th} loan is defined as following:

$$Y_i = \begin{cases} 0 & \text{if average delay per installment} \leq 2 \text{ days} \\ 1 & \text{if average delay per installment } > 2 \text{ days} \end{cases} \tag{3.1}$$

Next, missing values need to be treated. Beale and Little (1975) note that one popular approach, mean imputation for continuous variables, generally creates acceptable results. Van Gestel et al. (2006) adopt median imputation for continuous variables and mode imputation for categorical variables. Due to the high data quality, missing values only had to be corrected for one variable in this paper: *job experience*. Considering the categorical nature of *job experience*, mode imputation was applied.

The final data preprocessing step is outlier handling. Well-known, common approaches include the winsorized and trimmed means (Wainer, 1976). As the occurence of outliers in the data used for this paper is rather limited, with no signs of correlated outliers, a simple percentile approach was adopted. All observations under (or above) a 0.5% (99.5%) percentile were replaced by these limits.

3.2 Methodologies for Credit Scoring

There are three main approaches for credit scoring (Thomas, 2000): (i) judgmental, (ii) statistical and (iii) non-statistical, non judgmental. This paper focuses on the statistical approach, which is based on historical data and includes methodologies as discriminant analysis and logistic regression.

Discriminant analysis is a computationally efficient procedure, but is hampered by the assumption of normally distributed data (Sharma, 1996; Vogelgesang, 2003). As the models presented in this paper include multiple dummy variables, the normality assumption is violated and therefore discriminant analysis has not been adopted.

Logistic regression employs maximum likelihood estimators which require computationally more demanding procedures than discriminant analysis and linear regression do. However, logistic regression models are not constrained by the assumption of normally distributed data (Sharma, 1996) and they model a probability; their output is a percentage term which is directly interpretable and usable to perform operational actions such as setting cut-off values. Due to these benefits, logit models have been adopted in this paper.

3.3 Explanatory Variable Treatment and Selection

Once data has been prepared and a statistical model has been chosen, the next step is to decide on the treatment of the explanatory variables. Afterwards, the explanatory variable selection process needs to be considered.

3.3.1 Treatment of Explanatory Variables

Several of the explanatory variables in a credit scoring context are typically categorical (e.g. *purpose*, *loan officer*, *beginning month*). According to Thomas (2000), there are two options to implement categorical variables in a scoring model. First, a binary (dummy) variable can be created for each possible category of an explanatory variable. Such implementation permits the modeling of non-linear behavior. This approach is often used in credit scoring models and is also adopted in model 1 of this paper (see section 4.2). Crook et al. (1992) note that such a dummy approach can considerably reduce the degrees of freedom available in the model. In addition, near-singularity problems might arise for dummy coded variables when executing the logistic regression calculations, as happened for the variables *branch* and *loan officer* in this study. Furthermore, dummy coded variables risk to overfit the data on which the model is built as a coefficient is created for each category present in the data set, independent of representativity of the actual category. Therefore, another approach, weight of evidence (WoE), works with one variable for all categories of an explanatory variable (Crook et al., 1992; Thomas, 2000; Hand and Henley, 1997). With b_i defined as the number of defaulted loans that belong to the i-th group, g_i is defined similarly for the non-defaulted loans. B and G are the total number of defaulted and non-defaulted loans present in the whole sample, defined as:

$$B = \sum_{i=1}^{n} b_i \text{ and } G = \sum_{i=1}^{n} g_i \tag{3.2}$$

In the first WoE step, to avoid overfitting, categories are put together in n groups based on similarity of $g_i/(g_i + b_i)$. Next, each of the newly created n groups receives a coding based on its distribution of defaulted and non-defaulted loans. Hence, every weight of evidence variable is composed of n values, one for each of the groups. Boyle et al. (1992) propose different implementations for the coding procedure: (i) g_i/b_i, (ii) $g_i/(g_i + b_i)$,

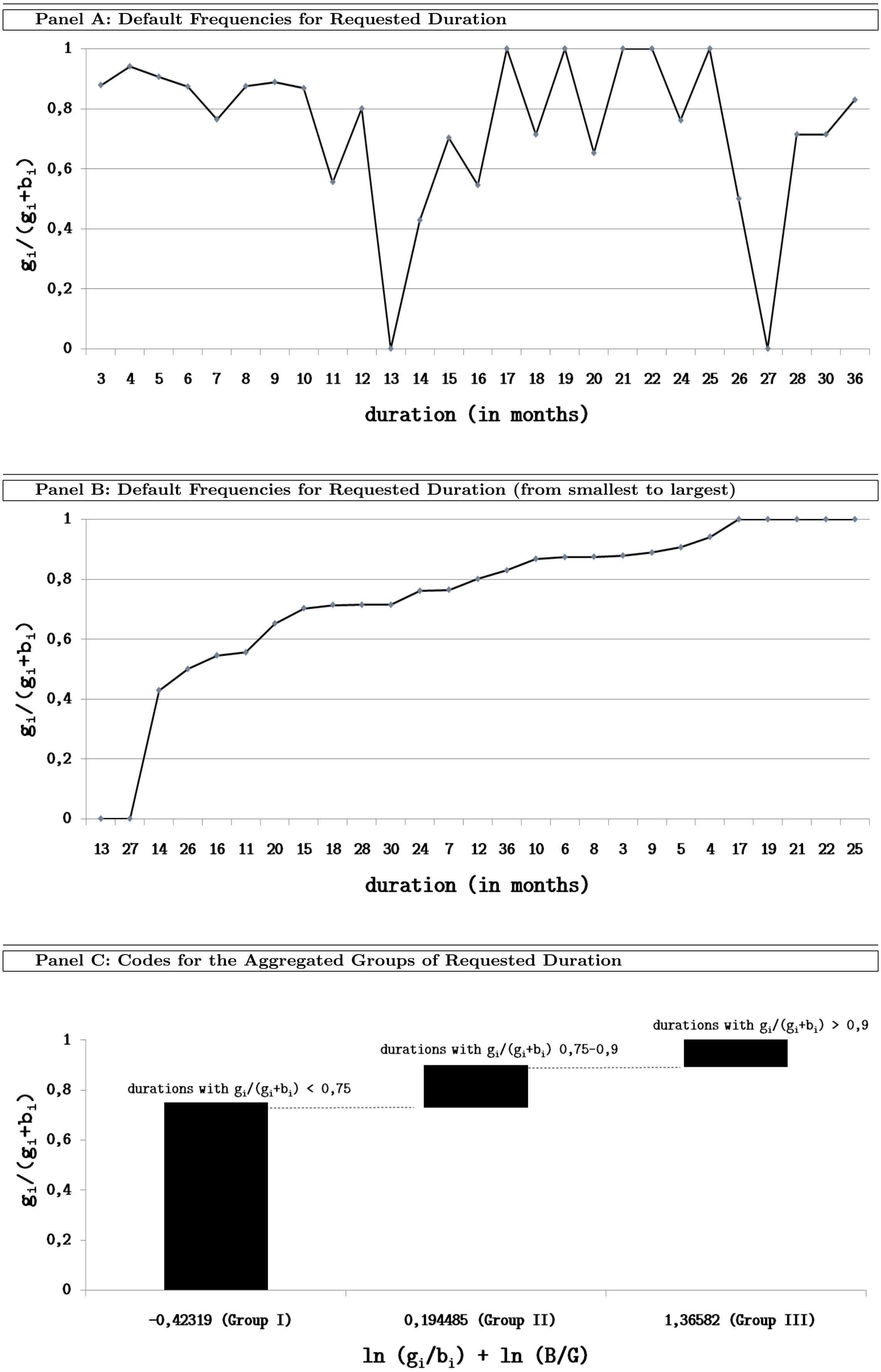
(iii) $b_i/(g_i + b_i)$, (iv) $\log(g_i/(g_i + b_i))$, (v) $\ln(g_i/b_i) + \ln(B/G)$. This study follows Crook et al. (1992) and Kleimeier and Dinh (2007) in the adoption of the fifth alternative for model 2 (see section 4.2). A potential weakness of the weight of evidence approach is that the explanatory variable coding is based on the dependent variables, which might cause overfitting on the sample data and result in inferior performance when tested out-of-sample.

Also for continuous explanatory variables (e.g. *age* and *amount*), there are different treatment options. First, the variable can be modeled as a linear straight line or as a more complex, e.g. quadratic, curve (Thomas, 2000). Secondly, as Van Gestel et al. (2005) describe, one might need to transform continuous variables in order to improve model fit and normality. Thirdly, another option is to convert the continuous variables into categorical ones (Thomas, 2000). Such conversion makes sense as this might allow to capture non-monotonous patterns, possibly resulting in a better discrimination among defaulted and non-defaulted loans (Crook et al., 1992; Boyle et al., 1992). Panel A of figure 1 presents the variable *requested duration* which follows a non-monotonous pattern. A first approach to discretize continuous variables is based on expert knowledge and is used in model 1 of this paper. The continuous variables are broken down in categories based on expert experience present among the microlender staff, and a dummy variable is created for each category. Another approach to convert continuous into categorical variables is statistically based. The categorization of a continuous variable can be chosen so that the default risk in the created categories is as homogeneous as possible. Based on similarity of $g_i/(g_i + b_i)$, the individual values of the continuous variable can be grouped (Crook et al., 1992). Grouping values together must be done such that the aggregated values appear sufficiently often in the data set in order to obtain statistically robust results (Boyle et al., 1992). This aggregation process creates groups for each originally continuous variable and is applied for model 2 presented in this paper. Consequently, following the example of Crook et al. (1992), the weight of evidence approach as described above for categorical variables is applied on the converted variables. Figure 1 presents a graphical illustration of the weights of evidence categorization approach for the variable *requested duration*.

3.3.2 Selection of Explanatory Variables

Including all variables would make the model unnecessarily large and deter clients when confronted with the required number of questions. Therefore authors typically adopt explanatory variable selection.

Hand and Henley (1997) describe three approaches on this matter. First, expert knowledge can be used to select the right variables. Secondly, statistical procedures as the forward and backward selection based on R^2 can be implemented. A combination of the forward and backward approaches, the stepwise approach, also exists. As a third



Panel A: Default Frequencies for Requested Duration

Panel B: Default Frequencies for Requested Duration (from smallest to largest)

Panel C: Codes for the Aggregated Groups of Requested Duration

Figure 1: Weight of Evidence Categorization Approach Applied on the Continuous Variable Requested Duration.

approach, Henley and Hand propose to select variables by using a measure which indicates the difference between the distributions of the defaulted and non-defaulted loans on that variable. Other authors, such as Verstraeten and Van den Poel (2005) also refer to the importance of the Receiver Operating Characteristic (ROC) Curve and its summary index Area Under the ROC Curve (AUC) in the explanatory variable selection process. The ROC Curve gives a graphical representation of the discriminatory power of a scoring system. ROC and AUC are further explained in the next section. Baesens et al. (2009) propose a heuristic variable selection procedure, based on AUC, which removes in each consecutive step the variable which causes the smallest decrease in AUC. Based on an expert decision, the trade-off between strong AUC performance and number of variables is established.

In this paper, both the stepwise selection procedure and the AUC heuristic selection procedure are employed and results are compared.

3.4 Validation

In line with Van Gestel et al. (2006), three main requirements are considered for the validation of a credit scoring model: stability, readability and discriminatory power.

- **Stability.** A stable model requires well determined coefficients with high confidence and similar results on performance characteristics if tested in- and out-of-sample.
- **Readability.** A model is said to be readable when its coefficients can be interpreted easily.
- **Discriminatory Power.** This is defined by the Basel Committee on Banking Supervision (2005) as the ability to correctly rank observations on the basis of default probability by assigning scores.

Stability is measured in two ways. First, we impose that p-values need to be below 5% and preferably below 1% for all coefficients included in the final models (2006). Secondly, a performance measure such as AUC needs to score similarly on in- and out-of-sample tests (Tasche, 2005).

The readability measure consists of a comparison between the a priori expected and the estimated sign of coefficient. This approach can also be found in Cantor and Packard (1996) and Van Gestel et al. (2006). The more the signs differ, the less readable the model is said to be. The a priori expected signs are based on intuition of the Bosnian microlender staff.

To test for discriminatory power, several measures are employed to assess the binary logit models presented in this paper. Tables 6 and 7 in the appendix present the Percentage Correctly Classified (PCC), Sensitivity (SENS), Specificity (SPEC), Kolmogorov-Smirnov (KS), AUC and Accuracy Ratio (AR) measures for binary logit models. The ROC and Cumulative Accuracy Profile (CAP) curves are also presented. When analyzing the strengths and weaknesses of the different measures (see right column of tables 6 and 7), AUC and AR come out as the most comprehensive measures to assess the discriminatory power of binary logit models. To the best of our knowledge, these two measures have not been reported before in credit scoring studies for microfinance. As credit scoring is in the first place about optimal discrimination between good and bad loans rather than description of a sample, these two measures will be used as the principal indicators of quality for the binary logit models presented.

4 Data and Model Description

This section presents the data set and data features and describes the constructed scoring models.

4.1 Data Set and Features

The data covers the period from June 2001 to November 2008. In total 6722 finished, individual loans are included in the data set. As recommended by Tasche (2005), a division in a 70% training set (4705 entries) and a 30% test set (2017 entries) is used for out-of-sample validation of the discriminatory power of the models. Following the risk definitions provided in section 3.1, $21,7\%$ of all loans is considered bad when risk is defined as a binary (good or bad) variable. This is a relatively high percentage due to the strict risk definitions employed at the cooperating microlender.

It should be remarked that the analysis in this paper is based on clients with approved loans. No generalizations can be made for a random sample of Bosnian micro-entrepreneurs as such a sample would also include rejected applicants, for which the behavior if they had been accepted is unknown. The problem of obtaining the default risk profile of rejected applicants is a well-documented problem called reject inference. Hand and Henley (1993) conclude that reliable reject inference is generally impossible and also Crook and Banasik (2004) reaffirm that useful reject inference depends on multiple parameters. As reject inference has not been dealt with in this study, the models developed in this paper can only be applied to those borrowers who also have been approved under the microlender's standard loan approval process.

In total 16 variables are considered for this study; several of these can also be found in other microfinance scoring models such as Schreiner (2004) and Vigano (1993). The variables can be grouped in three main categories: borrower, loan and lender characteristics.

Table 2 presents all variables with their Greek reference symbols, their categories, and a description of each variable. The categories column is only applicable for model 1, which creates a dummy variable for each variable category (see section 4.2). Concerning the description column of table 2, it should be noted that the expected effect of the variable on default risk is included based on intuition of the Bosnian microlender staff.

Overview of Variables		
Variable Name (Part of Model 1 and/or 2)	Variable Categories (only applicable for Model 1)	Description of Variable (including Bosnian microlender staff expectations of default risk influencing behavior)
Borrower Characteristics		
β_1 = Age (1, 2, 3)	20-24 (β_{11}), 25-34 (β_{12}), 35-49 (β_{13}), 50-64 (β_{14}), $\geq$ 65 (β_{15})	Age of applicant in years. Older applicants are expected to have a default risk-reducing effect as they would be more risk adverse, empirically confirmed by Boyle et al. (1992).
β_2 = Job Experience (1, 2, 3)	Unknown (β_{21}), 0 (β_{22}), <1 (β_{23}), 1-2 (β_{24}), 3-9 (β_{25}), 10-14 (β_{26}), $\geq 15(\beta_{27})$	Job experience of applicant in years. More experience is expected to have a decreasing effect on default risk as this would indicate more stability.
β_3 = Net Earnings of Business (1, 2, 3)	0 (β_{31}), 1-299 (β_{32}), 300-699 (β_{33}), 700-1499 (β_{34}), $\geq 1500(\beta_{35})$	Estimate by loan officer (LO) of net earnings of business in BAM on a monthly basis if loan would be disbursed. Higher net earnings are expected to have a decreasing effect on default risk as more room for repayment is available.
β_4 = Business Capital (1, 2, 3)	0 (β_{41}), 1-999 (β_{42}), 1000-4999 (β_{43}), 5000-14999 (β_{44}), 15000-49999 (β_{45}), $\geq 50000(\beta_{46})$	Estimate by LO of value of applicant's business in BAM (total assets − debt). No consensus could be reached among the Bosnian microlender staff on the default risk effect of this variable.
β_5 = Business Register (1, 2, 3)	Yes (registered) (β_{51}), No (unregistered) (β_{52})	Specifies if applicant business is officially registered. Yes is expected to have a decreasing effect on default risk as this generally indicates more stability.
β_6 = Net Earnings of Household (1, 2, 3)	1-299 (β_{61}), 300-699 (β_{62}), 700-1499 (β_{63}), $\geq 1500(\beta_{64})$	Estimate by LO in BAM of family's free cash flow. Higher net earnings are expected to have a decreasing effect on default risk as more room for repayment is available.
β_7 = Household Capital (1, 2, 3)	0-4999 (β_{71}), 5000-19999 (β_{72}), 20000-49999 (β_{73}), 50000-99999 (β_{74}), $\geq 100000(\beta_{75})$	Estimate of total value of applicant's household (assets-debt) in BAM. No consensus could be reached among the Bosnian microlender staff on the default risk effect of this variable.
β_8 = Other Debt (1, 2, 3)	0 (β_{81}), 1-199 (β_{82}), 200-999 (β_{83}), 1000-4999 (β_{84}), 5000-19999 (β_{85}), $\geq 20000(\beta_{86})$	Estimate by LO of size of other loans (in BAM) taken up by applicant. As other loans reduce room for repayment, this variable is expected to have a default risk-increasing effect.
Loan Characteristics		
β_9 = Purpose (1, 2, 3)	Trade (β_{91}), Manufacturing (β_{92}), Household (β_{93}), Services (β_{94}), Merchandise (β_{95}), Agriculture (β_{96})	Gives destination of micro-loan. Merchandise is the category destined for buying business and household equipment at partner organizations of the microlender. Services and trade are assumed to be the default risk-increasing categories due to their inherent volatility, while agriculture is assumed to be the most safe category due to higher social control and typical lower volatility.
β_{10} = Amount (1, 2, 3)	<1000 (β_{101}), 1000-1999 (β_{102}), 2000-2999 (β_{103}), 3000-4999 (β_{104}), 5000-9999 (β_{105}), $\geq 10000(\beta_{106})$	Measures size of loan requested in BAM. As the incentive to deviate increases for bigger loans, an increasing default risk effect is expected.
β_{11} = Requested Duration (1, 2, 3)	1-12 (β_{111}), 13-18 (β_{112}), 19-24 (β_{113}), 25-60 (β_{114})	Describes requested duration of loan in months. A longer duration is expected to signal insufficient short term capacity or to be associated with higher uncertainty about future solvability, causing an increasing default risk effect.
β_{12} = Cycles (1, 2, 3)	1 (β_{121}), 2 (β_{122}), 3 (β_{123}), 4-5 (β_{124}), 6-12 (β_{125})	Indicates history of applicant at lender. 1 indicates that no other loans have been disbursed before. Expected decreasing effect on default risk as a repeat disbursement is seen as a quality stamp.
β_{13} = Beginning Month (1, 2, 3)	January - December (β_{131} - β_{1312})	Description of month in which loan application was filed. Bosnian microlender staff expects winter months to have increasing default risk effect as more unforeseen circumstances can take place.
β_{14} = Year of Initiation (1, 2, 3)	2001-2008 (β_{141} - β_{148})	Indication of year in which loan application was filed. Earlier years are expected to have an increasing default risk effect as less experience among lender staff is assumed.
Lender Characteristics		
β_{15} = Branch (2)	28 branches over Bosnia-Herzegovina	Indicates branch in which loan application was filed. Rurally located branches are expected to have a decreasing default risk effect due to more social control.
β_{16} = Loan Officer (2)	95 loan officers	Indicates name of officer who filed loan application and is primarily used as a proxy for loan officer experience. Loan officers with several years of experience in the organization are expected to have a default risk-decreasing effect.

Table 2: Overview of Variables. BAM stands for Bosnian Convertible Mark, the pegged currency of Bosnia-Herzegovina (1,95583 BAM = 1 Euro).

Clarifications for the expectations are also included. Due to the application of the non-intuitive weight of evidence procedure, it makes no sense to interpret the expected default risk effect for the variables of model 2.

4.2 Model Description

Two models were developed to analyze whether credit scoring is applicable for microlenders: a binary logit model based on dummy coded variables and a binary logit model based on weight of evidence coded variables.

General Binary Logit Model (Model 1)

$$\pi_i = E(Y_i = 1) = \frac{1}{1 + e^{-logit_i}}, \tag{4.1}$$

$$logit_i = \beta_0 + \beta_{11}x_{i11} + \beta_{12}x_{i12} + ... + \beta_{148}x_{i148} \tag{4.2}$$

- Y_i: binary dependent variable (see equation 3.1)
- β_0: intercept, β_i: regression coefficient (see table 2)
- x_i: dummy coded explanatory variable (see section 3.3.1)

Model 1 is based on a logistic regression, with a binary (good/bad) input as dependent variable. 14 out of the 16 variables in the data set are included as candidate explanatory variables (see table 2).

As section 3.3.1 describes, there are several options for the treatment of the explanatory variables. For the continuous explanatory variables, model 1 chooses to make the transformation towards categorical variables. The grouping needed for this transformation is based on expertise of the Bosnian microlender staff. For the categorical explanatory variables, model 1 opts for an approach also taken by Schreiner (2004) in his Bolivian study: creation of a dummy variable for each variable category. Concerning variable selection, two versions of model 1 are created: one based on stepwise selection, the other one based on the heuristic AUC selection procedure (see section 3.3.2). The validation of model 1 is based on an out-of-sample approach with a 70% training and a 30% test set.

Binary Logit Model with Weight of Evidence Coding (Model 2)

$$\pi_i = E(Y_i = 1) = \frac{1}{1 + e^{-logit_i}}, \tag{4.3}$$

$$logit_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{16}x_{i16} \tag{4.4}$$

- Y_i: binary dependent variable (see equation 3.1)
- β_0: intercept, β_i: regression coefficient (see table 2)
- x_i: weight of evidence coded explanatory variable (see section 3.3.1)

Model 2 is very similar to model 1: it also employs logistic regression, it has a binary coded dependent variable, it employs the same two explanatory variable selection techniques, and uses the same validation measures as model 1.

The main discrepancy between model 1 and 2 is related to the explanatory variables. Model 2 is able to use all 16 explanatory variables available in the data set. While near-singularity problems were reported when converting the categories of *branch* and *loan officer* into dummy variables for model 1, these problems did not occur when adopting the weight of evidence procedure for model 2. Both categorical and continuous explanatory variables pass through the weight of evidence procedure as explained in section 3.3.1 and illustrated in figure 1. The objective of this approach is to obtain results with higher discriminatory power as more variables are taken into account, and to avoid the potential overfitting risk inherent to the dummy procedure.

5 Empirical Results

Section 5.1 presents the estimation output of the models, by describing general tendencies and comparing the two models. Section 5.2 discusses validation aspects.

5.1 Model Estimation

To estimate the two logistic regression models, we first need to select the explanatory variables based on the stepwise and the AUC heuristic approach (see the illustration in graph 2). Based on the results of the selection procedures, presented in tables 3 and 4, the following two general tendencies are observed:

- Five variables (*other debt*, *purpose*, *requested duration*, *amount* and *beginning month*) are significant according to all selection approaches, for both models. Apart from the variable *other debt*, the values of these variables are not based on estimates by the Bosnian loan officers. Their strong significance indicates an important role for this type of variables in the prediction of a loan applicant's repayment capacity.

- Two variables (*net earnings of business* and *business register*) are never significant for any of the selection approaches in one of the two models. Other indicators of the financial position of the borrower (*business capital*, *net earnings of household* and *household capital*) are only significant in model 2. These observations are most likely attributable to a combination of poor accuracy of the estimates made by loan officers about the financial position of clients, a weak relationship between the dependent variables and the explanatory variables, and improper categorizations of the variables. As the categorizations were made in cooperation with the microfinance institution, the weak performance of these variables might suggest the need for new, statistically-based, categorizations for the variables in addition to the practical intuition of the microlender staff.

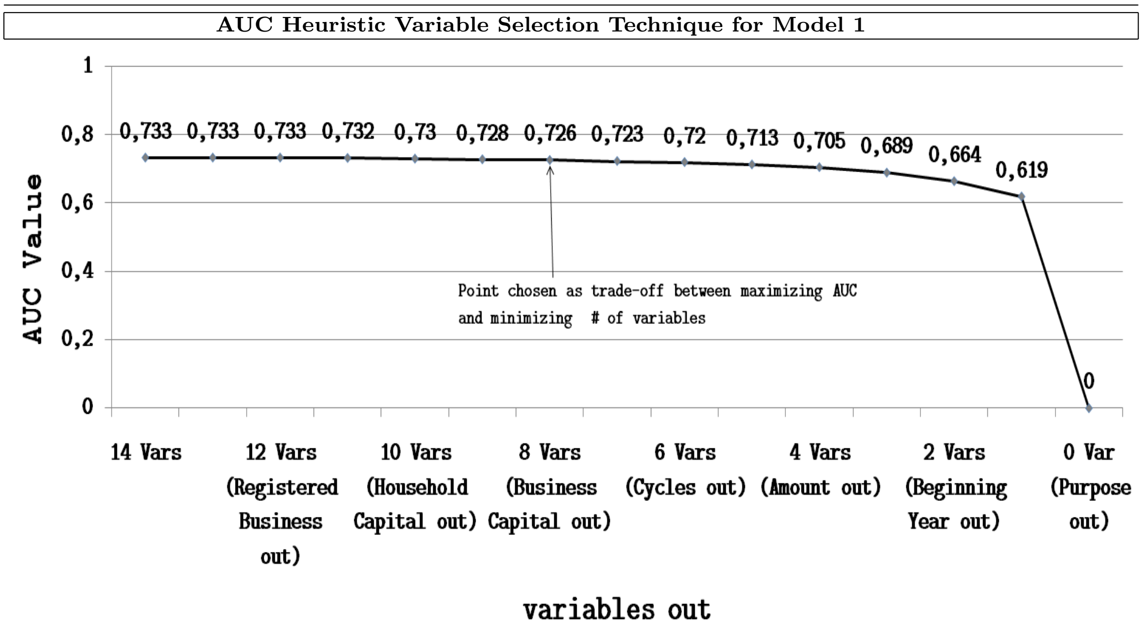


Figure 2: AUC Heuristic Variable Selection Technique for Model 1

When comparing the estimated models 1 and 2, three observations arise. First, the interpretation of model 2 is different from model 1. While model 2 uses a non-intuitive construction of its explanatory variables based on the weights of evidence approach, model 1 uses a more intuitive approach with a dummy variable for each category (see section 3.3.1). Due to this difference, it makes no sense to compare the coefficients of the two models. Another aspect of non-comparability concerns the p-values. While the p-values presented in table 4 determine the statistical significance for the variable as a whole, the p-values in table 3 determine the significance of the category versus the reference category (presented as the last dummy for each variable). Therefore, a direct comparison of the p-values is not possible. A second observation is that three variables (*business capital*, *net earnings of household*, *household capital*), which are not significant in model 1, appear to be significant in model 2. Also two variables (*cycles* and *year of initiation*) which are not significant in model 2 are significant in model 1. These differences indicate that complementary opportunities might exist to create a model which combines best of both worlds. Finally, model 2 contains two variables (*branch* and *loan officer*) which could not be included in model 1 as they caused near-singularity problems when each variable category was treated as a dummy variable. Again, the adoption of these two variables via the approach used in model 2 indicates complementary opportunities to create a best of both worlds model.

In conclusion, model 1 and 2 reveal similar information for five variables, while for the other ones complementary opportunities might exist between the two models.

5.2 Model Validation

Subsection 3.4 structured the validation of a model according to three principles: stability, readability and discriminatory power. The above two models will be evaluated on these three criteria.

Stability. Stability is measured in two ways. First, p-values of coefficients included in the final versions of the models studied are below 5% and in many cases even below 1% significance level based on the stepwise approach. A second explanatory variable selection approach was also applied: the AUC heuristic approach. The selected variables are presented in tables 3 and 4 and discussed above. A second way of measuring stability is comparing the in and out-of-sample performance on the most important measure, AUC. As presented in table 5, the in- and out-of-sample performances for model 1 when adopting the stepwise and AUC heuristic approaches are in line with each other. Model 2 posts in- and out-of-sample results which are diverging. The cause of this divergence for 2 model is probably overfitting of the model on the sample training set. Especially as the weight of evidence coding procedure of model 2 has avoided creating more than three codes per explanatory variable (see graph 1), this result points out an unexpected weakness of model 2.

Readability. The readability performance of the models is measured via the comparisons of the (a priori) expected and estimated signs in table 3. Due to the non-intuitive character of the coefficients of model 2, it makes no sense to include expected signs for this model. For model 1, the expected sign column makes sense because an intuitive interpretation is possible. The expected sign of a variable category should be interpreted relative to the other categories of the same variable. A plus (or double plus) sign indicates that the category of a certain variable is expected to have a (strong) risk-increasing effect compared to the other categories of the same variable with a minus (or double minus). The single and double sign expectations were determined by the domain experts of the Bosnian microlender. For two variables (*business capital*, *household capital*) no clear expectations existed. Five out of the 14 candidate variables (*age*, *business register*, *amount*, *cycles*, *beginning month*) report an estimated sign which is completely in line with the expectations. This proves that the staff intuition for these variables is correct. For 4 variables (*job experience*, *net earnings of business*, *net earnings of household*, *other debt*) small differences between expected and estimated signs are observed. In the case of *job experience* this difference forms a surprise and gives a new insight on the risk exposure. The small differences for the two net earnings variables and *other debt* are most likely attributable to the use of absolute values for these variables, without any correction for the (business

Category Sizes, Estimated Coefficients, Standard Errors and Significance Levels for Model 1

Variable		Category Size (%)	E.S.	Coeff.	S.E.	p-value
Intercept β_{10}				-2.4394	0.6120	$<.0001$
Age β_1	20-24 β_{11}	3,85%	++	0.5575	0.3186	0.0801^{2}
	25-34 β_{12}	23,58%	+	0.2502	0.2633	0.3419^{2}
	35-49 β_{13}	43,83%	-	0.1992	0.2565	0.4373^{2}
	60-64 β_{14}	25,97%	–	0.0385	0.2607	0.8827^{2}
	$\geq$ 65 β_{15}	2,77%	–	0.0000	N/A	N/A
Job Experience β_2	unknown β_{21}	1,01%	+	0.3782	0.7571	0.6174^{3}
	0 β_{22}	5,64%	++	0.1547	0.3202	0.6291^{3}
	<1 β_{23}	1,86%	++	0.6440	0.2830	0.0228^{3}
	1-2 β_{24}	17,30%	+	0.2955	0.1597	0.0642^{3}
	3-9 β_{25}	44,09%	-	0.2280	0.1385	0.0996^{3}
	10-14 β_{26}	17,54%	–	0.0512	0.1549	0.7410^{3}
	$\geq$ 15 β_{27}	12,56%	–	0.0000	N/A	N/A
Net Res. of Bus. β_3	0 β_{31}	9,91%	++	-0.9469	0.6972	0.1744
	1-299 β_{32}	11,71%	+	-0.4107	0.2351	0.0806
	300-699 β_{33}	41,03%	-	-0.4751	0.1787	0.0078
	700-1499 β_{34}	22,55%	–	-0.2009	0.1504	0.1818
	$\geq$ 1500 β_{35}	14,80%	–	0.0000	N/A	N/A
Business Capital β_4	0 β_{41}	9,30%	N/A	-0.2800	0.9688	0.7725
	1-999 β_{42}	1,31%	N/A	0.8750	0.4632	0.0589
	1000-4999 β_{43}	18,12%	N/A	0.3144	0.1804	0.0813
	5000-14999 β_{44}	28,41%	N/A	0.0121	0.1540	0.9375
	15000-49999 β_{45}	28,44%	N/A	-0.0387	0.1329	0.7709
	$\geq$ 50000 β_{46}	14,42%	N/A	0.0000	N/A	N/A
Business Register β_5	Yes (Reg) β_{51}	67,09%	-	-0.1077	0.0911	0.2372
	No (Unreg) β_{52}	32,91%	+	0.00000	N/A	N/A
Net Res. of Househ. β_6	1-299 β_{61}	9,09%	++	0.0512	0.2585	0.8430
	300-699 β_{62}	43,11%	+	0.0582	0.1798	0.7462
	700-1499 β_{63}	32,79%	-	0.1309	0.1503	0.3838
	$\geq$ 1500 β_{64}	15,01%	–	0.0000	N/A	N/A
Household Capital β_7	0-4999 β_{71}	3,17%	N/A	0.7828	0.2757	0.0045
	5000-19999 β_{72}	4,34%	N/A	0.1509	0.2186	0.4899
	20000-49999 β_{73}	32,77%	N/A	0.1904	0.1365	0.1632
	50000-99999 β_{74}	47,25%	N/A	0.0895	0.1222	0.4636
	$\geq$ 100000 β_{75}	12,47%	N/A	0.0000	N/A	N/A
Other Debt β_8	0 β_{81}	77,34%	–	-1.2369	0.2740	$<.0001^{2,3}$
	1-199 β_{82}	1,55%	-	-0.6790	0.3840	$0.0770^{2,3}$
	200-999 β_{83}	6,72%	+	-0.7775	0.2962	$0.0087^{2,3}$
	1000-4999 β_{84}	8,27%	+	-0.5645	0.2829	$0.0460^{2,3}$
	5000-19999 β_{85}	4,36%	++	-0.4637	0.2887	$0.1082^{2,3}$
	$\geq$ 20000 β_{86}	1,76%	++	0.0000	N/A	N/A
Purpose β_9	Agriculture β_{91}	12,17%	–	-0.3011	0.1543	$0.0511^{2,3}$
	Household β_{92}	9,00%	+	1.1243	0.8299	$0.1755^{2,3}$
	Manufacturing β_{93}	34,48%	-	-0.2118	0.1189	$0.0747^{2,3}$
	Merchandise β_{94}	0,79%	-	-0.0158	0.8008	$0.9843^{2,3}$
	Services β_{95}	23,45%	++	0.2452	0.1091	$0.0246^{2,3}$
	Trade β_{96}	20,11%	++	0.0000	N/A	N/A
Amount β_{10}	<1000 β_{101}	7,69%	–	-1.2504	0.3345	$0.0002^{2,3}$
	1000-1999 β_{102}	37,59%	–	-1.0571	0.2532	$<.0001^{2,3}$
	2000-2999 β_{103}	16,59%	-	-0.8035	0.2394	$0.0008^{2,3}$
	3000-4999 β_{104}	18,19%	+	-0.5977	0.2134	$0.0051^{2,3}$
	5000-9999 β_{105}	12,94%	++	-0.4861	0.1872	$0.0094^{2,3}$
	$\geq$ 10000 β_{106}	6,99%	++	0.0000	N/A	N/A
Requested Duration β_{11}	1-12 β_{111}	60,76%	–	0.6739	0.2431	$0.0056^{2,3}$
	13-18 β_{112}	26,44%	-	1.0863	0.2311	$<.0001^{2,3}$
	19-24 β_{113}	8,91%	+	0.7113	0.2402	$0.0031^{2,3}$
	25-60 β_{114}	3,90%	++	0.0000	N/A	N/A
Cycles β_{12}	1 β_{121}	67,99%	++	1.0809	0.3853	$0.0050^{2,3}$
	2 β_{122}	19,15%	-	0.9780	0.3857	$0.0112^{2,3}$
	3 β_{123}	6,81%	-	0.9508	0.4038	$0.0185^{2,3}$
	4-5 β_{124}	4,33%	–	0.4799	0.4222	$0.2556^{2,3}$
	6-12 β_{125}	1,73%	–	0.0000	N/A	N/A
Beginning Month β_{13}	January β_{131}	6,81%	+	-0.0260	0.1906	$0.8915^{2,3}$
	February β_{132}	8,35%	+	0.0794	0.1738	$0.6477^{2,3}$
	March β_{133}	9,65%	+	-0.0723	0.1673	$0.6657^{2,3}$
	April β_{134}	8,82%	+	-0.0156	0.1700	$0.9269^{2,3}$
	May β_{135}	8,79%	-	-0.2082	0.1768	$0.2390^{2,3}$
	June β_{136}	8,66%	-	-0.4914	0.1847	$0.0078^{2,3}$
	July β_{137}	7,80%	-	-0.3952	0.1836	$0.0313^{2,3}$
	August β_{138}	7,57%	-	-0.5385	0.1907	$0.0047^{2,3}$
	September β_{139}	7,81%	-	-0.2279	0.1793	$0.2038^{2,3}$
	October β_{1310}	7,94%	+	-0.1273	0.1742	$0.4649^{2,3}$
	November β_{1311}	8,27%	+	-0.0425	0.1704	$0.8031^{2,3}$
	December β_{1312}	9,52%	+	0.0000	N/A	N/A

Category Sizes, Estimated Coefficients, Standard Errors and Significance Levels for Model 1 (continued)

Variable		**Category Size (%)**	**E.S.**	**Coeff.**	**S.E.**	**p-value**
Year of Initiation β_{14}	2001 β_{141}	0,80%	+	0.3203	0.8180	0.6954[2,3]
	2002 β_{142}	3,51%	+	0.8483	0.3458	0.0142[2,3]
	2003 β_{143}	7,59%	+	1.3830	0.2882	<.0001[2,3]
	2004 β_{144}	8,52%	-	1.5146	0.2853	<.0001[2,3]
	2005 β_{145}	24,28%	-	1.2986	0.2663	<.0001[2,3]
	2006 β_{146}	23,09%	-	1.1440	0.2645	<.0001[2,3]
	2007 β_{147}	25,42%	-	0.7916	0.2584	0.0022[2,3]
	2008 β_{148}	6,78%	-	0.0000	N/A	N/A

Table 3: Category Sizes, Estimated Coefficients, Standard Errors and Significance Levels for Model 1. E.S. is the a priori expected sign, Coeff. is the coefficient, S.E. is standard error, p-value refers to significance of the respective category versus the reference category. [1] refers to 5% significance of the whole variable based on the stepwise approach, [2] to 1% significance on the stepwise approach, [3] to selection of the variable based on the AUC heuristic approach. See section 3.3.2 for more information on the explanatory variable selection techniques.

Estimated Coefficients, Standard Errors and Significance Levels for Model 2

Variable	**Coefficient**	**S.E.**	**p-value**
Intercept	-1.2699	0.0402	<.0001
Age	-0.9074	0.1684	<.0001[2,3]
Job Experience	-0.5602	0.1369	<.0001[2,3]
Net Earnings of Business	-0.0614	0.0962	0.5234
Business Capital	-0.4839	0.0934	<.0001[2,3]
Business Register	0.0582	0.1294	0.6532
Net Earnings of Household	-0.3530	0.0949	0.0002[2,3]
Household Capital	-0.6946	0.0963	<.0001[2,3]
Other Debt	-0.5619	0.0794	<.0001[2,3]
Purpose	-0.3329	0.0966	0.0006[2,3]
Amount	-0.1813	0.0930	0.0512[1]
Requested Duration	-0.8793	0.1267	<.0001[2,3]
Cycles	-0.1555	0.3162	0.6229
Beginning Month	-0.9140	0.2255	<.0001[2,3]
Year of Initiation	-0.1169	0.1161	0.3141
Branch	-0.3012	0.1355	0.0262[1,3]
Loan Officer	-0.8044	0.0842	<.0001[2,3]

Table 4: Estimated Coefficients, Standard Errors and Significance Levels for Model 2. S.E. is standard error, p-value refers to significance of the respective category versus the reference category. [1] refers to 5% significance of the whole variable based on the stepwise approach, [2] to 1% significance on the stepwise approach, [3] to selection of the variable based on the AUC heuristic approach. See section 3.3.2 for more information on the explanatory variable selection techniques.

or household) capital size involved, loan amount or purpose of the loan. For 3 variables (*purpose*, *requested duration*, *year of initiation*), serious divergences between expected and estimated signs are observable. For *purpose*, it comes as a surprise that household is by far the most risky category. For *requested duration*, the most risky categories are the medium-term durations, rather than the long-term durations. The long-term duration on the other hand forms the category with the strongest risk-decreasing effect. For *year of initiation*, the assumption that in earlier years less experience would lead to more risk seems to be untrue. In conclusion, it can be stated that model 1 and 3 generally score well on the readability performance as most default risk expectations of the staff proved to be correct. However for a few variables, the readability performance was weak, which involves an opportunity for the microlender to learn about its risk exposure.

Discriminatory Power. Concerning discriminatory power, first the Kolmogorov-Smirnov (KS), PCC, SENS and SPEC measures should be analyzed. A valuable side-

Performance of Models 1 and 2								
Performance Measure	**Model 1 - STPW**		**Model 1 - AUC**		**Model 2 - STPW**		**Model 2 - AUC**	
	In-Sample	Out of Sample	In-Sample	Out of Sample	In-Sample	Out of Sample	In-Sample	Out of Sample
PCC(1)	0.7843	0.7848	0.7834	0.7823	0.8042	0.7769	0.8053	0.7754
SENS(1)	0.7978	0.8039	0.7985	0.8031	0.8201	0.8038	0.8197	0.8016
SPEC(1)	0.5294	0.4833	0,5197	0.4634	0.6343	0.4315	0.6452	0.4161
PCC (2)	0.6380	0.6386	0.6829	0.6857	0.6812	0.6371	0.6332	0.5964
SENS (2)	0.8816	0.8810	0.8747	0.8720	0.8979	0.8632	0.9052	0.8741
SPEC (2)	0.3410	0.3322	0.3707	0.3614	0.3805	0.3202	0.3484	0.3049
Kolmogorov-Smirnov Statistic	0.3394	0.3134	0.3416	0.3224	0.4003	0.2670	0.4025	0.2667
Area Under ROC Curve (AUC)	0.7257	0.7066	0.7257	0.7051	0.7685	0.6810	0.7675	0.6789
Accuracy Ratio (AR)	0.4514	0.4133	0.4514	0.4102	0.5370	0.3620	0.5351	0.3579
Number of Variables Included	8		8		12		11	

Table 5: Performance of Models 1 and 2. STPW refers to the model obtained after running the stepwise explanatory variable selection technique with 5% significance level. AUC refers to the model obtained after running the heuristic AUC explanatory variable selection technique. See section 4.2 for explanation of the variable selection techniques. (1) refers to the 50% cut-off for all four models, (2) refers to the 19.82% cut-off for Model 1-STPW, 22.23% for Model 1-AUC, 20.66% for model 2-STPW, 17.57% for model 2-AUC. These second cut-offs are optima determined by means of the Kolmogorov-Smirnov statistic graph, see graph 3 for an illustration.

effect of the KS statistic, as illustrated in figure 3, is that it can determine the cut-off value at the point where the KS distance is maximal. This cut-off value allows the calculation of PCC, SENS and SPEC at another level in addition to the standard 50% point. Noteworthy is the higher SENS value when adopting the KS-determined cut-off point rather than the standard 50% level. A high SENS value can be understood as few defaulting applicants receiving a loan, which is an important indication for the practical usefulness of a scoring system. The trade-off inherent to this higher SENS value is a lower SPEC value, which indicates more lost business opportunities. Overall, neither the KS nor the 50% cut-off levels succeeded in very strong PCC, SENS or SPEC results such as the 99% PCC performance report in the study of Kleimeier and Dinh (2007). Secondly, the AUC measure should be analyzed. This measure was selected as the most important indicator of predictive accuracy in section 3.4. Based on figure 4 and table 5, the observation is made that model 2 has the best in-sample performance, but model 1 the best out-of-sample performance. For model 1, the differences between the versions estimated via the stepwise or the AUC heuristic explanatory variable selection approach are marginal. For model 2, the AUC heuristic approach performs slightly better as it succeeds in creating a model with only 11 variables and an almost equal AUC value as the 12 variable model of the stepwise approach. Tasche (2005) remarks that credit scoring models often have an AUC value in the range $0.75 - 0.9$. Model 1 and 2 score in the lower range or even slightly below, which leads to the conclusion that in terms of predictive accuracy, credit scoring models for microfinance can not compete yet with traditional credit scoring models for credit cards, consumer loans etc.

In conclusion, model 1 performs best on the validation tests. It scores best on stability and discriminatory power while it posts satisfactory readability results. Model 2 has weak stability performance and satisfactory discriminatory power.

Kolmogorov-Smirnov Graph based Cut-Off Determination Illustrated for Model 1

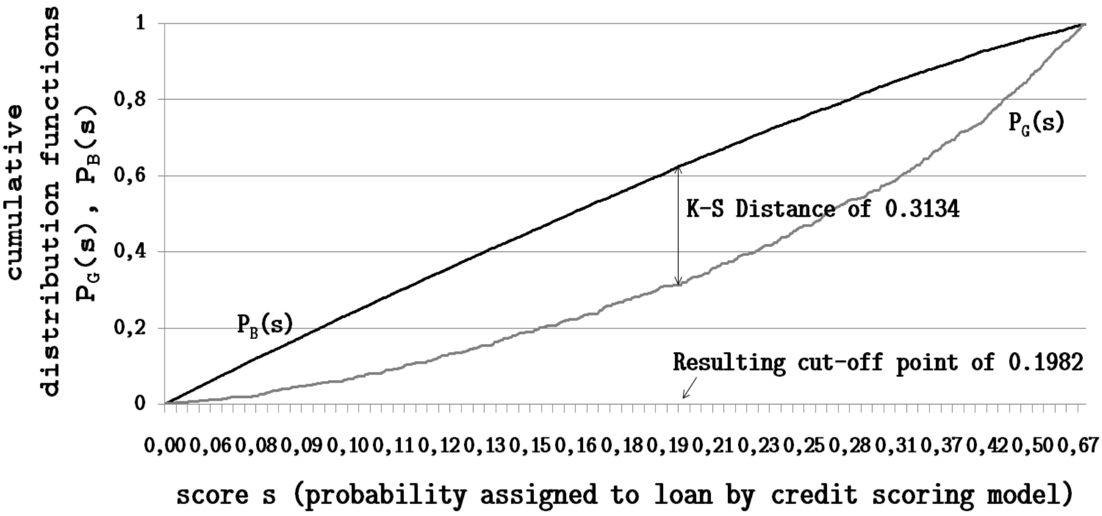


Figure 3: Kolmogorov-Smirnov Graph based Cut-Off Determination Illustrated for Model 1.

6 Discussion

This section discusses whether credit scoring should be adopted in microfinance institutions, in light of the above results.

6.1 Should Credit Scoring be Adopted in Microfinance?

Two hypotheses are linked to the central research question of this paper. The first hypothesis states that credit scoring can replace the traditional credit process. Strong stability, readability and discriminatory power of the scoring system are necessary to validate this statement. The results section clearly indicates model 1 as the best model with a strong stability and sufficient readability. Concerning discriminatory power however, even model 1, with an optimal out-of-sample AUC performance of 0.7066, is performing at the lower end of Tasche's (2005) range (0.75 – 0.9). Due to this limited discriminatory power performance, it would be impossible to replace the current credit processes at microlenders with an automated credit scoring system and the first hypothesis is thus rejected. The following two reasons for the limited discriminatory power are not only applicable on the particular Bosnian microlender studied here, but can be generalized to other microlenders:

- Limited data, both in number of loans included and quality of explanatory variables. Especially the variables concerning the financial position of the borrower seem to have a weak explanatory power, which is not surprising in light of their estimated and thus non-exact nature. While for traditional financial environments objective data is available, this is often not the case for microfinance.

ROC Performances for Model 1 and 2 (In and Out of Sample)

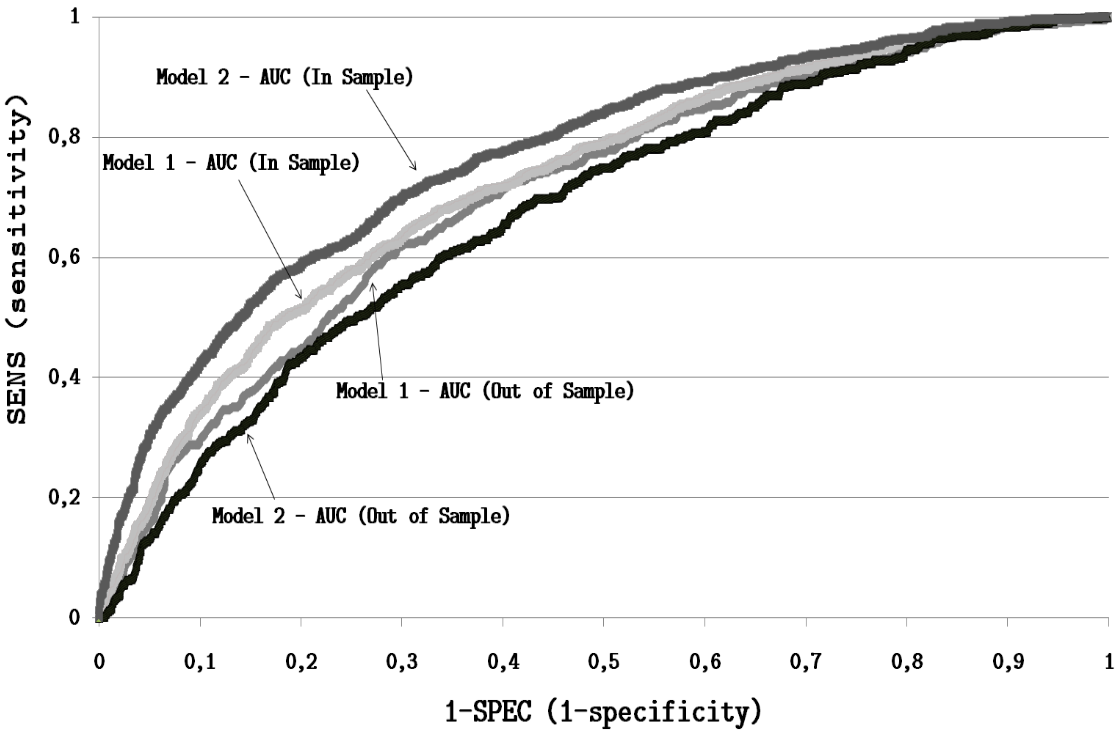


Figure 4: ROC Performances for Model 1 and 2 (In and Out of Sample). The ROC performances displayed are based on the versions of Model 1 and 2 when using the AUC heuristic explanatory variable selection approach.

- Significant uncontrollable risks are inherent to microfinance. For example, 32.91% of the loan applicants in the data set use the loan for black economy purposes, which is subject to possible fines. Also, interviews with the Bosnian microlender indicated that multiple clients change their occupation during the duration of the loan due to external circumstances (e.g. entrance of foreign department stores, weather circumstances). The case of a repaying borrower who started as a trader, then had to stop and rent out his building and went milking cows on the farm of a family member is illustrative.

Hypothesis 2 states that credit scoring can become a tool to refine the credit process of microfinance institutions. The requirements to confirm this hypothesis are similar to the ones of hypothesis 1, only the discriminatory power constraint is more relaxed. As the discriminatory power of optimal model 1 finds itself at the lower end of Tasche's (2005) range, hypothesis 2 can be confirmed. The introduction of credit scoring as a refinement tool rather than a panacea can be generalized to other microfinance institutions due to the following reasons:

- The loan officers in every microfinance institution should see the scoring system as a help rather than as a competitor. Loan officers can always manipulate a credit scoring system when they input the data. Especially as they might feel threatened by the introduction of such a system, it is crucial to make the scoring model a partner of the loan officer, allowing him or her to boost his performance.

- An unconstrained adoption of credit scoring causes a tunnel view towards lending. A scoring system is always based on the past and especially as microfinance clients are subject to significant uncontrollable risks, it would be unwise to exclude the human interaction part in the credit process.

6.2 Next Steps

Three key steps are of particular importance in the further development of credit scoring systems for microfinance. A first step concerns model optimization. As Thomas (2007) discusses, well-performing credit scoring models are often combinations of different models which are specifically tailored to a limited part of the population. In the case of this paper, an integration of the complementary parts of model 2 into model 1 could create improved discrimination power for the resulting model. Second, reject inference constitutes another major challenge. Even though it is often forgotten or briefly mentioned, the reject inference problem considerably limits credit scoring application to data sets of already approved loans. Advances in solving the reject inference problem for microfinance credit scoring purposes are definitely needed as this will help in transforming credit scoring for microfinance from a marked improvement into a real breakthrough. Finally, more credit scoring studies on European, Asian and even worldwide microfinance data sets are needed to confirm, generalize and refine the current insights. This would bring the knowledge on credit scoring for microfinance on a more equal footing with other credit scoring domains.

7 Conclusion

Three key findings are identified at the conclusion of this study, leading to practical recommendations for the microlender studied and other microfinance institutions.

1. The number of published credit scoring studies for microfinance is limited. There is no consensus as to the applicability of the concept – perhaps because many verdicts are heuristic rather than supported by quantitative evidence. Since no studies have been published for Eastern Europe-Central Asia and Middle East-North Africa, there is a need to extend the geographical reach of credit scoring studies towards these regions.

2. Credit scoring models in microfinance have the same ultimate goal as credit scoring models in other domains: optimal discrimination between good and bad loans. Therefore, best practices from other domains such as the weight of evidence coding

approach and the area under the ROC curve performance measure should be further introduced in microfinance credit scoring models.

3. The discriminatory power performance of credit scoring systems for microfinance remains too weak to justify a complete reversal of the traditional credit process towards scoring. However, credit scoring should become a refinement tool in the current process as it has already proven to be stable, easy to use, and also to have a certain discriminatory power. Improvements of the discriminatory power via model combinations, reject inference research, and more practical evidence, may gradually increase the role of credit scoring in the credit process.

A Appendix: Performance Measures and Graphs for Credit Scoring Models

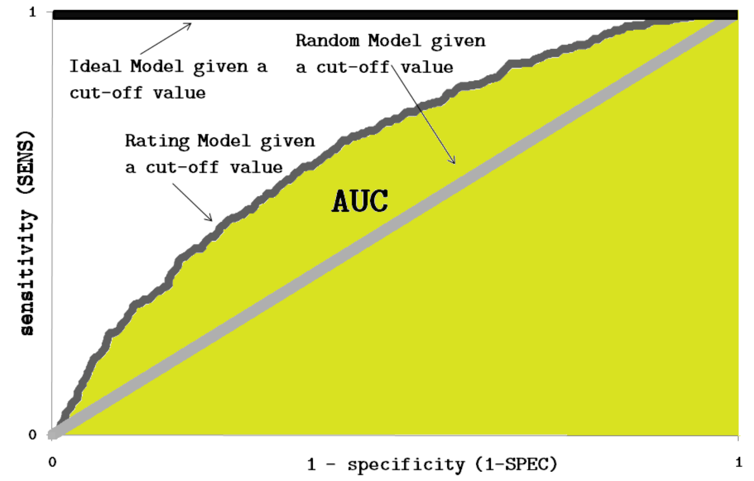
This appendix provides detailed overviews of the most important performance measures and graphs for binary logit credit scoring models. For each performance measure, a definition and a discussion of the strengths and weaknesses is provided. For the graphs, a presentation is given and the link with the performance measures is identified.

Performance Measures for Binary Logit Credit Scoring Models (Part I)		
Name (Source)	Definition	Strenghts and Limitations
Percentage Correctly Classified Observations, PCC (Kleimeier and Dinh, 2007; Provost et al., 1998)	PCC is a measure to assess the discriminatory power of a credit scoring model (range: 0% ≤ PCC ≤ 100%): $PCC = \frac{(TP+TN)}{(TP+FN+FP+TN)} \quad \text{(A.1)}$ where: TP stands for the True Positives (non-defaults predicted as non-defaults) in the data sample, FP for the False Positives (non-defaults predicted as defaults), FN for the False Negatives (defaults predicted as non-defaults), TN for the True Negatives (defaults predicted as defaults).	PCC assumes equal FN and FP misclassification costs and is dependent on the cut-off value.
Sensitivity, SENS (Baesens et al., 2003)	SENS is a measure to assess the discriminatory power of a credit scoring model specifically for non-default-predicted loans (range: 0% ≤ SENS ≤ 100%): $SENS = \frac{TP}{(TP+FN)} \quad \text{(A.2)}$	SENS is dependent on the cut-off value and is limited to non-default-predicted loans. SENS focuses on the proportion of correctly predicted successes among the loans which were predicted to be succesful.
Specificity, SPEC (Baesens et al., 2003; Kleimeier and Dinh, 2007)	SPEC is a measure to assess the discriminatory power of a credit scoring model specifically for default-predicted loans (range: 0% ≤ SPEC ≤ 100%): $SPEC = \frac{TN}{(FP+TN)} \quad \text{(A.3)}$	SPEC is dependent on the cut-off value and is limited to default-predicted loans. SPEC focuses on the proportion of correctly predicted failures among the loans which were predicted to be failures. SPEC is an unimplementable measure in practice, as a lender cannot retrieve whether its declined loan applicants will actually default or not.
Kolmogorov-Smirnov Statistic, KS (Thomas, 2007; Sun and Wang, 2005)	KS assesses the discriminatory power of a credit scoring model by measuring the maximum distance between the cumulative probability functions of defaulted and non-defaulted loans (range: 0 ≤ KS ≤ 1): $KS = \max_s \lvert F(s\lvert G) - F(s\lvert B)\rvert = \max_s \lvert SENS + SPEC - 1\rvert \quad \text{(A.4)}$ where: $F(s\lvert G)$ and $F(s\lvert B)$ are the cumulative probability functions of non-defaulted and defaulted loans and s refers to the score assigned to individual loans.	KS only describes the optimal situation, at a cut-off score which is usually much higher than any realistic cut-off score.

Table 6: Performance Measures for Binary Logit Credit Scoring Models (Part I).

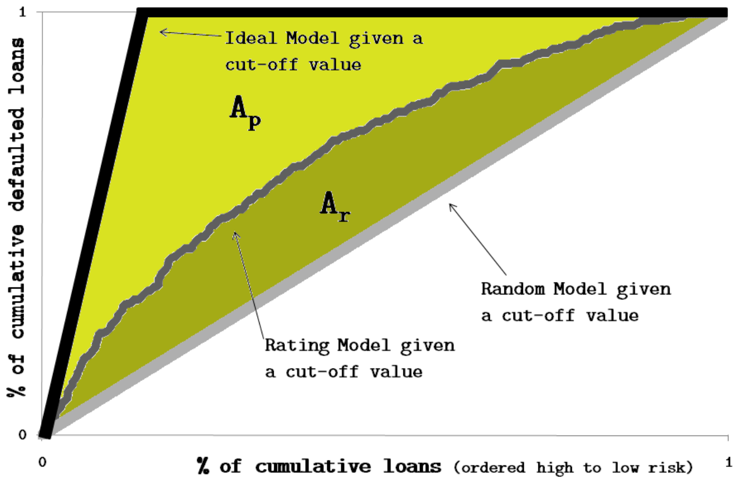
Panel A: Performance Graphs for Binary Logit Credit Scoring Models

Receiver Operation Characteristic Curve, ROC



The ROC curve represents for all possible decisions the 'hit rate' (SENS) and 'false alarm rate' (1-SPEC) (Tasche, 2005; Blöchlinger and Leippold, 2006).

Cumulative Accuracy Profile Curve, CAP



The CAP curve represents to what extent the ideal case (cumulative % of defaulted loans equaling the cumulative % of all loans multiplied by default rate) is realized for each % level (Tasche, 2005).

Panel B: Performance Measures for Binary Logit Credit Scoring Models (Part II)

Name (Source)	Definition	Strenghts and Limitations
Area Under the ROC Curve, AUC (Tasche, 2005)	AUC measures the discriminatory power of a credit scoring model by assessing the area under the ROC curve of the rating model, which can be interpreted as the probability that a good loan receives a better score than a bad loan (range: 0.5 ≤ AUC ≤ 1): $AUC = \int_0^1 SENS(1 - SPEC)d(1 - SPEC)$ (A.5)	AUC is a summary index for the ROC curve. As a measure, AUC is independent of misclassification costs or class distributions and it represents all cut-off values, which makes it comparable over different models.
Accuracy Rate, AR (Tasche, 2005; Kraft et al., 2004)	AR assesses the discriminatory power of a credit scoring system by measuring the area under the CAP curve of the rating model relative to the area under the CAP curve of the perfect model (range: 0 ≤ AR ≤ 1): $AR = \frac{A_r}{A_p}$ (A.6)	AR is a summary index for the CAP curve. AR and AUC are linear transformations of each other ($AR = 2AUC - 1$), both making discriminatory power performance comparable over different models and cut-off values.

Table 7: Performance Graphs and Measures for Binary Logit Credit Scoring Models (Part II).

References

Baesens, B., Van Gestel, T. and Thomas, L. (2009). *Credit Risk Management: Basic Concepts*. Oxford: Oxford University Press.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6), pp 627–635.

Basel Committee on Banking Supervision (2006). *Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive Version*. Basel: Bank of International Settlements.

Beale, E. and Little, R. (1975). Regression with missing x's: A review. *Journal of the Royal Statistical Society Series B (Methodological)*, 37(1), pp 129–145.

Blöchlinger, M. and Leippold, M. (2006). Economic Benefit of Powerful Credit Scoring. *Journal of Banking and Finance*, 30(3), pp 851–873.

Boyle, M., Crook, J., Hamilton, R. and Thomas, L. (1992). Methods for credit scoring applied to slow payers. *in* L. Thomas, J. Crook and D. Edelman (eds), *Credit Scoring and Credit Control*. Oxford: Oxford University Press. pp 75–90.

Cantor, R. and Packer, F. (1996). Determinants and Impact of Sovereign Credit Ratings. *Economic Policy Review*, 2(2).

Capon, N. (1982). Credit Scoring Systems: A critical analysis. *Journal of Marketing*, 46(2), pp 82–91.

Crook, J. and Banasik, J. (2004). Does Reject Inference Really Improve the Performance of Application Scoring Models?. *Journal of Banking and Finance*, 28(4), pp 857–874.

Crook, J., Hamilton, R. and Thomas, L. (1992). A comparison of discriminations under alternative definitions fo credit default. *in* L. Thomas, J. Crook and D. Edelman (eds), *Credit Scoring and Credit Control*. Oxford: Oxford University Press. pp 217–245.

Dennis, W. (1995). Fair Lending and Credit Scoring. *Mortgage Banking*, 56(2), pp 55–59.

Desai, V., Crook, J. and Overstreet, G. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, 95(1), pp 24–37.

Diallo, B. (2006). Un modele de 'credit scoring' pour une institution de micro-finance Africaine: le cas de Nyesigiso au Mali.

Freytag, C. (2008). Credit scoring: Why scepticism is justified. *in* I. Matthus-Maier and J. von Pischke (eds), *New Partnerships for Innovation in Microfinance.* Berlin Heidelberg: Springer. pp 233–235.

Hand, D. and Henley, W. (1993). Can Reject Inference Ever Work?. *Journal of Management Mathematics*, 5(1), pp 45–55.

Hand, D. and Henley, W. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 160(3), pp 523–541.

Kleimeier, S. and Dinh, T. (2007). A credit scoring model for Vietnam's retail banking market. *International Review of Financial Analysis*, 16(5), pp 471–495.

Kraft, H., Kroisandt, G. and Muller, M. (2004). Redesigning Ratings: Assessing the Discriminatory Power of Credit Scores under Censoring. *Working Paper Series - Fraunhofer Institut fur Technology und irtschaftsmathematik*, .

Kulkosky, E. (1996). Credit Scoring could have a downside, experts say. *American Banker*, 161(208), pp 8.

Morduch, J. (2000). The Microfinance Schism. *World Development*, 28(4), pp 617–629.

Navarrete, E. and Navajas, S. (2006). Basel II and Microfinance. *Inter-American Development Bank Microenterprise Development Review*, 9(1).

Provost, F., Fawcett, T. and Kohavi, R. (1998). The Case Against Accuracy Estimation for Comparing Classifiers. *in* J. Shavlik (ed.), *Proceedings of the Fifteenth International Conference on Machine Learning.* San Francisco, CA: Morgan Kaufmann. pp 445–553.

Reinke, J. (1998). How to lend like mad and make a profit: A micro-credit paradigm versus the Start-up Fund in South Africa. *Journal of Development Studies*, 34(3), pp 44–61.

Robinson, M. (2001). *The Microfinance Revolution.* Washington, DC: World Bank.

Schreiner, M. (2003). Scoring: The Next Breakthrough in Microfinance. *CGAP Occasional Paper*, (7).

Schreiner, M. (2004). Scoring arrears at a Microlender in Bolivia. *Journal of Microfinance*, 6(2), pp 65–88.

Sharma, M. and Zeller, M. (1997). Repayment Performance in Group-Based Credit Programs in Bangladesh: An Empirical Analysis. *World Development*, 25(10), pp 1731–1742.

Sharma, S. (1996). *Applied Multivariate Techniques.* New York, NY: John Wiley Sons.

Sun, M. and Wang, S. (2005). Validation of Credit Rating Models - A Preliminary Look at Methodology and Literature. *Review of Financial Risk Management*, 2(94), pp 1–15.

Tasche, D. (2005). Rating and probability of default validation. *in* Basel Committee on Banking Supervision (ed.), *Working Paper No. 14 - Rating and Probability of Default Validation.* Basel: Bank of International Settlements.

Thomas, L. (2000). A Survey of Credit and Behavioural Scoring; Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2), pp 149–172.

Thomas, L. (2007). Measuring the Discrimination Quality of Suites of Scorecards: ROCs Ginis, Bounds and Segmentation.

Van Gestel, T., Baesens, B., Van Dijcke, P., Garcia, J., Suykens, J. and Vanthienen, J. (2006). A process model to develop an internal rating system: Sovereign credit ratings. *Decision Support Systems*, 42(2), pp 1131–1151.

Van Gestel, T., Baesens, B., Van Dijcke, P., Suykens, J., Garcia, J. and Alderweireld, T. (2005). Linear and Non-linear Credit Scoring by Combining Logistic Regression and Support Vector Machines. *Journal of Credit Risk*, 1(4).

Verstraeten, G. and Van den Poel, D. (2005). The impact of sample bias on consumer credit scoring performance and profitability. *Journal of the Operational Research Society*, 56(8), pp 981–992.

Vigano, L. (1993). A credit-scoring model for development banks: An African case study. *Savings and Development*, 17(4), pp 441–482.

Vogelgesang, U. (2003). Microfinance in Times of Crisis: The Effects of Competition, Rising Indebtness, and Economic Crisis on Repayment Behaviour. *World Development*, 31(12), pp 2085–2114.

Wainer, H. (1976). Robust Statistics: A Survey and Some Prescriptions. *Journal of Educational and Behavioral Statistics*, 1(4), pp 285–312.

Yobas, M., Crook, J. and Ross, P. (2000). Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics*, 11(2), pp 111–125.

Zeller, M. (1998). Determinants of repayment performance in credit groups: The role of program design, intra-group risk pooling, and social cohesion. *Economic Development and Cultural Change*, 46(3), pp 599–620.