Improved Credit Scoring with Multilevel Statistical Modelling

Inauguraldissertation zur Erlangung
des akademischen Grades
Doktor rerum politicarum
der Technischen Universität Dortmund

Alesia S. Khudnitskaya

December 2010

ADVISOR:	Prof. Wälter Krämer, PhD (Institut für Wirtschafts- und Sozialstatistik)
CO-ADVISOR	Prof. Peter Recht, PhD (Operations Research und Wirtschaftsinformatik)
FACULTY:	Technische Universität Dortmund, Wirtschafts- und Sozialwissenschaftliche Fakultät
LOCATION:	Dortmund
CANDIDATE:	Alesia S. Khudnitskaya, M.Sc. (Financial Economics)
TITLE:	Improved Credit Scoring with Multilevel Statistical Modelling.
DATE:	December 17, 2010

Contents

List of tables

List of figures

Acknowledgements

I would like to express sincere gratitude to my supervisor, Prof. Dr. Walter Krämer. With his enthusiasm and inspiration, and his great effort to explain things clearly and simply, he helped me a lot. I am very grateful to my co-advisor, Prof. Dr. Peter Recht, and his colleagues for sharing ideas and proving sound comments during the introductory presentation in June.

I would like to thank Dr. Matthias Arnold for relevant comments, fruitful discussions and sound advices on dissertation writing. My graduate studies would not have been the same without the social and academic challenges provided by all my colleagues in the Institut für Wirtschafts- und Sozialstatistik (IWUS) at the Technische Universität Dortmund. They created a stimulating environment in which to learn and grow.

It is a great pleasure to thank many people at Ruhr Graduate School in Economics and Rheinisch-Westfälisches Institut for providing excellent opportunity for enhanced training and for the financial assistance during this PhD program. Many thanks to my RGS cohort fellows and, especially, to the senior fellow Pavel Stoimenov who sincerely shared his experience and gave many perceptive comments on the topic.

I am forever grateful to my mother, Zoya Khudnitskaya, and my father, Dr. Stanislav Khudnitskii, for their understanding, endless patience and encouragement when it was most required. To them I dedicate this dissertation. In addition, I wish to thank my uncle, Prof. Dr. Uriy Ushko, who was exemplary for me in many ways. Taking part in his course in Corporate Finance helped me to find focus and guide my research interests towards financial economics.

1 Introduction

“ We cannot direct the wind… but we can adjust the sails ”
Dolly Parton

This dissertation contributes to the literature on credit scoring. It introduces a new type of credit scoring model which specifies a multilevel structure to the data. To my knowledge, multilevel credit scoring models have never been applied in retail banking for credit scoring. These scorecards are improved alternatives to the conventional scoring techniques which include discriminant analysis and logistic regression scorecards.

The multilevel scoring model assesses credit worthiness of applicants for a loan by forecasting their probability of default. I introduce and fit several versions of the multilevel models which vary by the degree of complexity and are designed to answer different questions in application credit scoring. In addition, this thesis proposes a new way of data clustering for a multilevel structure which is more intuitive and relevant for efficient credit worthiness assessment.

Credit scoring plays an important role in the general lending practice within a bank. Therefore, recently, the majority of credit scoring models are based on prominent statistical theory (Anderson (2007), Crook (2005)). This is a logical further development of the subjective credit rating provided by the human judgment alone. These scoring models are also called predictive statistical scoring models. They are used to assess the relative likelihood of the future event of interest, based on some historical knowledge and past experience. The process of scoring involves collecting of relevant information about borrowers and then applying it in order to discriminate the population of applicants for a loan into two parts: accepted and rejected customers. Credit scoring models are also called scorecards. I will use these denotations interchangeably in this thesis.

The motivation for the topic and the core idea of this dissertation are closely related to the main advantages of improved credit scoring and its application into the decision-making process in retail banking. The main advantages are the accuracy gain and cost-saving. Improving credit scoring techniques helps to increase operating efficiency by increasing the predictive quality and reducing misclassification errors. From the cost-saving prospective, it also leads to profit growth and gives a higher return on capital. Accordingly, this thesis proposes several alternative specifications of the multilevel scorecards and demonstrates that these models outperform standard scoring models by providing a higher forecasting accuracy.

In credit scoring the main goal is to define factors which influence riskiness of individuals who apply for a bank loan. Accordingly, I introduce a particular type of multilevel structure which is relevant for a more efficient credit scoring. The main advantage of this structure is that it makes use of information on unobserved characteristics which impact credit worthiness of borrowers additionally to the observed characteristics such as income, marital status and credit history. Accounting for unobserved determinants of default in a credit scoring model is important and helps to increase the accuracy of the model predictions. The scorecard assumes that these unobserved characteristics of credit worthiness are random-effects. This thesis introduces two types of multilevel structures which allow including random-effects at the higher-level of the hierarchy. The first structure nests applicants for a loan within second-level groups, microenvironments. Each microenvironment determines the living area of a borrower with a particular combination of socio-economic and demographic conditions. Microenvironment-specific effects impact the riskiness of borrowers additionally to the observed personal characteristics. Importantly, clustering within microenvironments differs from simple geographical grouping. The difference is that microenvironments can include individuals from different cities or regions if their living area conditions are similar.

The second type of multilevel structure extends the first. It cross-classifies individuals with different classifications according to similarities in particular characteristics of their occupational activities, living area condition and infrastructure of shopping facilities in their residence areas.

It is important to mention that in this dissertation I mainly focus on application credit scoring which is implemented on the first stage of the decision-making process in retail banking: when individuals apply for a loan and a lender has to decide whether to accept or reject a borrower. In this case, lenders begin scoring by making an assessment of prospective customers according to their capacity to borrow, credit history and derogatory information, capital (credit resources) and conditions of a credit deal. These assessments are based upon the lenders own experience, taking into consideration not only the historical information, but also a forward-looking view of the borrowers' prospects. Then this information is used in a credit scoring model which as a result provides a credit score. An application credit score provides the numerical assessment of borrower's credit worthiness and is regularly measured by probability of default. When application scores are estimated lenders choose the cut-off point which discriminates the population of borrowers into two categories. Applicants above the cutoff point are going to be granted a loan and applicants below the cutoff point are rejected.

The quality of application credit scoring models should be of primary importance for a retail banker as these scores are applied to a new cohort of customers in the first place. Application credit scores also help to choose the most reliable borrowers from the population of all customers who apply for a loan to a bank.

In spite of that, application credit scores are one of the most important for a bank, there exist other types of credit scores depending on where and how they are used. The most common are behavioral scores, Bureau scores and customer scores (Hand (1997), Baesens (2005)). I do not discuss these alternatives in this dissertation.

There are a number of credit scoring techniques which aim to assess credit worthiness. The most commonly applied methods are logistic regression scoring, probit models, decision trees and multiple discriminant analysis (Anderson (2007). The primary differences between these techniques involve the assumptions regarding the explanatory variables and the ability to model binary outcomes. In addition to the multilevel scorecards, I also fit a logistic scorecard in order to compare the predictive quality between the multilevel scoring models and a benchmark logit.

1.1 Literature overview

This subsection reviews the literature on multilevel modeling and discusses the main fields of recent application of multilevel models. In general, multilevel models combine features of known models such as variance component models, mixed effects models and random-effects models in panel data analysis. Variance component models are also called hierarchical linear models. It is assumed that the data used in the variance analysis is grouped within one or more hierarchical categories. According to Kreft and de Leeuw (1998) variance component models were mentioned for the first time by the astronomer Airy (1879). However, the substantial work was done by Fisher (1918) who introduced the term "analysis of variance" in the literature and developed variance component models. Tipett (1931) was the first to employ linear models in the analysis of variance. He considered the problem of selecting the optimal sampling design for particular experimental situations for a one-way random model.

A further extension of a variance component model is a mixed-effects model which puts distinction between fixed and random effects in the model. A mixed-effects regression was introduced by Eisenhart (1947) and Henderson (1953) who also developed best linear unbiased estimates of fixed effects and best linear unbiased predictions of random effects (BLUP).

Longitudinal or panel data model is a kind of mixed-effects model. A panel model assumes that the same characteristics are measured repeatedly over time for the same set of individuals or households. A comprehensive review of panel data models and estimation approaches are discussed in detail in Chamberlain (1984), Hsiao (2007) and Verbeke and Molenberghs (2006). Strenio et al. (1983) were the first to relate panel data models and multilevel models.

In general, a multilevel model is a more advanced form of a mixed-effects model which includes fixed and random-effects at different levels of the model hierarchy. These statistical models imply that the data for the analysis is nested within groups. In the simple multilevel model with two levels observations are treated as level-one units which are clustered within level-two units, groups. Nested data structure or hierarchical structure is typical in social sciences and behavioral economics. The most prominent example in the literature where data has a hierarchical structure comes from the field of education where pupils or students are nested within schools or classes (Goldstein (2003), Blatchford (2002), and Steele (2007)). The motivation for this kind of grouping is that it is assumed that individual units from the same group share more similarities than units from different groups. Goldstein et al. (1996, 1999) and Burkholder and Harlow (2003) apply multilevel modeling to analyze pupils' examination results. They emphasize that pupils from the same school share more common characteristics than pupils randomly drawn from a population of pupils. The similarities in characteristics are explained by school-specific internal rules and customs, teaching methods and leisure activities. All these characteristics determine school specifics which make pupils within one school more similar to each other compared to pupils from other schools.

In an organizational behavior study, typical examples of hierarchical structures include employees-within-firms and firms-within-cities (Staw, Sandelands and Dutton (1981)). Browne and Prescott (2006) discuss the application of multilevel data structures in health economics and pharmaceutical industry. In particular, they apply a two-level structure (patients-within-hospitals) to examine the differences between hospitals in their rates of post-operative complications. In political science Gelman (2007) uses a hierarchical structure to analyze voting preferences during the presidential election in 2000.

2 Multilevel Hierarchical Credit Scoring Model

This chapter introduces a new type of a credit scoring model which has a multilevel structure. Multilevel credit scoring models have never been applied in retail banking for credit worthiness assessment. Here, I demonstrate that the multilevel scoring model is an improved alternative to a conventional logistic scoring regression which is regularly applied in retail banking. In addition, the chapter proposes a new type of clustering for a hierarchical two-level structure which is more intuitive and efficient in the application to credit scoring. This structure explores living area-specific effects which are viewed as unobserved determinants of default. Including area-specific effects in the models improves the accuracy of the forecasts and allows evaluating the impact of the particular group-level characteristics on default.

I introduce several versions of the credit scoring models which can be used in retail banking for a credit worthiness assessment of customers. Importantly, the thesis mainly focuses on application credit scoring. It implies that a scorecard is primarily used for forecasting the probability of default of a customer who applies for a bank loan and for whom a detailed credit history is collected. Accordingly, I do not discuss other types of credit scoring models here. However, the approach can easily be extended to the behavioural or relationship scoring models.

The chapter is divided into three parts: theory, empirical application and discussion of the results. The first section 2.1 presents the multilevel structure and gives a motivation for the particular type of a hierarchical structure. A detailed description of the data used in the empirical analysis is given in section 2.2. The data sample contains credit histories of borrowers which are collected from three different sources: personal data, Credit Bureau reports and socio-economic data for the living area of a borrower. I also use statistical data for regional economic accounts (counties and states) provided by the Bureau of Economic Analysis (BEA).

The empirical part of the chapter (section 2.3) specifies the multilevel credit scoring models and applies them to the credit history data. The scorecards vary by the degree of complexity. I begin by presenting the simplest version with only a random-intercept and then elaborate it by including more random-effects and group-level characteristics.

The data sample is divided into two parts: a training sample and a testing sample. The training data sample is applied to fit the scorecards and the testing data sample is used for the postestimation diagnostics.

I apply a ROC curve analysis to check the predictive accuracy of the estimated scoring models. The ROC curve plots and related metrics conclude the presentation of the empirical results for the scorecards in each subsection. In addition, I perform several other statistical tests which aim to assess the discriminatory power of the models. I summarize and compare the performance between the multilevel scorecards and a conventional scoring model in Chapter 3. In addition, this chapter discusses the main limitations and drawbacks associated with an application of the ROC curve (AUC) to credit scoring and proposes alternative methods for evaluating forecasting accuracy.

2.1 Microenvironment and multilevel structure

The scope for the application of multilevel structures is wide. It allows addressing various questions and fitting models of different complexity. In credit scoring the main goal is to define factors which influence riskiness of individuals who apply for a bank loan. Accordingly, I introduce a particular type of multilevel structure which is relevant for a more efficient credit scoring. The main advantage of this structure is that it makes use of the information on unobserved characteristics which impact credit worthiness of borrowers additionally to the observed characteristics such as income, marital status and credit history. Accounting for unobserved determinants of default in a credit scoring model is important and helps to increase the accuracy of the model predictions. The scorecard assumes that these unobserved characteristics of credit worthiness are random-effects.

I define a two-level hierarchical structure for a scoring model which includes random-effects. The structure nests applicants for a loan within microenvironments. I use the term *microenvironment* to determine the living area of a borrower. Each microenvironment represents a particular combination of socio-economic and demographic conditions. In this two-level structure borrowers are treated as the level-one units which are nested within the level-two units, the microenvironments.

2.1.1 Clustering algorithm

The grouping of borrowers within microenvironments is done according to the similarities in the economic and demographic conditions in their residence areas. In order to nest the borrowers within microenvironments I use area descriptive data as well as BEA data on regional economic accounts. The economic determinants of grouping include living area income, housing wealth and the percentage of retail stores, furniture outlets, gas stations and autohouse sales in the total sales in the market. The socio-demographic determinants of grouping are the share of individuals with a college degree in the living area and the share of African-American (Hispanic) residents in the district.

I apply non-hierarchical clustering algorithm, k-means, to nest the borrowers within microenvironments. This algorithm was first used by MacQeen in 1967, though the idea goes back to Hugo Steinhaus in 1956. The procedure follows a simple and easy way to classify a data set through a certain number of clusters fixed a priori. The main idea is to find k centroids, one for each cluster.

The next step is to take each point belonging to a given data set and associate it to the nearest centroid by minimizing an objective function.

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-means algorithm aims to partition n observations into k sets $S = \{S_1, S_2, \ldots, S_k)$ $(k<n)$ in order to minimize the within-cluster sum of squares (MacQeen (1967)):

$$arg\ min_S \sum_{i=1}^{k} \sum_{x_j \epsilon S_i} \left\| x_j - c_i \right\|^2 ,$$

where c_i is the mean of points in S_i.

The k-means clustering follows an iterative refinement technique. Given an initial set of k centroids $c_1^{(1)}, \ldots, c_k^{(1)}$, which may be specified randomly or defined a priori, the algorithm proceeds by alternating between two steps (Mackey (2003)): assignment step and update step.

At the assignment step each observation is located to the cluster with the closest mean:

$$S_i^{(t)} = \{x_j : \left\| x_j - c_i^{(t)} \right\| \leq \left\| x_j - c_{i*}^{(t)} \right\| \ for\ all\ i^* = 1, \ldots, k\}.$$

When no point is pending, the first step is completed and an early grouping is done. Given k new clusters the update step recalculates new centroids in the clusters and then reestimates new distances.

$$c_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \epsilon S_i^{(t)}} x_j\ .$$

The procedure is repeated until centroids do not move anymore. The clustering algorithm applies squared-Euclidian distance to k-means which is calculated as follows

$$Distance_{Euclidian} = \sqrt{\sum_d (x_{j,d} - c_{i,d})^2}\ ,$$

where d is the dimension of x_j or the number of input variables (determinants of clustering) which are used to nest borrowers within microenvironments.

The k-means clustering is sensitive to the choice of initial cluster centres. Therefore, I define a grouping variable which provides initial clustering. To create the grouping variable, I sort the data by median income in the living area and split the data into 70 equal-size clusters. In addition, I normalize the input variables used in the clustering by subtracting mean and dividing by standard deviation. This makes the variables comparable as initially they are measured on different scales.

After clustering is done I adjust the number of clusters by combining together clusters which have small number of observations and similar centroids. The final set of clusters consists of 61 microenvironments. It should be mentioned that it possible to specify other number of clusters or define other determinants of clustering for a multilevel scorecard. This choice should generally be determined by a researcher. In this thesis I focus on the two-level structure with 61 microenvironments because I suggest that this is a reasonable amount of second-level clusters given the quantity and quality of the credit history data.

2.1.2 Microenvironments: aims and advantages

I define 61 microenvironments within which all borrowers are clustered. Each microenvironment includes borrowers with a unique combination of economic and demographic conditions in their living areas. In the credit scorecard the unobserved microenvironment-specific characteristics are captured by the random-effects and the observed living area characteristics are given by the group-level variables.

There are several reasons why including information on the microenvironments in the credit scoring model is important and advantageous. First, it shows that borrowers from dissimilar living areas are exposed to different risk factors which impact their probabilities of default. It is evident that poor living areas with an undeveloped infrastructure of shopping facilities have higher unemployment rates and crime rates, contain a lower share of individuals with a college degree and have a lower level of real estate wealth (percentage of families who own a house). In such environments, individuals have a higher chance to experience adverse events such as damage of a property, severe income cut, loss of the main job or health problems. All these factors are living area-specific and influence the riskiness of borrowers who reside in such regions. It is also true that the quality of borrowers differs between low income areas and high income regions. The share of borrowers with a problematic credit debt and the share of individuals with derogatory credit history are higher in poor regions as compared to richer areas. This is because the amount of unobserved area-specific risks is much higher in a microenvironment with unstable economic conditions. I do not list all risks as it is assumed that the microenvironment random-effects aggregate the information on all unobserved determinants of default and explain the area-specific hazards. Importantly, these area-specific effects impact probability of default given the personal characteristics of borrowers. For two applicants with exactly the same personal characteristics, probabilities of default are different and depend on the microenvironment in which they reside.

Second, recognizing the two-level structure which nests borrowers within microenvironments allows exploring the impact of the microenvironment-level characteristics on default. The microenvironment-level information is given by the group-level variables. I define several second-level variables which characterize the conditions within a living area. A graphical illustration in chapter 3 provides a discussion about the impact of living area income and real estate wealth on the riskiness of borrowers from poor and rich areas. In order to explore the effect of the socio-demographic conditions in the region on default, I specify higher education and the share of African-American residents as the second-level variables.

I provide a descriptive table in order to interpret the term microenvironment. Table 2.1 reports the economic and demographic characteristics of the living area conditions within high income, average income and low income microenvironments. The region-specific characteristics are area income, real estate wealth, share of college graduates and share of African-American residents.

The first three microenvironments represent economically stable living areas where the average income is high and the majority of families own a real estate property. These areas also contain a higher share of high-skilled individuals (college degree) and a lower share of African-American and Hispanic residents. The last two columns describe the living conditions in poor microenvironments where the average level of income is low and only a minor share of families have a real estate property. The differences in the living area economic and demographic conditions between poor and rich microenvironments are huge. This implies that the exposure to the microenvironment-specific risks also varies considerably across the regions.

Living area characteristics	*Microenvironment ID*				
	8	***6***	***39***	***17***	***52***
Average area income, $ annually	75 000	55 250	42 940	24 360	18 420
Housing wealth (% of families who own a house)	80.40	57.00	81.20	30.70	3.20
Share of college graduates , %	27.90	21.00	15.90	7.40	1.10
African-American + Hispanic residents,%	2.60	8.80	15.30	29.60	98.80
Median age	38.0	40.0	34.6	32.1	29.1

Table 2.1. *Descriptive summary for five microenvironments. Each microenvironment determines a unique combination of economic and demographic conditions in the living area of a borrower.*

The main aim of multilevel modelling is to make inference about a population. It is assumed that there is a population of microenvironments within which all borrowers reside. Observing a sample from this population helps to explore the parameter values in the population. Accordingly, in a credit scoring model the unobserved microenvironment-specific effects are assumed to be drawn from a population. These effects are viewed as random and in the scorecards they are captured by the random-intercepts and random-coefficients (other than intercept).

Importantly, I emphasize that the two-level structure which nests borrowers within microenvironments is a more efficient alternative to the conventional type of structure where individuals are nested in groups according to their geographical locations. The geographical grouping suggests clustering of individuals within small regions, cities or states. The main difference between these two types of clustering is that the former structure is more relevant in application to credit scoring because it recognizes that borrowers within one group are similar in terms of their living area conditions. It implies that a particular combination of economic and demographic characteristics within a microenvironment impacts the riskiness but not a geographical location itself. Accordingly, within one microenvironment it is possible to have applicants from different regions or cities if their living area conditions are essentially the same. For instance, in the case of Germany, a geographical clustering of borrowers shows nesting of individuals within one of the 429 urban or rural districts (Landkreise, 313, und Stadtkreise, 116). This kind of grouping represents only residence locations of the borrowers and does not clarify which characteristics of their neighbourhoods impact the probability of default. Alternatively, if I nest individuals within microenvironments, then it is clear which combination of the area-specific characteristics such as area income, unemployment rate, share of college graduates or foreign residents influence the probability. A microenvironment may contain borrowers from different rural districts or cities if their living area conditions are similar. In this case individuals within one microenvironment are exposed to the same triggering default factors and living area risks (poor regions, high crime rates, bad labour market, etc.) which impact their riskiness and probability of default.

For instance, if we compare zip-code areas in Dortmund and Essen cities, there are good and bad areas within each of the cities. Pure geographical grouping would nest borrowers within clusters taking into account only their location in Essen or in Dortmund. This creates many clusters (areas in Dortmund and Essen) which include regions with very similar, almost identical, economic and demographic conditions. In this case geographical grouping is inappropriate because it leads to wrong inferences about the between-groups variance. In construct, nesting of applicants within microenvironments resolves this problem by combining areas with similar economic conditions in one cluster. In addition, it reduces the number of overall clusters and increases the precision of the parameters' estimates.

In summary, the main weakness of a geographical grouping is that it is inappropriate if there are regions with similar economic conditions, like Dortmund and Essen. In this case, the multilevel structure recognizes two different living area-effects which are the same in reality. Grouping borrowers within microenvironments helps to resolve this problem.

2.2 Data and variables

In this section I describe the data used in the empirical application of the multilevel credit scorecards and list the predictor variables used in the models. The dataset is part of American Express credit card database analyzed by W.Greene (1992). The full dataset contains 13 444 records on credit histories of individuals who applied for a loan in the past. In this sample 10499 applications are accepted for a loan and 2945 are rejected. The outcome variable (default or not default) is observed for the subsample of accepted borrowers and personal characteristics plus auxiliary information are available for the full sample of borrowers. In the empirical analysis I use the subsample of applicants who were granted a loan. There are 996 defaulters within this sample. Default occurs when a credit account is more than 6 months past due.

In addition, I collect data on regional economic accounts provided by the Bureau of Economic Analysis (BEA) in the United States (www.bea.gov). The BEA information includes annual personal income, full and part-time employment, taxes and gross domestic product. The regional-account data is measured at the county, metropolitan area and state levels. I apply this data in order to define the microenvironment-level characteristics and create the group-level variables which are then used in the multilevel credit scoring models in section 2.3.

The credit history data combines information on the applicant for a loan collected from three different sources: personal data, credit Bureau report and living area descriptive data. The personal data is collected through application forms which borrowers fill in when they apply for a credit. It includes socio-demographic characteristics such as family composition, age, level of education, annual labour income, additional income, occupational field, monthly expenditures, accommodation ownership, employment duration in months, duration at current and previous living addresses and other information.

The credit bureau report contains detailed information on the past credit history of a customer. It provides information on major and minor derogatory reports, shows the history of previous credit file searches or enquiries, lists the past experience with a lender (such as banking saving and checking accounts or personal loans), shows the number of open (active) trade accounts and gives a detailed overview of the currently issued credit cards and revolving credit lines. Consumer credit enquiries is a notice in a credit profile of a borrower which shows how many times a customer applied for a new credit (mortgage, auto loan, or credit card) prior to the current application. Credit inquiries appear in a credit profile whether the applications were approved or not. Given information on enquiries lenders can determine if a borrower has been trying to secure new lines of credit recently or obtain a loan to consolidate the past due bills.

Living area descriptive data characterize the economic and demographic conditions in the borrowers' neighbourhoods. These characteristics are measured for the areas defined by the 5-digit zip-code. The major benefit of using a micro-level statistical data is that it provides a better representation of the living area conditions within the bigger regions or states. States, regions and large cities usually combine micro-areas with very dissimilar conditions.

The living area descriptive data contains the following characteristics: per capita income in the market, population growth rate, buying power index, unemployment rate, percentage of African-American (Hispanic) residents and detailed information on the shopping facilities available in the living area. The characteristics of the shopping facilities include the share of retail store, gasoline company, furniture outlet, dining place and drug store sales in the total retail sales in the market.

The summary and descriptive statistics of the credit history data is provided in Table 2.2.

Table 2.2. *Summary and descriptive statistics of the credit history data.*

Variable	*Measure*	*Mean*	*Std.dev*	*Min*	*Max*
Personal information					
Labour income[1]	Continuous, thd of $	34.2	17.7	1.3	100.0
Additional income[1]	Continuous, thd of $	4.1	9.1	0	10.0
Age	Number	33.4	10.2	18	88
Number of dependents in the family	Numerical	1.02	1.2	0	9
Duration in months at current address	Numerical, in mths	55.31	63.08	0	576
Duration in months at previous address	Numerical, in mths	81.3	80.5	0	600
House owner / renter	Indicator	0.45	0.49	0	1
Average revolving balance	Continuous, thd of $	5.2	7.5	0	19

Occupation:					
Military	Indicator	0.022	0.14	0	1
Retail trade	Indicator	0.078	0.26	0	1
High-skilled professionals	Indicator	0.115	0.31	0	1
Management	Indicator	0.074	0.26	0	1
Clerical staff	Indicator	0.088	0.28	0	1
Proprietors	Indicator	0.057	0.23	0	1
Construction, transportation and others	Indicator	0.620	0.48	0	1
Credit Bureau report					
Derogatory information:					
Major reports	Numerical	0.460	1.40	0	22
Minor reports	Numerical	0.290	0.76	0	11
Previously had a loan with a lender	Indicator	0.073	0.26	0	1
Dollar amount of average revolving balance	Numerical	52.81	75.90	0	190
Miscellaneous data:					
Department credit card	Indicator	0.150	0.34	0	1
Gasoline credit card	Indicator	0.028	0.16	0	1
Number of credit enquiries (applications for a loan)	Numerical	1.400	2.20	0	56
Number of trade lines 30 days past due	Numerical	0.055	0.26	0	3.0
Number of 30 day-delinquencies in last 12 months	Numerical	0.365	1.24	0	21
Banking accounts:					
Checking account	Indicator	0.297	0.45	0	1
Savings account	Indicator	0.034	0.18	0	1
Checking and savings	Indicator	0.661	0.47	0	1
Number of current trade accounts	Numerical	6.42	6.10	0	50
Number of open trade accounts	Numerical	6.050	5.20	0	43
Number of active trade accounts	Numerical	2.280	2.60	0	27
Average revolving credit balance	Continuous, thsd $	5.28	7.5	0	190
Living area descriptive data					
Real estate wealth (share of families which own a house)	Percentage	53.9	28.2	0	100
Per capita area income[1]	Continuous, thds $	28.34	10.4	0	75.1
Demographic characteristics:					
African-American residents	Percentage	11.7	20.5	0	100
Spanish residents	Percentage	7.7	13.1	0	96
Employment in the living area	Percentage	40.99	108.0	0	65.2
College graduates	Percentage	10.7	8.5	0	54.9
Average age of residents	Continuous	33.2	5.4	0	65
Index of buying power in market (5 digit zip code)	Index	0.014	0.009	0	0.113
Population growth rate (annual)	Percentage	22.4	18.7	-6.1	70.68
Infrastructure of shopping facilities:					
Apparel	Percentage	2.43	2.43	0	33.3

Autohouses	Percentage	1.49	1.32	0	33.3
Gas	Percentage	1.76	1.79	0	99
Dining places	Percentage	6.58	3.95	0	99.1
Drug stores	Percentage	1.30	1.77	0	15.2
Build material outlets	Percentage	1.12	1.23	0	33.3
Furniture	Percentage	1.86	2.51	0	99

[1] *Income, additional income and per capita area income are measured in $1000 units and are censored at 100.*

The main aim of the thesis is to develop a multilevel credit scorecard and to show that this kind of a credit scoring model has a higher predictive accuracy than a conventional scorecard. Here, I refer to a logistic regression scorecard as a standard credit scoring model. Accordingly, in order to compare out-of-sample predictive performance between the multilevel scorecards and a logistic regression, I randomly split the sample into two subsets. First, I assign random numbers to each observation in the sample and then draw a random sample without replacement. The first subset is the training sample which is used in the estimation stage. It contains 60% of observations. The second subset is the testing sample which is applied in the forecasting stage. The description summary for the training and testing datasets is given in Table 2.3.

	Full sample	*Training sample*	*Testing sample*
Default	996	571	425
Non-default	9503	5748	3755
Observations	10499	6319	4180

Table 2.3. *Description of the training and testing subsets.*

There are 37 variables in the data sample which can be used in a credit scoring model. I apply a forward selection method in order to choose the best performing predictors which should be included in a scoring model. For more technical details on variable selection techniques I refer to the paper by Burnham and Anderson (2002).

According to the forward selection method variables are included in the model one by one until they decrease AIC or BIC criteria (Akaike, 1974; Schwartz, 1978). I use a logistic regression in the selection process. The resulting set of explanatory variables consists of 12 variables which are then used in the credit scoring models within this chapter. Importantly, these explanatory variables are individual-level variables. Microenvironment-level variables included in the credit scorecards in section 2.3 are not given in this set.

Additionally, I report the correlation matrix between all explanatory variables in order to check for multicollinearity. The table with the correlation coefficients is provided in Appendix I. The results confirm that multicollinearity is not a problem here: the correlation between explanatory variables is low (the highest absolute value is 0.15).

2.3 Empirical analysis

This section provides an empirical analysis for the multilevel credit scoring models. I introduce and fit several versions of the credit scorecards which differ by the composition of the random-effects and group-level variables. All scoring models are specified with a two-level structure where borrowers are the level-one units which are nested within microenvironments, the level-two groups. The two-level structure allows recognizing the microenvironment-specific effects which are given by the random-effects in the models.

I begin by providing the empirical results for the multilevel credit scoring model which includes only a random-intercept for the microenvironments. Then, this scoring model is elaborated to include more area-specific effects. Subsection 2.3.2 introduces a credit scorecard which additionally to the microenvironment-specific intercept specifies several group-level variables. Including group-level characteristics improves the estimation of the area-specific intercepts. The scoring model in subsection 2.3.3 extends the random-intercept scorecard and allows the coefficients of two individual-level variables to vary across microenvironments. Finally, I present a very flexible version of a multilevel scoring model which includes multiple random-coefficients, group-level variables and interacted variables (interactions between the individual-level and group-level variables).

As discussed earlier I apply the training data sample for the model fitting. The testing dataset is used for the calculation of the postestimation diagnostics which include different assessments of the scorecards' performance and the predictive accuracy check. In particular, I apply a ROC analysis and several other statistical measures to test the forecasting quality of the scorecards and compare the discriminatory power between scoring models as given in section 2.4. Additionally, I report the classification table for the optimal cut-off point, sensitivity and specificity pairs and calculate areas under the ROC curve.

The credit scoring models presented in this chapter are fitted in Stata and MLwiN (StataCorp. 2007, Centre for Multilevel Modelling, University of Bristol (2009)) by maximum likelihood. There are several approaches proposed in the literature to estimate a logistic regression and its extension – a multilevel logistic regression. In the thesis I follow the literature and apply two of the most frequently used techniques: maximum likelihood and Bayesian Markov chain Monte Carlo. Bayesian MCMC is applied to fit the complex credit scoring models with cross-classified structures which are presented in chapter 4. It should be mentioned that the estimation methods are not the main topic of this dissertation. Accordingly, I do not provide a comprehensive description of the technical details of the estimation within this chapter. Instead, I denote a single chapter 5 'Estimation' where a brief summary is given for the estimation with maximum likelihood and Bayesian MCMC. Chapter 5 also reviews the main advantages/disadvantages of the estimation approaches and discusses which method is more appropriate for fitting a multilevel scoring model with different combination of random-effects.

In order to compare the predictive performance between the multilevel scoring models and a conventional scorecard I report the empirical results for the logistic regression scorecard first. The logistic regression scorecard is specified in [2.1]. The dependent variable y_i is binary (0,1). It equals one for the defaulted borrowers and zero for the non-defaulters. The model assesses credit worthiness of an applicant for a loan by estimating the probability of default.

$$\Pr(y_i = 1|x_i) = \quad Logit^{-1}(\beta_0 + \gamma_1 Income_i + \gamma_2 Dependents_i + \gamma_3 Trade_{accounts_{ij}}$$

$$+ \quad \gamma_4 Bank_{ij} + \gamma_5 Enquiries_i + \gamma_6 Professional_i + \gamma_7 DR_i +$$

$$+ \quad \gamma_8 R_{credits_i} + \gamma_9 Credit_i + \gamma_{10} Past_{due_i} + \gamma_{11} Own_i). \qquad [2.1]$$

I apply a forward selection to choose the variables which are then included in the scorecards. The resulting set contains 12 individual-level variables. I will use the same set of variables for all credit scoring models within the chapter. The explanatory variables are annual income of a borrower ($Income_i$), the number of dependents in the family ($Dependents_i$), the number of current trade accounts ($Trade_{accounts_i}$), a dummy variable for customers who use bank savings and checking accounts ($Bank_i$), the number of previous credit enquiries ($Enquiries_i$), a dummy variable for high-skilled professionals ($Professional_i$), the number of derogatory reports in the credit profile of a borrower (DR_i), the average revolving credit balance ($R_{credits_i}$), a dummy variable for borrowers who have previous experience with a lender such as a personal loan or a credit card ($Credit_i$), the number of 30-days delinquencies on the credit accounts in the past 12 months ($Past_{due_i}$) and a dummy variable for the borrowers who own a house or a flat (Own_i). The logistic regression scorecard is estimated in Stata by maximum likelihood.

Variable	*Coefficient*	*Std.error*	*z*	*P>/z/*
Total Income	-0.043	0.003	-14.06	<0.001
Number of dependents	0.088	0.023	3.54	<0.001
Trade accounts	-0.049	0.005	-7.90	<0.001
Bank accounts (ch/ savings)	-0.346	0.061	-5.57	<0.001
Enquiries	0.392	0.012	30.80	<0.001
Professional	-0.369	0.106	-3.42	<0.001
Derogatory Reports	0.625	0.023	27.19	<0.001
Revolving credit balance	0.013	0.003	3.38	<0.001
Previous credit	-0.091	0.030	3.03	0.005
Past due	0.306	0.055	5.54	<0.001
Own	-0.053	0.066	-0.80	0.420
Constant	-1.380	0.104	-13.20	<0.001

Table 2.4. *Estimation results for the logistic regression credit scorecard.*

The estimation results are presented in Table 2.4. The first column in the table gives a variable name. The second and the third columns provide coefficient estimates and standard errors. The forth column reports a t-test (or z-test) to check the null hypothesis that a coefficient equals zero. The last column provides the p-value for the corresponding two-sided test.

The interpretation of the coefficients in the case of a generalized linear model is not straightforward as in the linear case. To interpret the estimates I calculate marginal effects of the explanatory variables by taking the first derivative, dy/dx. Table 2.5 presents the results. I set the continuous variables at their mean values ($\bar{x}$) while calculating marginal effects. In the case of an indicator variable the marginal effect is the change in the probability given a discrete change of a dummy variable from 0 to 1. The other dummy variables are specified to take a value of one while calculating marginal effects. I denote dummy variables by * in the table.

Variable	dy/dx	*Std.err.*	$\bar{x}$
Total Income	-0.0033	0.0002	30.11
Number of dependents	0.0068	0.0010	1.02
Trade accounts	-0.0038	0.0004	7.17
Bank accounts (ch/ savings)*	-0.0860	0.0050	
Enquiries	0.0301	0.0010	1.42
Professional*	-0.0910	0.0060	
Derogatory Reports	0.0481	0.0020	0.46
Revolving credit balance	0.0009	0.0002	5.28
Previous credit *	-0.0220	0.0060	
Past due	0.0240	0.0040	0.15
Own*	-0.0140	0.0050	

Table 2.5. *Marginal-effects for variables in the logistic regression scorecard evaluated at the mean value, $\bar{x}$. The mean values of independent variables are given in the forth column.*

The results in Table 2.5 confirm that the probability of default decreases with higher income, higher number of trade accounts, if a borrower holds any banking savings and checking accounts and if an applicant is a college graduate. In particular, a ten thousands increase in total income decreases the probability by 3.3% given that the other continuous variables are taken at their mean values and the dummy variables are equalized to 1. The probability of default is 2.2% smaller if a borrower had previous experience with a lender such as a personal loan or credit card. It is also true that high-skilled professionals are less risky than other borrowers. In contrast, unsatisfactory credit history records such as a high number of derogatory reports or delinquencies on the past credit obligations have a positive impact on the riskiness of customers.

An alternative to this method is a rule of thumb 'divide-by-four' which is a quick way of calculating marginal effects. According to this rule the marginal effect of a continuous variable can be approximated by dividing its estimated coefficient by 4. This gives an upper bound for the change in the dependent variable given a unit change in a predictor variable (Gelman (2007)).

In order to check the goodness-of-fit of the logistic scorecard I apply several postestimation diagnostics. Table 2.6 provides the results for the likelihood ratio test, Hosmer-Lemeshow chi-squared test, pseudo R-squared, Akaike information criterion (AIC) and Bayesian information criterion (BIC). Later, I will apply these results to compare the goodness-of-fit and predictive quality between different multilevel credit scoring models and the logistic scorecard. I refer to Schwarz (1978) and Akaike (1974) for more technical details on information criteria.

Postestimation statistics		*p-value*
Likelihood ratio chi-square test	1730.6	<0.001
Hosmer-Lemeshow goodness-of-fit test	40.50	<0.001
Pseudo R^2 (full model)	0.2775	
Pseudo R^2 (reduced model)	0.0234	
AIC	4583.13	
BIC	4667.07	

Table 2.6. *Postestimation diagnostics for the logistic regression scorecard.*

As might be expected, the likelihood ratio test for the logistic scorecard shows that a logit model with a full set of variables performs much better than a reduced form model with only an intercept.

The pseudo R^2 and Hosmer-Lemeshow chi-square statistics assess the goodness-of-fit of the logistic scorecard. The Hosmer-Lemeshow statistics tests the null hypothesis that there is no difference between the observed and predicted values of the dependent variable (Hosmer and Lemeshow (2000)). Based on the result for the test given in the table I reject the null hypothesis. This implies that the model does not fit the data on the acceptable level.

The pseudo R^2 is McFadden's adjusted R^2 which provides a logistic regression analogy to the standard R^2 in OLS regression. For more technical details on the calculation I refer to Agresti and Zheng (2000) and McFadden (1973). It is evident that the pseudo R^2 for the scoring model with the full set of explanatory variables is higher compared to the reduced model with only an intercept. However, the value of the pseudo R^2 for the full model is still rather low suggesting that the logistic scorecard poorly predicts the outcome and further improvements are possible.

The main aim of any scoring model is to measure credit worthiness of a borrower by forecasting the probability of default. Accordingly, concluding the presentation of the empirical results for the logit scorecard I provide the assessment of the model performance. I apply the same assessments to each credit scorecard discussed in this chapter and then summarize and compare the results for the different scorecards in chapter 3.

I begin by evaluating the discriminatory power of the logistic regression scorecard. For this purpose the classification is given in Table 2.7. This table summarizes the performance of the fitted scoring model given a specified cut-point (Fawcett (2005)). A cut-point is a threshold which is used to discriminate borrowers' scores (or predicted probabilities) into two classes: default (D) and non-default (ND). Table 2.7 reports results for two cut-off points (probabilities of default), $c_1 = 0.15$ and $c_2 = 0.50$. These thresholds are frequently used in credit scoring.

Given the cut-off point $c_2 = 0.5$, four possible outcomes are observed. If the outcome is positive and the scoring model classifies it as positive (true positive, 78); if it is classified as negative but the true outcome is positive, it is counted as a false negative (347). If the outcome is negative and it is classified as negative, it is counted as true negative (3630); if it is classified as positive, it is counted as false positive (125). The predictions along the major diagonal in the table represent the correctly classified outcomes and the off-diagonal elements are the misclassified outcomes or errors.

	True (c_1 = 0.15)			**True** (c_2 = 0.5)		
Classified	D	ND	Total	D	ND	Total
D	180	664	844	78	125	203
ND	245	3091	3336	347	3630	3977
Total	425	3755	4180	425	3755	4180
Correctly classified, %			78.25			88.70
False D rate for true ND (FPR), %			17.68			3.34
False ND rate for true D (FNR), %			57.60			81.63
ROC curve metrics:						
Area under the ROC (AUC)						0.707
Standard error (AUC)						0.008
95% Confidence Interval (AUC)						[0.698; 0.716]

Table 2.7. *The classification table for the logistic credit scorecard given the cut-off points for probability of default:* $c_1 = 0.15, c_2 = 0.5$. *Summary results for the ROC curve analysis, area under the ROC curve.*

Based on the results in the classification table several common metrics can be calculated. The overall rate of correct classifications equals *(1-MR),* where MR is the misclassification rate which shows the proportion of incorrectly predicted outcomes. For the cut-point $c_1 = 0.15$ the calculation of the misclassification rate is provided in [2.2] where the denominator is the sum of the false negative and false positive classifications.

$$MR = \frac{FN+FP}{TP+FN+FP+TN} = 21.75\%. \qquad [2.2]$$

The most common way of reporting the accuracy of a binary prediction is to analyze the true (false) positive and true (false) negative outcomes separately. This implies that a falsely classified negative prediction may have different consequences than a false positive one. In retail banking, a falsely classified non-defaulter is much less costly for a lender than an incorrectly classified defaulter. The false positive rate gives the proportion of the false positive outcomes in the total negative as shown in [2.3]. The false positive rate is applied to calculate specificity which equals (1-FPR). The false negative rate shows the proportion of incorrectly classified negative predictions as given in [2.4]. True negative rate is also called sensitivity. It is derived using FNR: $sensitivity = 1 - FNR$.

$$FPR = \frac{FP}{FP+TN} = 17.68\%, \qquad [2.3]$$

$$FNR = \frac{FN}{FN+TP} = 57.60\%. \qquad [2.4]$$

It should be noted that the resulting values of the misclassification rate, true positive rate and true negative rate, are very sensitive to the choice of a particular cut-off point. Table 2.7 shows that the false negative rate at the cut-off point $c_1 = 0.15$ is 57.60% which is 23.83% smaller than the false negative rate at the cut-off point $c_2 = 0.5$. In order to provide a more general representation of the sensitivity/specificity pairs I apply a ROC-curve analysis to the fitted model predictions.

The receiver operating characteristic (ROC) curve is a technique which visualizes the performance of a predictive model. This method has long been applied in medicine for a diagnostic testing (Zou (2002), Swets (2000)). A standard ROC plot is a two-dimensional graph which provides a graphical illustration of the true positive rate (on the ordinate axis) versus false positive rate for all possible cut-off points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold. A model which provides a perfect discrimination has a ROC curve which passes through the upper left corner (100% sensitivity, 100% specificity). Accordingly, the predictive model is considered to show a higher discriminatory power if its ROC curve is closer to the upper left corner than curves for the other models (Zweig and Campbell (1993)).

In application to retail banking, a ROC curve shows the relative trade-off between benefits a lender gets by correctly classifying defaulters (true positive) and costs he acquires by incorrectly classifying non-defaulters (false positive). The receiver operating characteristics curve for the logistic regression scorecard is illustrated in Figure 2.1.

The diagonal line in the plot represents the case when a credit scoring model randomly assigns borrowers into a class of defaulters and non-defaulters. In this case the model is expected to predict one half of the positive and one half of the negative outcomes correctly. The ROC curve for the logistic scorecard is over the diagonal line.

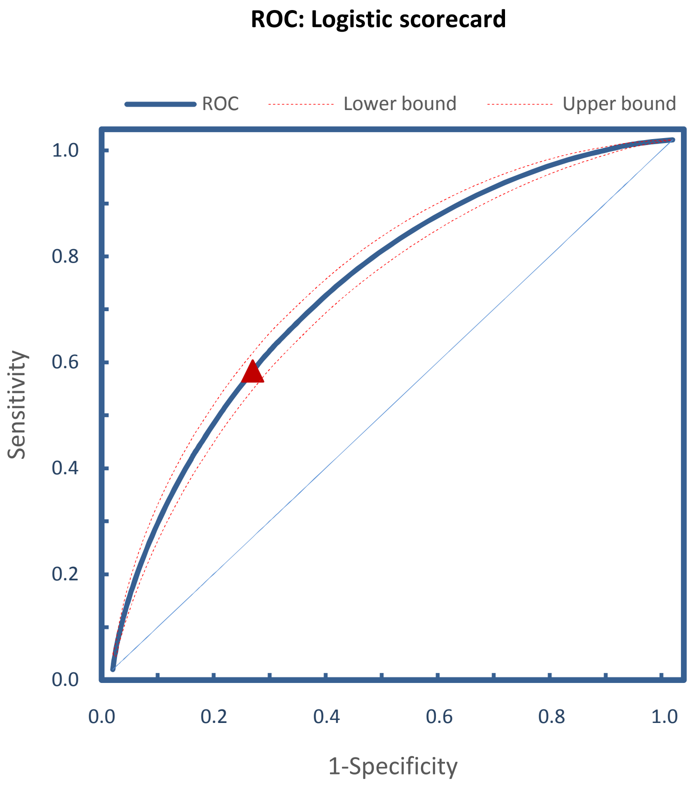


Figure 2.1. *ROC curve and pointwise confidence bounds for the logistic regression scorecard.*

The discriminatory power of a model is usually evaluated by calculating the area under a ROC curve (Hanley and McNeil (1982), Bradley (1997) and Hand (2005)). If S_{ND} and S_D are the scores given to the randomly and independently chosen individuals from D (defaulters) and ND (non-defaulters) respectivly, then $AUC = p(S_D > S_{ND})$.

If the ROC curve is defined as $y = f(x)$, where y is the true positive rate and x is the false positive rate, then $x(t) = p(S > t|ND)$ is the false negative rate given a threshold t and $y(t) = p(S > t|D)$ is the true positive rate given a cut-off point t. Area under the ROC curve is the average true positive rate, taken over all possible false positive rates within the range (0;1).

$$AUC = \int_0^1 y(x)dx.$$

Given that the slope of the ROC curve at the point with threshold value t is $\frac{dy}{dx}$ and $x \to 0$ as $\to \infty$, and $x \to 1$ as $t \to -\infty$, the AUC can be calculated as follows:

$$\begin{aligned} AUC &= \int_{+\infty}^{-\infty} y(t)\frac{dx}{dt}dt \\ &= -\int_{+\infty}^{-\infty} p(S > t|D)\ p(t|ND)dt \\ &= \int_{-\infty}^{+\infty} p(S > t|D)\ p(t|ND)dt \\ &= \int_{-\infty}^{+\infty} p(S_D > t \ and \ S_{ND} = t|D)\ dt \\ &= \int_{-\infty}^{+\infty} p(S_D > S_{ND}|t)\ dt \\ &= p(S_D > S_{ND}). \end{aligned}$$

The AUC is always between 0 and 1. An optimal credit scorecard precisely separates between defaulted and non-defaulted borrowers. In this case the area under the ROC curve equals 1. A suboptimal scorecard would randomly assign probabilities to credit applicants. The ROC curve for a random guess model is given by the diagonal line and the area under the curve equals 0.5 in this case.

The AUC can be directly related to the Gini coefficient (Gini) which is based on the Lorenz curve and its accuracy ratio (Keenan and Sobehart (1999), Engelmann et al. (2003)). The only difference between a ROC curve and a Lorenz curve is that the former plots false positive rate versus true positive rate and the latter graphs true positive rate given the percentage of borrowers (Engelmann et al. (2003)). Similarly to the AUC, the quantitative measure of discriminatory power of a scoring model can likewise be based on the area between the Lorenz curve and the diagonal line. It is called the Gini coefficient and equals twice this area as follows:

$$\begin{aligned} Gini &= (2AUC - 1) \\ &= AR\,/Pr(ND), \\ AR &= \frac{Gini}{Gini_{optimal}} = \frac{Gini}{Pr(ND)}, \end{aligned}$$

where $p(D)$ and $p(ND)$ are the probabilities of default and non-default correspondingly (Kraft et al. (2002)).

The ROC curve metrics and the correct classification rate are given in Table 2.7. It also reports the AUC value, the Gini coefficient and the accuracy ratio for the logistic regression scorecard. The area under the ROC curve of the logit scorecard equals 0.707 which is fairly better than AUC of a random guessing model (AUC=0.5), however, further improvements are possible. Additionally, I report the standard error of the area under the ROC and the 95% confidence interval for the AUC value. The 95% interval for the AUC is [0.698; 0.716]. The standard error is computed by applying a nonparametric approach which is described in detail by Hanley and McNeil (1982).

An important feature of the AUC is that it allows comparing the discriminatory power of different credit scoring models fitted to the same dataset. In this sense the AUC measure can be applied to select a model which shows the best performance. The next chapter provides a comparison of the AUC measures. In addition, I test whether the differences between the areas are significant using a z-test as described in detail in Hanley and McNeil (1983) and DeLong (1988).

2.3.1 Microenvironment-specific intercept scorecard

The following four subsections introduce several versions of the multilevel credit scorecards which differ by the combination of random-effects and group-level variables. Similarly to the logit scorecard, I conclude the presentation of the empirical results by providing the assessment of the predictive accuracy and other postestimation diagnostics. These results are of particular interest in the thesis as they aim to summarize the advantages of the multilevel scoring models over the logistic regression scorecard. The summary of the ROC curve analysis, the discriminatory power check and goodness-of-fit tests are presented in chapter 3.

The credit scoring model in this section is an extension of the logistic regression scorecard. It applies a two-level structure to the logistic scoring model and allows the intercept to vary at the second level of the hierarchy. The model specifies a random-intercept to determine the microenvironment-specific effects which represent the living areas with different economic and socio-demographic conditions. Including a random-intercept in the scorecard helps to relax the main assumption of the logistic regression of the conditional independence among responses for the same cluster (microenvironment) given other explanatory variables.

The two-level credit scorecard with a varying-intercept and the individual-level explanatory variables is presented in [2.5].

$$\begin{aligned} Pr(y_i = 1 | x_i, u_{j,0}) = \ & Logit^{-1}(\alpha_{j[i]} + \gamma_1 Income_i + \gamma_2 Dependents_i + \gamma_3 Trade_{accounts_i} \\ & + \gamma_4 Bank_i + \gamma_5 Enquiries_i + \gamma_6 Professional_i + \gamma_7 DR_i + \gamma_8 R_{credit_i} \\ & + \gamma_9 Credit_i + \gamma_{10} Past_{due_i} + \gamma_{11} Own_i), \qquad [2.5] \end{aligned}$$

$$\alpha_{j[i]} = \alpha_0 + u_{j,0}\,, \qquad [2.6]$$

$$u_{j,0} | x_i \sim N\,(0, \sigma_u^2), \qquad for\ microenvironment\ j = 1, .., 61.$$

Given the explanatory variables, the random-intercept follows a normal distribution with mean α_0 and variance σ_u^2. This is a standard assumption in multilevel modelling which implies that random-effects for the microenvironment j are drawn from a normal distribution with mean α_0 and variance σ_u^2. Importantly, it is possible to define other types of probability distributions for the random-intercept if there is prior knowledge about the distributional type of the random-effects from similar studies or from other sources. Unfortunately, this is not the case here as the hierarchical structure with borrowers-within-microenvironments has never been explored in the literature on credit scoring before. Therefore, this thesis follows the widely accepted practice and assigns a normal distribution to the microenvironment-specific effects.

The main difference between the formulation of the scoring model in [2.5] and [2.6] and the logit scorecard in [2.1] is that in the former model the intercept is allowed to vary across groups at the second-level and it is specified with the subscript j . The varying-intercept is modelled as given in [2.6].

The second-level model for the random-intercept includes a population average intercept γ_0 and a random term $u_{j,o}$. The residual $u_{j,o}$ determines the unobserved characteristics of a microenvironment which influence the probability. This area-specific effect is not included in the single-level logistic scorecard. The microenvironment-specific effect can be viewed as the aggregated impact of the unobserved determinants which explains why some borrowers are more risky (have higher probability of default) than others. In other words, the random-effect helps to account for the unobserved heterogeneity in the probabilities of default between borrowers within different microenvironments.

Consider two applicants for a loan (with the same personal characteristics) whose microenvironments are very different in terms of economic conditions: poor and rich living areas. Accordingly, the exposure to risk in the poor living area with low income, high unemployment, and bad infrastructure of shopping facilities is higher compared to the rich living area where average income is high, unemployment is low and infrastructure of shopping facilities is well-developed. In this case the microenvironment-specific effects are unobserved characteristics which impact the probability of default additionally to individual-level characteristics of a borrower.

In the credit scoring model the exposure to the poor (rich) area-specific risks is captured by the random-effect $u_{j,o}$. This implies that the intercept for the particular microenvironment j differs from the population average intercept by the value $u_{j,o}$. In order to illustrate this graphically I plot the fitted model lines for ten randomly chosen microenvironments. The graph is given in Figure 4.2.

The abscissa axis in Figure 4.2 shows the linear part of the prediction given by $x'\hat{\beta} = Logit[Pr(y = 1|x)]$ without the microenvironment-specific effect $\hat{u}_{j,0}$

and the ordinate axis illustrates the linear predictor with the area-specific effect as given by $(x'\hat{\beta} + \hat{u}_{j,0}) = Logit[Pr(y = 1|x)]$.

This graphical illustration of the microenvironment-specific lines confirms that a credit scoring model with the two-level structure provides more flexible modelling of the probabilities than a conventional logistic regression.

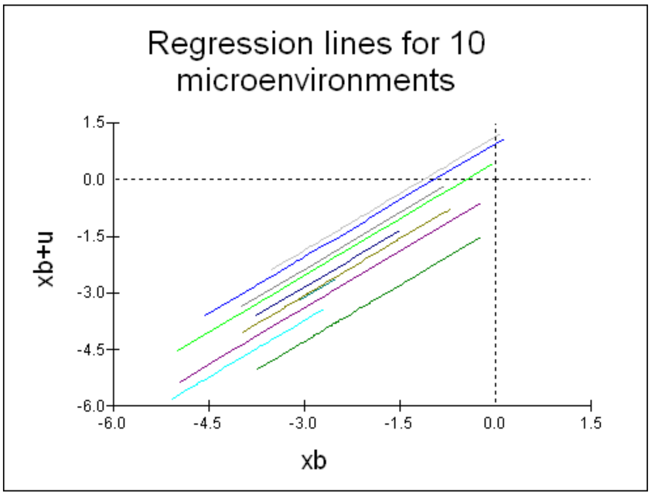


Figure 2.2. *Fitted model regression lines for ten randomly chosen microenvironments. The abscissa axis gives the linear part of prediction excluding the microenvironment-specific intercept. The x'b+u is the linear part of the prediction including the area-specific intercept.*

The estimation results for the two-level credit scoring model with microenvironment-specific intercept are presented in Table 2.8. The estimated coefficients for the varying-intercept scorecard in [2.5] are similar to the estimates obtained from the logistic scorecard in [2.1].

The last row in the table provides the estimate of the standard deviation of the random-intercept with the standard error given in the brackets. The standard deviation is large suggesting that there is a considerable variation in the area-specific intercepts among different microenvironments. Importantly, this variability was not captured in the logistic regression scorecard.

Variable	*Coefficient*	*Std.err.*	*z*	*P>/z/*
Total Income	-0.044	0.004	-9.88	<0.001
Number of dependents	0.113	0.033	3.45	<0.001
Trade accounts	-0.039	0.008	-5.01	<0.001
Bank accounts (ch/ savings)	-0.427	0.082	-5.19	<0.001
Enquiries	0.376	0.017	22.48	<0.001
Professional	-0.327	0.093	-3.50	<0.001
Derogatory Reports	0.622	0.030	20.65	<0.001
Revolving credit balance	0.015	0.004	3.46	<0.001
Previous credit	-0.059	0.019	3.16	<0.001
Past due	0.239	0.074	3.22	<0.001
Own	-0.321	0.109	-2.94	0.003
Constant	-1.270	0.211	-6.01	<0.001

Random-effects	*Estimate (Std.err.)*	*95% Confidence interval*
Standard deviation of intercept, (σ_u)	0.61(0.09)	[0.43; 0.81]
Random-intercept 95% CI, $(\alpha_{j[i]})$		[-2.50;-0.07]

Table 2.8. *Estimation results for the two-level credit scoring model with microenvironment-specific intercepts. The estimated standard deviation and its 95% confidence interval, 95% confidence interval for the random-intercept.*

The 95% confidence interval for the microenvironment-specific intercept is reported in the last row in Table 2.8. Given the normality assumption the confidence interval for the varying-intercept is calculated based on the following formula: $CI = [\widehat{\gamma_0} \pm 1.96 \cdot \hat{\sigma}_{u_o}]$, where $\widehat{\gamma_0}$ is the estimated population average intercept and $\hat{\sigma}_u$ is the standard deviation of the random-intercept. For a more detailed description on the calculation of confidence intervals for the random-effects in a multilevel model I refer to Rabe-Hesketh (2008). The confidence interval for the microenvironment-specific intercept equals $[-2.5; -0.07]$.

Similar to the logistic scorecard, I evaluate the performance of the multilevel credit scorecard by applying a ROC curve analysis. Figure 2.3 presents the ROC curve for the microenvironment-specific intercept scorecard given in [2.5]. Following Hilgers (1991), I also display 95% pointwise confidence bounds for the curve. The red triangle on the graph indicates the optimal cut-off point. This value provides a criterion which yields the highest accuracy (minimal false negative plus false positive rate).

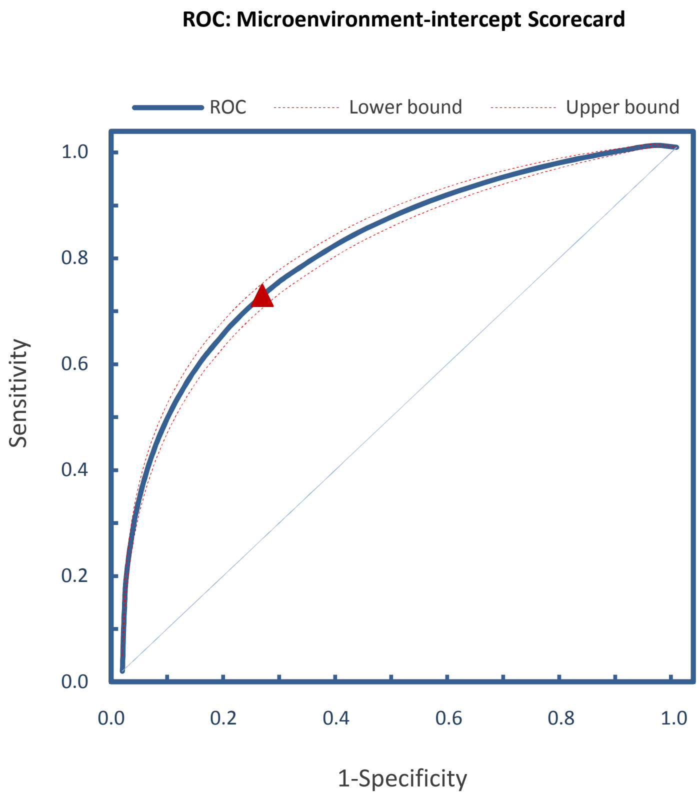


Figure 2.3. *ROC curve for the two-level credit scoring model with microenvironment-specific intercept. The optimal cut-off point is* $c_1 = 0.1373$*.*

I should mention that this optimal threshold is only optimal with respect to minimizing the total misclassification error. However, it is possible to compute other cut-off points which are optimal according to a specified rule or given a budget constraint. For instance, in a cost-benefit analysis an optimal cut-off point defines a threshold at which the costs are minimized (Krämer and Bücker (2009)). I do not provide a detailed discussion of these alternatives in the thesis because the decision about the cut-off point is generally driven by practical considerations within a bank. Given a scorecard a lender assesses the costs and benefits associated with different cut-off points and then decides which one satisfies the budget constraints and legislation requirements.

The summary results derived from of the ROC curve in Figure 2.3 and the classification table for the optimal cut-off point are presented in Table 2.9.

	True		
Classified ($c_1 = 0.1373$)	D	ND	Total
Default	293	1002	1295
Non-default	132	2753	2885
Total	425	3755	4180
Correctly classified, %			72.87
Sensitivity, %			69.00
Specificity, %			73.31
ROC curve metrics:			
Area under the ROC (AUC)			0.801
Standard error (DeLong)			0.005
95% confidence interval			[0.794;0.808]
Gini coefficient			0.602
Accuracy ratio			0.663

Table 2.9. *Summary metrics for the ROC curve of the microenvironment-specific intercept model and the classification table for the optimal cut-off point:* $c_1 = 0.1376$.

The optimal threshold for the microenvironment-specific intercept model in [2.5] is $c_1 = 0.1373$ (minimal misclassification error). Selecting a threshold above 0.1373 increases the proportion of true negative classifications (increased specificity) but decreases the fraction of true positive classifications (reduced sensitivity). Selecting a cut-off below 0.1373 refers to the case when a scoring model predicts a higher fraction of true positive outcomes (increased sensitivity) but a smaller fraction of true negative outcomes (decreased specificity).

The area under the ROC curve is 0.8015 which is higher than in the case of the logistic regression scorecard. The Gini coefficient and the accuracy ratio are also increased. It implies that specifying microenvironment-specific intercepts improves the discriminatory power of the credit scoring model.

The 95% confidence interval for the AUC shows the bounds in which the true area under the ROC curve lies with 95% confidence ([0.794; 0.808]). Importantly, this interval is narrow and does not overlap with the confidence interval for the logistic regression scorecard.

2.3.2 Group-level variables in the two-level credit scorecard

This subsection introduces the two-level credit scoring model which includes group-level characteristics. The scorecard is presented in [2.7]. It expands the random-intercept scorecard given in [2.5] by inserting the microenvironment-level characteristics in the second-level model for the varying-intercept $\alpha_{j[i]}$. The microenvironment-level variables are denoted by $z'\beta$ in the second-level model. Specifying group-level characteristics in a scorecard helps to explore the impact of the microenvironment-level information on the probability of default. It also improves the estimation of the area-specific intercepts.

Similarly to the previous case, the area-specific intercept is modelled as given in [2.8]. Additionally to the population average intercept α_0 and the random term $u_{j,0}$ the model for the varying-intercept now includes four microenvironment-level variables $z_{j,m}$, for *m=1,..,4.* The group-level variables $z_{j,m}$ vary across *J=61* microenvironments but take the same value for all borrowers $i = 1,..,n_j$ within a given microenvironment j.

$$\begin{aligned} Pr\left(y_{ij} = 1 \middle| x_{ij}, u_{j,0}\right) = \quad & Logit^{-1}(\alpha_{j[i]} + \gamma_1 Income_i + \gamma_2 Dependents_i + \gamma_3 Trade_{accounts_i} \\ & + \gamma_4 Bank_i + \gamma_5 Enquiries_{ij} + \gamma_6 Professional_i + \gamma_7 DR_i \\ & + \gamma_8 R_{credit_i} + \gamma_9 Credit_i + \gamma_{10} Past_{due_i} + \gamma_{11} Own_i), \qquad [2.7] \end{aligned}$$

$$\begin{aligned} \alpha_{j[i]} &= \alpha_0 + z'\beta + u_{j,0}, \\ z'\beta &= \beta_1 Area\_Income_j + \beta_2 AA_{residents_j} + \beta_3 Stores_j + \beta_4 College_j, \qquad [2.8] \\ u_{j,o}\, |x_i\,, z_j &\sim N\,(0, \sigma_u^2). \end{aligned}$$

Microenvironment-level variables characterize the economic and demographic conditions in the borrowers' residence areas. The variables are $Area_{Income_j}$- average income in the living area j (measured in thousands of dollars); $Stores_j$ -percentage of retail, furniture, building materials and auto store sales in the total retail sales in the market; $College_j$ - percentage of college graduates in the residence area and $AA_{residents_j}$ – the share of African-American (Hispanic) residents in the region.

The two-level credit scoring model with the microenvironment-level variables and a varying-intercept is fitted in Stata by using maximum likelihood. Table 2.10 provides the estimated coefficients of the individual and group-level variables, and the standard deviation of the area-specific intercept.

The fixed-effect estimates of the individual-level variables are essentially the same as in the scorecard presented in [2.5]. This is quite reasonable as including the microenvironment-level characteristics only modifies the random-intercept model. The standard deviation of the microenvironment-intercept is smaller than in the credit scoring model without group-level variables. This is due to the fact that the second-level characteristics partly explain the variation between microenvironments.

The estimated coefficients for the microenvironments-level variables show the impact of the living area conditions on the riskiness of applicants for a loan. Higher per capita income has a negative effect on the riskiness of a borrower. Similarly, the living area share of individuals with a university degree negatively impacts the probability of default. The result is intuitive and implies that the effect of higher education on default is negative not only at the borrower-level but also at the microenvironment-level.

In contrast, the impact of the variable share of African-American residents on default is significant and positive. The coefficient of $AA_{residents_j}$ explains how the demographic composition of residents in the area influences the probability of default. It is evident that borrowers within microenvironments with a large share of African-American and Hispanic residents have higher exposure to area-specific risks which trigger default.

Variable	*Coefficient*	*Std.err.*	*z*	*P>/z/*
Total Income	-0.041	0.004	-9.34	<0.001
Number of dependents	0.114	0.033	3.47	<0.001
Trade accounts	-0.038	0.008	-5.02	<0.001
Bank accounts (ch/ savings)	-0.426	0.082	-5.19	<0.001
Enquiries	0.373	0.017	22.40	<0.001
Professional	-0.332	0.096	-3.47	<0.001
Derogatory Reports	0.615	0.030	20.51	<0.001
Revolving credit balance	0.015	0.004	3.45	<0.001
Previous credit	-0.060	0.018	3.16	0.004
Past due	0.221	0.068	3.25	<0.001
Own	-0.285	0.100	-2.85	0.004
Constant	-0.860	0.210	-4.09	<0.001
Microenvironment-level variables, $\alpha_{j[i]}$				
Living area per capita income	-0.017	0.008	-2.12	0.033
Share of African-American residents	0.012	0.003	4.00	<0.001
Share of college graduates	-0.034	0.014	-2.42	0.015
Infrastructure of shopping facilities	0.037	0.029	1.27	0.204

Random-effects	*Estimate (Std.err.)*	*95% Confidence interval*
Standard deviation of intercept, σ_{u_o}	0.38(0.08)	[0.24; 0.59]

Table 2.10. *Estimation results for the two-level random-intercept model with microenvironment-level explanatory variables. The random-intercept variance is given in the last row in the table.*

The effect of the infrastructure of shopping facilities on default is positive. One possible interpretation of the result may be that a good access to various department stores and shopping malls provokes spending and initiates borrowing. In addition, I use in the empirical analysis the credit history data on the consumer loans which individuals regularly use for making small purchases of durable goods, buying cars or covering medical bills.

I apply a ROC curve analysis to assess the classification performance of the credit scorecard with group-level characteristics and a varying-intercept. Figure 2.4 shows the ROC curve and pointwise confidence bounds.

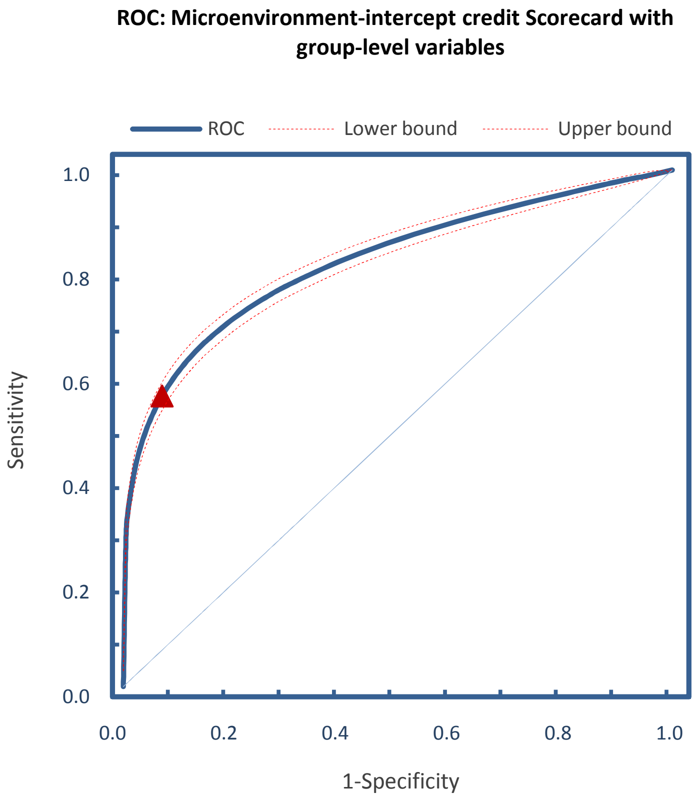


Figure 2.4. *ROC curve for the two-level credit scoring model with an area-specific intercept and group-level variables. The optimal cut-off point is indicated by the red triangle (*$c_1 = 0.2264$*).*

The summary of the ROC curve analysis, the Gini coefficient and a classification table for the optimal cut-off point are provided in Table 2.11.

The area under the ROC curve and the Gini coefficient are increased. The AUC is 0.017 higher than in the case of the credit scoring model without the microenvironment-level variables. The difference is not large; however, the 95% confidence intervals for the AUC values do not overlap which implies the areas are significantly different from each other ([0.811; 0.825] versus [0.794; 0.808]).

Another important improvement of the current version of the credit scoring model over the scorecard without group-level variables is that the former model has a higher rate of correct classifications. The rate of correct classifications is calculated at the threshold which corresponds to the maximal sensitivity/specificity pair ($c_1 = 0.2264$). The specificity is also higher at this point.

	True		
Classified ($c_1 = 0.2264$)	D	ND	Total
Default	235	308	543
Non-default	190	3447	3637
Total	684	3755	4180
Correctly classified, %			87.16
Sensitivity, %			55.22
Specificity, %			91.81
ROC curve metrics:			
Area under the ROC (AUC)			0.818
Standard error			0.005
95% confidence interval			[0.811; 0.825]
Gini coefficient			0.636
Accuracy ratio			0.701

Table 2.11. *Summary for the ROC curve analysis and the classification table for the optimal cut-off point,* $c_1 = 0.2264$, *for the microenvironment-intercept scorecard with the group-level variables.*

2.3.3 Microenvironment-specific coefficients in the credit scoring model

In the case of the logistic credit scoring model, a coefficient estimate shows the population average effect of an explanatory variable which is fixed for all applicants and microenvironments. In this section I relax this assumption and show how to elaborate the varying-intercept scoring model by allowing the coefficients to vary across microenvironments.

Specifying area-level coefficients makes a scorecard more flexible and improves the estimation. The area-specific coefficients combine information on the unobserved microenvironment-specific characteristics which impact default. In other words, a random-coefficient can be interpreted as an interaction effect between the individual-level and area-specific effects.

I specify random-coefficients for the two individual-level variables $Enquiries_i$ and $Past_{due_i}$. The motivation for this choice is the following: I suppose that the impact of these variables varies considerably between microenvironments with stable and unstable economic conditions. In particular, the area-specific coefficient of $Enquiries_i$ allows to measure the impact of credit enquiries on default for the customers within poor and rich living areas.

Importantly, multilevel modelling assumes that random-coefficients are drawn from some population of the microenvironment-specific effects. Therefore, parameters of these random-effects represent population values. The estimated variances and covariances of the random-coefficients show the variability in the population. Thinking in terms of population is relevant for a more efficient credit scoring because lenders are primarily interested in developing scorecards which can be easily applied to a new cohort of applicants for a loan. These borrowers may be sampled from other microenvironments which are not present in the current dataset. In this case, the estimated variances and covariances of random-effects can be applied to predict new area-specific effects.

The credit scoring model with the microenvironment-specific coefficients is presented in [2.9]. The two-level structure of the scorecard remains unchanged. The varying-intercept is modelled by itself at the second-level. I include in this model four group-level predictors $z'\beta$ whose coefficients do not vary by group. The main difference from the previous scorecard is that the coefficients on individual-level variables are allowed to vary across microenvironments. Random-effects at the second-level follow a multivariate normal distribution with zero mean and variance-covariance matrix Σ_u as shown in [2.11].

Models for the area-specific coefficients of the explanatory variables $Enquiries_i$ and $Past_{due_i}$ are given in [2.10]. Similar to the random-intercept model, the random-coefficient model $\beta_{j[i]}^{enq}$ includes a fixed-effect of credit enquiries γ_{enq} and a random-term $u_{j,enq}$. The second-level model for $\beta_{j[i]}^{past}$ contains the intercept γ_{Past} and an area-specific term $u_{j,Past}$.

$$\begin{aligned} Pr(y_i = 1 | x_i, z_j, u_{j,Enq}, u_{j,past}) = \ & Logit^{-1}(\alpha_{j[i]} + \gamma_1 Income_i + \gamma_2 Dependents_i + \gamma_3 Trade_{accoun_i} \\ & + \ \gamma_4 Bank_i + \ \beta_{j[i]}^{enq} \ Enquiries_i + \gamma_6 Professional_i \\ & + \ \gamma_7 DR_i + \gamma_8 R_{credits_{ij}} + \gamma_9 Credit_i + \beta_{j[i]}^{past} \ Past_{due_i} \\ & + \ \gamma_{11} Own_i), \end{aligned} \qquad [2.9]$$

$$\begin{aligned} z'\beta \ &= \ \beta_1 Area\_Income_j + \beta_2 AA_{residents_j} + \beta_3 Stores_j + \beta_4 College_j, \\ \alpha_{j[i]} &= \ \gamma_0 + z'\beta + \ u_{j,0}, \end{aligned}$$

$$\begin{aligned} \beta_j^{\ enq} \ &= \ \ \gamma_{enq} + u_{j,enq,} \\ \beta_j^{\ past} \ &= \ \ \gamma_{past} + u_{j,past}, \end{aligned} \qquad [2.10]$$

$$(u_{j,enq} \ , u_{j,past}, u_{j,0} | x_i, z_j,) \sim N\left(\begin{pmatrix} \gamma_0 \\ 0 \\ 0 \end{pmatrix}, \Sigma_{\mathrm{u}} = \begin{bmatrix} \sigma_{enq}^2 & \rho\sigma_{enq}\sigma_{past} & \rho\sigma_{enq}\sigma_u \\ \rho\sigma_{enq}\sigma_{past} & \sigma_{past}^2 & \rho\sigma_{past}\sigma_u \\ \rho\sigma_{enq}\sigma_u & \rho\sigma_{past}\sigma_u & \sigma_u^2 \end{bmatrix} \right). \ [2.11]$$

Given the individual-level and microenvironment-level variables, random-coefficients are allowed to be correlated and the correlation coefficient is given by ρ in the variance-covariance matrix in [2.11].

Table 2.12 provides the estimation results for the scoring model with microenvironment-specific coefficients. As might be expected, the probability of default decreases with higher annual income, experience in using credit and debit banking accounts, number of trade accounts and if the borrower owns a house. The effect of the number of derogatory reports is positive. Having previous experience with a lender significantly decreases the riskiness of a borrower. The effect of having a house or holding banking deposit accounts is negative. This makes sense as real estate property or other assets indicate financial stability of a borrower. These borrowers are also more reliable and have a higher incentive not to fall into arrears. In the case of default their property can be repossessed and deposit accounts can be attached by a bank. Compared to the borrowers who rent accommodation, house owners are 5.1% less risky.

Variable	*Coefficient*	*Std.err.*	*z*	*P>|z|*
Total Income	-0.037	0.003	-12.43	<0.001
Number of dependents	0.131	0.023	5.60	<0.001
Trade accounts	-0.037	0.007	-4.96	<0.001
Bank accounts (ch/ savings)	-0.384	0.059	-6.56	<0.001
Enquiries	0.380	0.021	17.95	<0.001
Professional	-0.312	0.100	-3.11	0.002
Derogatory Reports	0.605	0.038	15.81	<0.001
Revolving credit balance	0.011	0.004	2.91	0.003
Previous credit	-0.061	0.018	-3.40	<0.001
Past due	0.243	0.053	4.58	<0.001
Own	-0.215	0.081	-2.65	0.008
Constant	-1.380	0.100	-13.76	<0.001
Microenvironment-level model, $\alpha_{j[i]}$				
Living area per capita income	-0.075	0.038	-1.97	0.048
Share of African-American residents	0.008	0.002	3.80	<0.001
Share of college graduates	-0.025	0.011	-2.24	0.025
Infrastructure of shopping facilities	0.009	0.008	1.18	0.238
Random-coefficients		*Estimate (Std.err.)*		*95% Confidence interval*
Std .deviation of $\beta_{j[i]}^{enq}$ (credit enquiries)		0.122(0.019)		[0.089; 0.167]
Std .deviation of $\beta_{j[i]}^{past}$ (Past due)		0.169(0.074)		[0.071; 0.401]
Std .deviation of $\alpha_{j[i]}$		0.283(0.079)		[0.129; 0.448]
Correlation($u_{j,enq}$, $u_{j,past}$)		0.79		

Table 2.12. *Estimation results for the two-level microenvironment-specific coefficients credit scoring model: coefficients of the individual and group-level variables, standard deviations with their 95% confidence intervals and the correlation coefficient.*

The fixed-effect of the variable $Enquiries_i$ is 0.38 on the logit scale which is similar to the result for the scorecard without random-coefficients. The standard deviation of the microenvironment-specific slope $\beta_{j[i]}^{enq}$ is 0.122 with error

0.019. This implies that the area-specific slopes differ by ±3% on the probability scale.

Following Hox (2002) I calculate the confidence interval for the varying-coefficients. The 95%-confidence interval for the area-specific coefficient of credit enquiries equals [0.15; 0.61]. This interval shows the range within which 95% of the varying-coefficients are falling.

Similarly, the fixed-effect of past due accounts $Past_{due_i}$ is 0.243. The estimated standard deviation of the varying-slope is $\hat{\sigma}_{past_{due}}$= 0.169 on the logit scale. Translating it to the probability scale shows that the microenvironment-specific coefficient explains the change in the probability over and above the population average value by approximately ±4.3%. The confidence interval for the varying-coefficient $\beta_{j[i]}^{past}$ shows that in 95% of times the area-specific coefficients of the variable past due accounts are going to lie within the interval [-0.08; 0.57].

I check the discriminatory power of the credit scoring model with varying-coefficients and group-level variables by applying a ROC curve as shown in Figure 2.5. Following Hilgers (1991) I also display 95% confidence bounds for the curve. The threshold which yields the maximal sum of true positive and true negative rates is indicated by the red triangle on the graph. At this threshold the misclassification error rate is minimized.

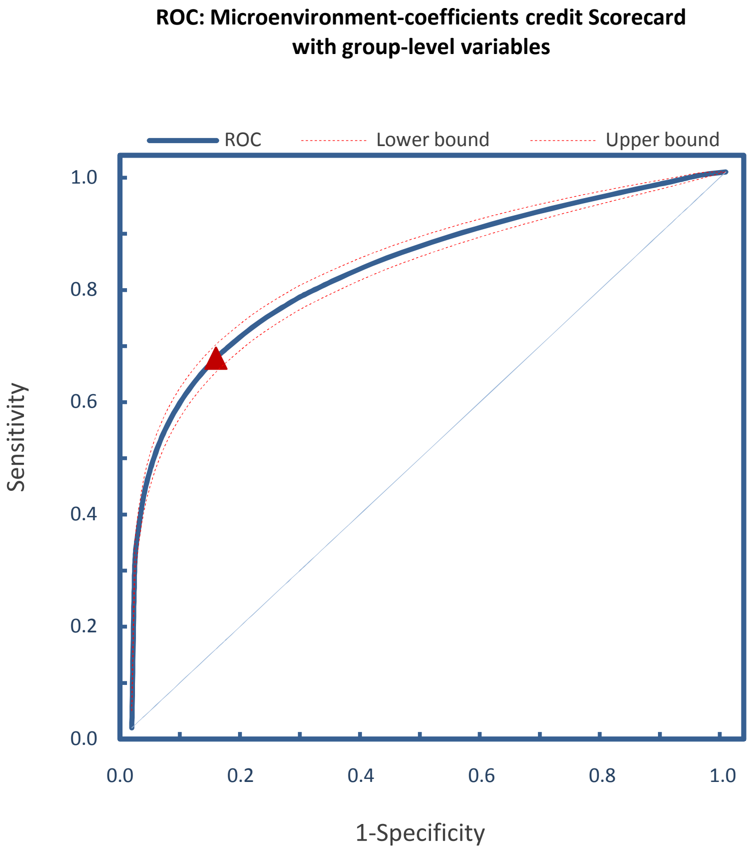


Figure 2.5. *ROC curve for the two-level credit scoring model with the area-specific coefficients and microenvironment-level variables. The optimal threshold (probability of default) is* $c_1 = 0.1406$.

The summary results derived from the ROC curve and the classification table for the optimal cut-off point ($c_1 = 0.1406$) are presented in Table 2.13. The area under the ROC curve is higher than in the case of the model without varying-coefficients. The AUC equals 0.824 and the 95% confidence interval for this value is [0.817; 0.83]. The confidence intervals for the microenvironment-coefficients model and the intervals for the area-specific intercept scorecard do not overlap which indicates that the current version of a scorecard improves the predictive accuracy. The Gini coefficient and the accuracy ratio are also increased.

Given the optimal cut-off point $c_1 = 0.1406$ the credit scoring model correctly classifies 80% of applicants for a loan. The true negative rate and the true positive rates are 81.9% and 65.6%, respectivly.

	True		
Classified ($c_1 = 0.1406$)	D	ND	Total
Default	279	680	959
Non-default	146	3075	3221
Total	425	3755	4180
Correctly classified, %			80.24
Sensitivity, %			65.60
Specificity,%			81.90
ROC curve metrics:			
Area under the ROC (AUC)			0.824
Standard error (DeLong)			0.005
95% confidence interval			[0.817; 0.830]
Gini coefficient			0.648
Accuracy ratio			0.714

Table 2.13. *Summary of the ROC curve analysis and the classification table for the optimal cut-off point:* $c_1 = 0.1406$, *for the two-level credit scoring model with the area-specific coefficients and microenvironment-level variables.*

2.3.4 Multiple random-coefficients credit scoring model

In this subsection I present a very flexible version of the credit scoring model which contains multiple random-coefficients, microenvironment-level variables and interacted variables. The scorecard extends the varying-coefficients scoring model presented in the previous subsection. Complementary to the previous structure, I specify two random-coefficients for the individual-level explanatory variables: the use of banking savings and checking accounts ($Bank_j$) and a house ownership indicator (Own_j).

The two-level model with multiple random-effects is presented in [2.12]. The scorecard includes four individual-level explanatory variables whose coefficients vary by microenvironment. The microenvironment-specific coefficients are modelled as shown in [2.14]. Each second-level model for the varying-coefficient includes a population average coefficient γ and the second-level residual u_j. Similarly, the varying-intercept model $\alpha_{j[i]}$ contains the constant term α_0, the second-level characteristics $z'\beta$ whose coefficients do not vary by group and the microenvironment-specific residuals $u_{j,0}$. The group-level coefficients in the microenvironment-intercept model are given by the 1×4 vector β. I fit the scorecard using the same set of the individual-level and group-level variables as in the previous scoring models.

The interactions between borrower-level and microenvironment-level variables are denoted by $k'\delta$ in [2.12]. I create three interacted variables which are $Past_{due_i} \cdot AA_{residents_{j[i]}}$ – the number of the delinquent credit accounts in the past measured at the borrower-level and the living area share of the African-American residents measured at the microenvironment-level; $Burden_i \cdot Stores_{j[i]}$ - access to various shopping facilities in the residence area and current credit burden of a borrower; and the variable $Adress_i \cdot Ownership_{Area,j[i]}$ - the interaction between housing wealth within a living area and the duration (in months) a borrower stays at his current living address. The main aim of the interacted variables is to explain the combined impact of the living area effects and individual-level characteristics on the probability of default.

The Own_i is a binary variable which takes a value of 1 when a borrower owns an accommodation and 0 otherwise. In the data sample the proportion of families who own a house is 53.9% (see the descriptive data table presented in section 2.2). The random-coefficient model of the variable Own_i is presented in [2.14]. It shows that the average impact of having a real estate property on the probability of default is γ_{own}. The $u_{j,own}$ is the microenvironment-level residual which explains the change in the probability over and above the population average value. The varying-coefficient model of the variable $Bank_i$ is similar. It includes the microenvironment-level residual $u_{j,bank}$ and the intercept γ_{bank}.

The variance-covariance matrix for the second-level random-effects is constrained to have an independent structure as illustrated in [2.15]. The reason for this specification is simple. I am primarily interested in estimating standard deviations of the microenvironment-specific effects and to a lesser extent in measuring the covariances between the varying-coefficients. Additionally, the independent structure of the variance-covariance matrix helps to speed up the estimation as the number of parameters is noticeably decreased. In this dissertation I do not provide a discussion about the alternative types of the variance-covariance matrix specification (such as exchangeable, identity or unstructured).

$$\begin{aligned} Pr\left(y_i = 1 \middle| x_i, u_j, z_j\right) \; = \; & Logit^{-1}\,(\,\alpha_{j[i]} + \gamma_1 Income_i + \gamma_2 Dependents_i \\ & + \gamma_3 Trade_{accounts_i} + \; {\beta_{j[i]}}^{bank} Bank_i + {\beta_{j[i]}}^{enq} Enquiries_i \\ & + \gamma_6 Professional_i + {\beta_{j[i]}}^{DR} DR_i + \gamma_8 R_{credits_i} + \gamma_9 Credit_i \\ & + \gamma_{10} Past_{due_i} + \; \beta_{j[i]}^{Own} Own_i + k'\delta), \end{aligned} \qquad [2.12]$$

$$\alpha_{j[i]} \; = \; \alpha_0 + \, z'\beta + \, u_{j,0},$$

$$z'\beta \; = \; \beta_1 Area\_Income_j + \beta_2\, AA_{residents_j} + \beta_3\, Stores_j + \beta_4 College_j,$$

$$\begin{aligned} k^{'}\delta \; = \; & \delta_1 Past_{due_i} \cdot AA_{residents_{j[i]}} + \, \delta_3 Adress_i \cdot Ownership_{Area,j[i]} \\ & + \, \delta_2 Burden_i \cdot Stores_j, \end{aligned}$$

$$\begin{aligned} {\beta_{j[i]}}^{Enq} &= \gamma_{enq} \; + \, u_{j,enq,} \\ {\beta_{j[i]}}^{DR} &= \gamma_{DR} \; + \, u_{j,DR,} \\ {\beta_{j[i]}}^{bank} &= \gamma_{bank} \; + \, u_{j,bank}, \\ \beta_{j[i]}^{Own} &= \gamma_{own} \; + \, u_{j,own}, \end{aligned} \qquad [2.14]$$

$$(u_{j,enq}\,, u_{j,DR}, u_{j,Bank},\, u_{j,Own}, u_{j,0} | x_i, z_j) \sim N \left(\begin{bmatrix} \alpha_0 + z'w \\ \gamma_{enq} \\ \gamma_{DR} \\ \gamma_{bank} \\ \gamma_{own} \end{bmatrix}, \Sigma_{\mathrm{u}} \right),$$

$$\Sigma_{\mathrm{u}} \; = \; \begin{bmatrix} \sigma_u^2 & & \cdots & & 0 \\ & \sigma_{enq}^2 & & & \\ \vdots & & \sigma_{DR}^2 & & \vdots \\ & & & \sigma_{bank}^2 & \\ 0 & & \cdots & & \sigma_{Own}^2 \end{bmatrix}. \qquad [2.15]$$

It is important to mention that the structure of the model in [2.12] is quite complex. It contains many random-effects, borrower-level and microenvironment-level variables and interacted variables. Accordingly, the maximum likelihood estimation of this scorecard is not an easy task. This is because in a multilevel logistic regression random-effects should be integrated out in a likelihood function which requires the application of numerical methods. An approximation of the likelihood produces decent results when the number of random-effects is low and the precision decreases as the number of random-effects increases. In this case it is better to apply Bayesian Markov chain Monte Carlo. This approach allows more flexibility in modelling random-effects in this credit scoring model. However, I do not apply a Bayesian MCMC to fit the credit scoring model in this subsection in order keep it comparable to the previous scorecards fitted in Stata by maximum likelihood.

The estimation results for the flexible version of the credit scorecard with multiple microenvironment-specific coefficients, group-level variables and interacted variables are provided in Table 2.14. The estimated standard deviations of the microenvironment-specific effects are presented together with their 95% confidence intervals.

The population average effects of the individual-level explanatory variables are very similar to the estimates from the previous credit scoring model. The standard deviation of the microenvironment-specific coefficient of credit enquiries equals 0.052 which is more than twice smaller than in the credit scorecard with only two varying-coefficients. A large variation is found between the coefficients of the variable Own_i. The standard deviations of the varying-coefficients of the number of derogatory reports DR_i and banking accounts $Bank_i$ are 0.175 and 0.48 on the logit scale.

The fitted model coefficients of the interacted variables are not precisely estimated which is not surprising, given I only have 61 level-two groups (microenvironments). Nevertheless, the impact of the interaction $Burden_i \cdot Stores_j$ on default is highly significant and positive. It shows that the effect of a higher credit burden differs across residence areas with different access to shopping facilities. Interestingly, this effect is more pronounced for over-indebted individuals who reside in microenvironments with a developed infrastructure of various department stores and shopping malls. The explanation is the following: in areas with highly developed infrastructure of shopping facilities, customers are offered a wider range of credit products because lenders locate more bank branches there in order to satisfy high demand for credit resources.

Variable	*Coefficient*	*Std.err.*	*z*	*P>|z|*
Total Income	-0.031	0.003	-9.92	<0.001
Number of dependents	0.133	0.024	5.64	<0.001
Trade accounts	-0.031	0.006	-5.16	<0.001
Bank accounts (ch/ savings)	-0.368	0.059	-6.28	<0.001
Enquiries	0.366	0.021	17.76	<0.001
Professional	-0.259	0.100	-2.60	0.009
Derogatory Reports	0.607	0.038	15.85	<0.001
Revolving credit balance	0.005	0.002	2.34	0.019
Previous credit	-0.170	0.069	-2.48	0.013
Past due	0.233	0.050	4.66	<0.001
Own	-0.260	0.112	-2.33	0.019
Constant	-1.890	0.286	-6.60	<0.001
Microenvironment-level model $\alpha_{j[i]}$				
Living area per capita income	-0.063	0.034	-1.86	0.062
Share of African-American residents	0.011	0.001	5.92	<0.001
Share of college graduates	-0.094	0.043	-2.15	0.031
Infrastructure of shopping facilities	0.012	0.005	2.12	0.034
Interactions				
$Past_{due_i} \cdot AA_{residents_{j[i]}}$	0.015	0.019		
$Burden_i \cdot Stores_j$	0.310	0.076		
$Adress_i \cdot Ownership_{Area,j[i]}$	-0.089	0.041		
Random-coefficients		*Estimate (Std.err.)*		*95% Confidence interval*
Std .deviation of $\beta_{j[i]}^{enq}$ (Credit enquiries)		0.052(0.016)		[0.028; 0.100]
Std .deviation of β_j^{DR} (Derogatory rep.)		0.175(0.085)		[0.068; 0.453]
Std .deviation of $\beta_{j[i]}^{bank}$ (Banking)		0.048(0.020)		[0.005; 0.164]
Std .deviation of $\beta_{j[i]}^{Own}$ (Own/rent)		0.664(0.097)		[0.501; 0.884]
Std .deviation of $\alpha_{j[i]}$		0.127(0.057)		[0.024; 0.269]

Table 2.14. *Estimation results for the flexible credit scoring model with multiple random-coefficients, microenvironment-level variables and interacted variables. The estimated standard deviations of the random-effects are reported together with their 95% confidence intervals.*

The results confirm that the impact of the interacted variable $Adress_i \cdot Ownership_{Area,j[i]}$, on probability of default is negative. This implies that in wealthy living areas with a high level of housing wealth the effect of the length of stay at the address on default is higher than in regions where the majority of families rent their accommodation.

I evaluate the discriminatory power of the flexible version of the two-level credit scorecard with microenvironment-specific coefficients, group-level variables and interactions by applying a ROC curve analysis as illustrated in Figure 2.6.

The optimal cut-off point is indicated by the red triangle on the ROC curve. The classification table given the optimal threshold $c_1 = 0.1496$, the summary results of the ROC curve analysis, the Gini coefficient and the accuracy ratio are presented in Table 2.15.

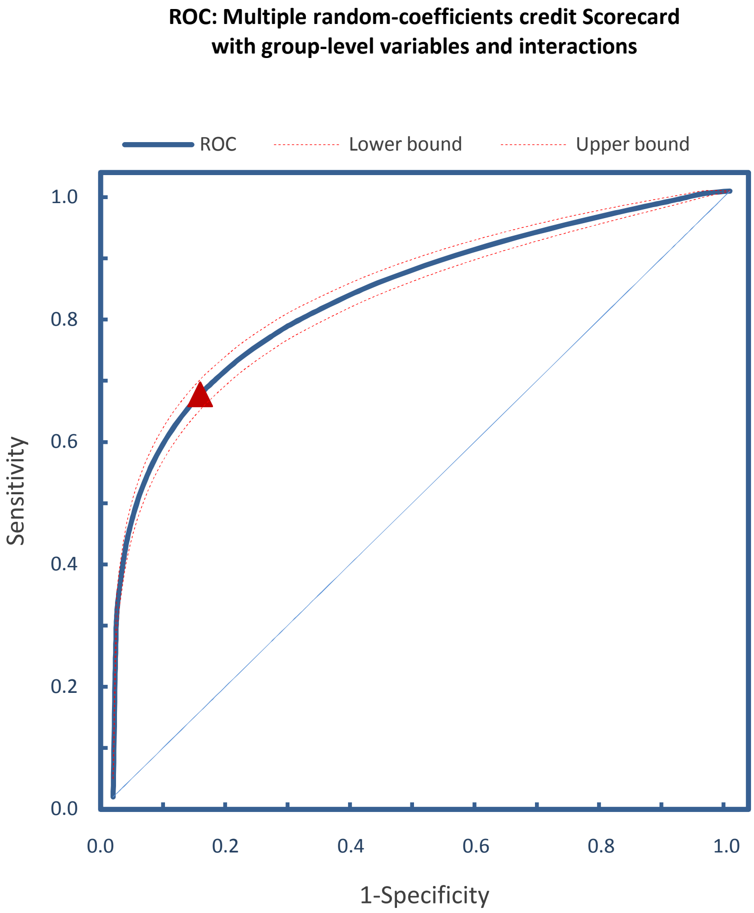


Figure 2.6. *ROC curve for the flexible credit scoring model with area-specific coefficients, group-level variables and interacted variables. The optimal cut-off point is* $c_1 = 0.1496$ *(threshold for the probability of default).*

The area under the ROC curve is increased to 0.825. The change in the estimated AUC value compared to the previous model is moderately small and the confidence intervals overlap. This is not surprising given the data limitations. The data sample is not large enough to provide the information required for a more precise estimation of a multilevel scorecard with many microenvironment-specific effects. Observing a larger sample on the credit histories of borrowers can improve the estimation and increase the predictive accuracy of a scorecard.

	True		
Classified ($c_1 = 0.1496$)	D	ND	Total
Default	273	623	896
Non-default	152	3132	3284
Total	425	3755	4180
Correctly classified, %			81.46
Sensitivity, %			64.12
Specificity, %			83.42
ROC curve metrics:			
Area under the ROC (AUC)			0.825
Standard error (DeLong)			0.005
95% confidence interval			[0.818; 0.831]
Gini coefficient			0.650
Accuracy ratio			0.715

Table 2.15. *Summary of the ROC analysis results, Gini coefficient, accuracy ratio and the classification table for the optimal cut-off point:* $c_1 = 0.1496$.

Given the optimal threshold c_1=0.1496 the credit scorecard correctly classifies 81% of applicants for a loan. I have to mention that this cut-off point implies that a lender equally weights true positive and true negative classifications which may not be the case in retail banking. I discuss alternative choices for an optimal threshold in the next chapter where I compare the predictive performance between different credit scoring models.

In summary, this chapter shows that specifying a multilevel hierarchical structure for the credit scoring model provides relevant information for a more accurate credit risk assessment of borrowers. It makes the scoring model more flexible and allows accounting for area-specific effects which are given by random-intercepts and random-coefficients. The microenvironment's random-effects are viewed as unobserved determinants of default which influence the riskiness of customers within a living area with a particular combination of economic and demographic conditions. Including these effects in the credit scoring model improves the predictive quality of a scorecard.

The other important advantage of a multilevel structure is that it allows exploring the impact of the microenvironment-level characteristics on the probability of default. The microenvironment-level information is given by the group-level variables and interactions. It is investigated that living area income and the share of college graduates have a negative impact on the probability of default. Controversially, a positive effect is found for the interaction of the credit burden of a borrower with good access to shopping facilities.

Scoring model	*Name*
Logistic credit scorecard in [2.1]	Scorecard 1
Microenvironment-intercept scorecard in [2.5]	Scorecard 2
Microenvironment-intercept scorecard with group-level variables in [2.7]	Scorecard 3
Microenvironment-coefficients scorecard with group-level variables in [2.9]	Scorecard 4
Multiple random-coefficients scorecard with group-level variables and interactions in [2.12]	Scorecard 5

Table 2.16. *Renamed credit scoring models.*

The next chapter provides the summary of the ROC curve analysis results and shows how to test the statistical significance of the differences between the AUC measures for the multilevel credit scoring models and the logistic regression scorecard. Additionally, I calculate and report several postestimation diagnostic statistics which aim to check the goodness-of-fit of the credit scoring models.

In this chapter I have presented the empirical results for five different credit scoring models which vary by the composition of the random-effects and group-level characteristics. In order to distinguish between different credit scoring models presented in this chapter I assign a name to each scoring model. It helps to simplify the presentation of the comparison results and shortens the notation. Table 2.16 lists the scorecards and assigns the new names to them. The postestimation statistical tests and the ROC curve in the next chapter are also going to be named according to the new names as given in the table.

3 Predictive accuracy and goodness-of-fit check

In this chapter I provide several postestimation diagnostic statistics which aim to check the predictive performance of the credit scoring models presented in the previous chapter.

In general, there are quite a few techniques discussed in the literature which can be used in order to check the goodness-of-fit and assess the discriminatory power of a regression. However, the number of possibilities decreases when a multilevel modelling is applied (Hox (2002)). The main complexity in a multilevel model which prevents application of the standard goodness-of-fit tests (Hosmer and Lemeshow, pseudo R^2) is that the model includes characteristics measured at different levels. Accordingly, in this thesis I compute and report the measures of the goodness-of-fit of an estimated scoring model which are appropriate for multilevel modelling. Following Farrell (2004) and Zucchini (2000) I calculate Akaike information criterion (AIC) as well as Bayesian information criterion (BIC). AIC and BIC are the tools for a model selection which combine both the measure of fit and complexity. Given two models are fitted on the same data, the model with the smaller value of the information criterion is considered to be better. If y is the data and K is the number of parameters θ in a model, then Akaike information and Bayesian information criteria can be defined as follows

$$AIC = -2 * \log[\, g(y|\theta)\,] + 2\,K,$$

$$BIC = -2 * \log[\, g(y|\theta)\,] + K \log(n),$$

where $g(y|\theta)$ is the likelihood and n is the number of observations. The mathematical details of the calculation of AIC and BIC are provided in Burnham and Anderson (2002), Akaike (1974) and Schwarz (1978).

I summarize the results of the ROC curves analysis for the multilevel credit scoring models and the logistic regression scorecard in section 3.1. This section provides a pairwise comparison of the AUC measures and tests the statistical significance of the differences in the AUC values between the different credit scorecards. Additionally, I briefly discuss the application of the ROC curve metrics for the evaluation of a scorecard performance in retail banking and describe alternative measures of the predictive accuracy check. In particular, I compute the area under a specific region of the ROC curve (a partial AUC) and show how to incorporate asymmetric costs in the regular ROC curve analysis.

Section 3.2 provides a comparison of a model fit by applying AIC and BIC criteria. It also checks the discriminatory power between credit scorecards by calculating Brier scores, logarithmic scores and spherical scores (Krämer and Güttler (2008)). These scalar measures of accuracy allow to compare the per observation error of the forecasts produced by the different scoring models. These techniques are relatively simple but at the same time they provide a transparent measure of the predictive quality.

I conclude the chapter by presenting a graphical illustration of the predicted probabilities and the fitted model results. Section 3.3 evaluates economic significance of the two-level structure and provides a discussion on the role of random-effects in a credit scoring model. In addition, I analyse the impact of the microenvironment-level characteristics on the riskiness of borrowers. It is explored that the quality of borrowers varies between living areas with dissimilar economic and socio-demographic conditions. Poor living areas contain a higher share of borrowers with a derogatory credit history and problematic debt than richer regions. Living area conditions matter for more accurate credit risk assessment.

3.1 Summary of ROC curve analysis

In order to compare the ROC curves and related metrics between the multilevel credit scoring models and the logistic regression scorecard I provide a summary plot in Figure 3.1. The plot combines five ROC curves for the credit scoring models which are presented in chapter 2. The curves are named according to the shortened notation as given in Table 2.16. The logistic regression scorecard is presented by the dashed line and it is given the nameROC^1. The ROC^2 and ROC^3 denote microenvironment-specific intercept scorecards with and without group-level variables. The curves ROC^4 and ROC^5 illustrate the performance of the credit scoring models with two random-coefficients and multiple random-coefficients.

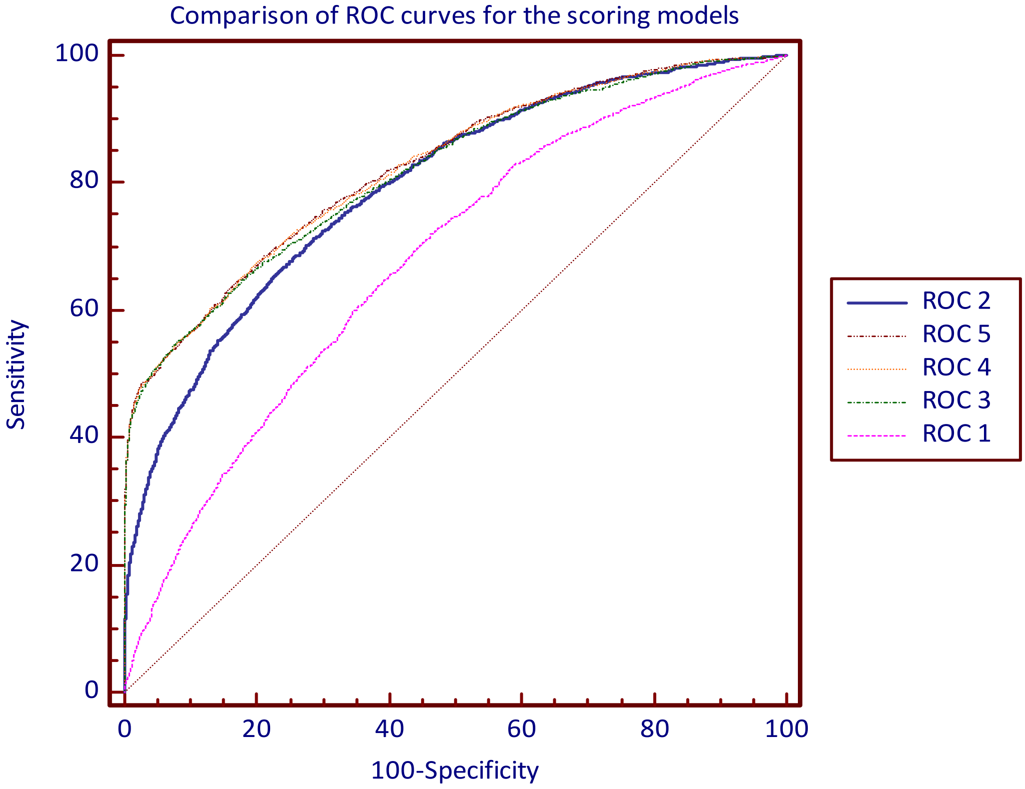


Figure 3.1. *The comparison of the ROC curves for the different credit scoring models presented in chapter 2.*

It is evident from the graph that the multilevel credit scoring models outperform the conventional logistic scorecard by showing a higher classification performance. Similarly, the comparison of the ROC curves between the multilevel models reveals that the scorecards with more microenvironment-specific effects provide higher predictive accuracy. The two-level scorecard with multiple random-coefficients and group-level variables has a ROC curve which lies above the other curves.

In order to give a meaningful interpretation to the graphical illustration of the ROC curves I make a pairwise comparison of the areas under the curves. The results are presented in Table 3.1. I use the logistic scorecard as a reference model and calculate the differences in the AUC measures as following: $\Delta AUC_i = AUC_{Logit} - AUC_{ROC_i}$, where AUC_{ROC_i} denotes the area under the ROC^l for *l=2,..,5*. The standard error of this difference is given by

$$SE_{AUC} = \sqrt{\left(SE_{AUC_1}\right)^2 + \left(SE_{AUC_2}\right)^2 - 2\rho SE_{AUC_1} SE_{AUC_2}},$$

as reported in the third column in the table (SE_{AUC} and ρ are estimated according to Delong (1988)).

ROC curve	$\Delta AUC = AUC_{ROC_i} - AUC_{Logit^*}$	*Standard error*	*95% confidence interval*	*z-statistics*	*p-value*
ROC^2	0.094	0.00566	[0.084;0.105]	16.65	<0.001
ROC^3	0.111	0.00623	[0.099;0.123]	17.81	<0.001
ROC^4	0.117	0.00615	[0.105;0.128]	18.98	<0.001
ROC^5	0.118	0.00623	[0.107;0.130]	19.02	<0.001

** Logistic regression scorecard: area under the ROC^{Logit} curve is AUC_{Logit}=0.707*

Table 3.1. *A pairwise comparison of the differences between the areas under the ROC^i and the ROC^{Logit}. The standard errors of ΔAUC are calculated according to Delong (1988).*

Following Hanley and McNeil (1984) I calculate the z-statistics in order to test if the differences (ΔAUC_i) are statistically significant. The z-statistics tests the null hypothesis that the difference between two AUC values is zero. The test results and the corresponding p-values are presented in the fifth and the sixth columns in the table. I also report the 95% confidence interval for the differences in the areas as reported in the forth column in the table. It is evident, that AUC values are significantly smaller for the logistic regression scorecard as compared t the multilevel scorecards 2-5. Between the multilevel scoring models the AUC values increase with the complexity of the models, although the differences in ΔAUC are significant only in some cases.

Next, I will take a closer look at a ROC curve analysis application to retail banking in general and discuss several alternative methods which help to assess the predictive accuracy of a scoring model. Under particular circumstances these alternative methods are more relevant and suitable than a standard ROC curve metrics.

In general, a ROC curve illustrates the performance of a model by plotting true positive rate against false positive rate. It is currently considered to be a benchmark method used to check the predictive quality of a model. Given a ROC curve, the predictive performance of a model is measured by computing the area under the curve. However, there are several limitations associated with the use of AUC as a standard measure of accuracy (Termansen et al. (2006), Austin (2007), Hosmer and Lemeshow (2000)). In this dissertation I only briefly discuss the main drawbacks of the AUC measure when it is applied in credit scoring.

First, ROC (AUC) ignores the predicted probability values and goodness-of fit of the estimated model (Ferri (2005)). The continuous forecasts of the probabilities are converted to a binary default-nondefault variable. This transformation neglects the information on how large is the difference between the threshold and the prediction. Hosmer and Lemesow (2000) show that it is possible for a poorly fitted model (which overestimates or underestimates all the predictions) to have a good discrimination power. They also provide an example where a well-fitted model has a low discrimination power.

A second limitation of the ROC curve and AUC is that they summarise a model performance over all regions of the ROC space including regions in which it is not reasonable to operate (Baker and Pinsky (2001)). For instance, in retail banking, a lender typically defines a threshold for the accept/reject decision within a range (0.1; 0.3). Therefore, he is rarely interested in summarizing the scorecard's performance across all possible thresholds as given by a ROC curve (AUC) and related metrics. In this case the left and central areas of the ROC curve are of less importance for him.

One solution to this weakness would be to compute an area under a portion of the ROC curve. Partial AUC is an alternative to the regular AUC measure which evaluates the discriminatory power of a model over a particular region of the ROC curve (Thompson and Zucchini (1989), Baker and Pinsky

(2001) and McClish (1989)). When it is applied in credit scoring, the partial AUC is simply the area under the partial ROC curve between two cut-off points or given a specific range for the specificity/sensitivity pairs. Computing a partial AUC is also helpful if a lender aims to satisfy a budget constrain or fulfil a banking legislation requirement. For instance, a partial AUC can be estimated over the region of the ROC curve between two cut-off points which yields the desired range of true positive rates.

I do not provide a detailed discussion on the calculation of a partial AUC in the thesis as the decision about an assessment of a particular region of a ROC curve should be guided by practical considerations within a commercial bank. Here, I only consider the case when a lender decides to evaluate a scorecard performance over the region of the ROC curve between two cut-off points.

On a ROC curve plot the performance of a predictive model is visualized by plotting TPR (true positive rate) versus FPR (false positive rate) over all possible cut-off points c. If the TPR given a threshold c is $TPR(c) = \Pr(Y > c|D) = S_D(c)$ and the corresponding $FPR\ is\ FPR(c) = Pr(Y > c|ND) = S_{ND}(c) = t$ then according to Pepe (2003) the area under the ROC curve from some point t_1 to the point t_2 is defined as following

$$\begin{aligned} pAUC &= \int_{t_1}^{t_2} ROC(t)dt \\ &= \int_{t_1}^{t_2} S_D\big(S_{ND}^{-1}(t)\big)dt \\ &= P\,r\{Y^D > Y^{ND}, Y^{ND} \in\ [S_{ND}^{-1}(t_1), S_{ND}^{-1}(t_0)]\ \}, \end{aligned}$$

where Y^{ND} and Y^D are continuous variables with survivor functions S_{ND} and S_D. In application to credit scoring Y^{ND} and Y^D would define the classification scores (or probabilities) assigned to the non-defaulted and defaulted customers. Figure 3.2 provides a graphical illustration of the partial area under the ROC curve between *FPR(c₂)* and *FPR(c₁)* where *c₁* and *c₂* are the cut-off points.

On the graph the partial area of the ROC curve is bounded above by the area of the rectangle that encloses it. This rectangle has sides of length 1 and $(FPR(c1) - FPR(c2))$ which leads to the following partial area

$$pAUC^{max} = FPR(c1) - FPR(c2).$$

This area is the maximum partial AUC given c_1 and c_2. The lower bound for the partial AUC is given by the trapezoid which lies below the 45° diagonal line on the ROC plot. The area of this trapezoid is

$$pAUC^{min} = \frac{((FPR(c_1)+FPR(c_2))}{2}(FPR(c_1) - FPR(c_2)),$$

$$pAUC^{max} > pAUC > pAUC^{min}.$$

Accordingly, the partial AUC given two cut-off points c_1 and c_2 lies between the maximum and minimum partial areas. In other words, $pAUC^{max}$ gives the area under the portion of a ROC curve of a perfect scoring model (100% sensitivity). Similarly, $pAUC^{min}$ provides a partial AUC of a random guessing.

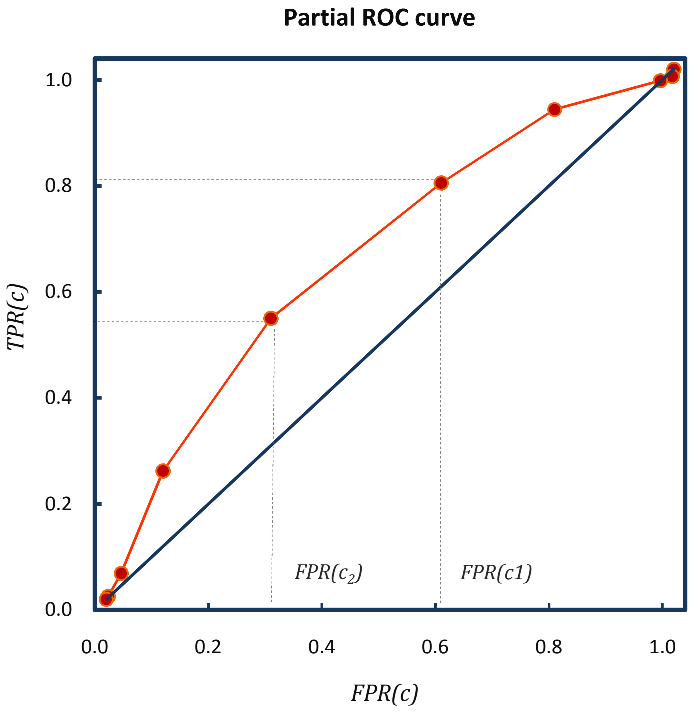


Figure 3.2. *Partial area under the ROC curve between FPR(c2) and FPR(c1).*

Next, I apply a partial AUC to the five scorecards in order to evaluate the areas under the portion of the ROC curve between the cut-off points c_1=0.1 and c_2 =0.3 and between c_1=0.1 and c_2 =0.2. To calculate a partial AUC I need to compute the sensitivity/specificity pairs corresponding to the cut-off points within the range [0.1, 0.3]. Table 3.2 presents the results. The sensitivity defines the true positive rate (TPR) and specificity gives the true negative rate (TNR).

The results in the table are interesting by themselves and show how the discriminatory power of a scorecard changes if the threshold for an accept/reject decision increases from 0.1 to 0.3. Given the cut-off point c_1 =0.1 the logistic scorecard correctly classifies 61.14% of true defaulters which is 10-17% smaller than the TPR predicted by the multilevel scoring models. If the cut-off point increases to c_2=0.3 the differences in the classification performance become even more sharp between the logit scorecard and the multilevel models. Given the threshold c_2=0.3 the logistic scoring model accurately forecasts only 14.56% of the true defaulters while scorecard 2 correctly classifies 42.34% of the true positive outcomes. The TP rates at the cut-off point 0.3 produced by the scorecards 4 and 5 are even higher. The table implies that the multilevel scoring models show better classification performance over the region of the ROC curve between the cut-off point c_1 and c_2.

Cut-offs	*Scorecard 1*		*Scorecard 2*		*Scorecard 3*		*Scorecard 4*		*Scorecard 5*	
	TPR	*TNR*	*TPR*	*TNR*	*TPR*	*TNR*	*TPR*	*TNR*	*TPR*	*TNR*
0.10	61.14	68.45	78.88	61.74	71.25	73.30	72.91	73.18	73.22	73.07
0.12	53.61	75.44	73.69	68.31	67.99	77.99	68.88	77.56	69.32	77.60
0.14	45.68	81.18	68.21	74.34	64.62	82.27	65.62	81.63	65.12	81.79
0.16	39.26	85.04	63.02	79.24	60.97	85.32	61.86	84.67	62.41	84.82
0.18	33.53	88.29	59.26	82.32	58.82	87.71	59.65	87.16	59.54	87.13
0.20	29.12	90.36	55.17	85.76	56.99	89.59	57.27	89.14	56.99	89.26
0.22	24.70	92.10	51.85	87.55	55.50	91.28	55.89	90.84	55.78	90.97
0.24	21.18	93.16	48.42	89.15	54.06	92.66	54.51	92.20	54.12	92.14
0.26	18.37	94.22	46.10	90.72	53.01	93.61	52.85	93.44	52.85	93.40
0.28	16.16	95.02	44.06	91.83	51.80	94.40	51.85	94.34	51.69	94.43
0.30	14.56	95.83	42.34	92.70	50.75	95.20	50.97	95.18	50.58	95.02

Table 3.2. *Sensitivity/specificity pairs corresponding to the cut-off points for probability of default within the interval* [0.1; 0.3].

The highest TN rate at the threshold c_1=0.1 is provided by scorecard 3. However, given the threshold c_2=0.3 the logistic regression scorecard slightly outperforms other scoring models and correctly classifies 95.83% of the true non-defaulters. The TNR provided by the scorecards 2-5 are only slightly smaller.

The partial areas under the ROC curve are presented in Table 3.3. I calculate and report partial areas for the two regions of the ROC space: between cut-off point $c_1 = 0.1$ and $c_2 = 0.3$ and between $c_1 = 0.1$ and $c_2 = 0.2$. Additionally to the pAUC values, the table provides the maximum and minimum bounds for the partial areas and the relative value of the partial AUC ($\frac{pAUC}{pAUC^{max}}$).

Cut-off points (interval)	*[0.1; 0.3]*				*[0.1; 0.2]*			
	$pAUC$	$pAUC^{max}$	$pAUC^{mi}$	$\frac{pAUC}{pAUC^{ma}}$	$pAUC$	$pAUC^{max}$	$pAUC^{min}$	$\frac{pAUC}{pAUC^{max}}$
Scorecard 1	0.1036	0.2738	0.0489	***0.394***	0.0988	0.2191	0.0451	***0.451***
Scorecard 2	0.1876	0.3096	0.0705	***0.631***	0.1609	0.2402	0.0630	***0.670***
Scorecard 3	0.1335	0.2190	0.0344	***0.635***	0.1044	0.1629	0.0302	***0.641***
Scorecard 4	0.1362	0.2200	0.0348	***0.645***	0.1038	0.1596	0.0300	***0.651***
Scorecard 5	0.1358	0.2195	0.0350	***0.645***	0.1054	0.1619	0.0304	***0.651***
Differences between the relative partial AUC values								
Scorecard 1 2				0.237				0.219
Scorecard 1 3				0.241				0.190
Scorecard 1 4				0.251				0.200
Scorecard 1 5				0.251				0.200

Table 3.3. *Partial areas under the portion of the ROC curve between the cut-off points c_1=0.1 and c_2= 0.3 and between c_1=0.1 and c_2= 0.2. The differences in the relative partial AUC values for the logit scorecard and the multilevel scoring models.*

Table 3.3 confirms that the multilevel scoring models outperform the logistic regression scorecard over the region of the ROC space between the cut-off points c_1 and c_2 . It is also true that the differences in the partial AUC values are higher than the differences in the total AUC given in Table 3.1. Given the thresholds c_1 and c_2 the scorecards 4 and 5 provide similar classification performance. Interestingly, given the region of the ROC space between the cut-off point $c_1 = 0.1$ and $c_2 = 0.2$ scorecard 2 shows the highest predictive accuracy yielding the relative partial area $\frac{pAUC}{pAUC^{max}}$=0.67.

The third important limitation of a standard ROC curve or the AUC value is that they do not account for the asymmetry of costs. The AUC implies that misclassifying a defaulter has the same consequence as incorrectly classifying a non-defaulter. However, this is not the case in retail banking where the costs of misclassification errors (false positive and false negative outcomes) are very asymmetric.

Generally, incorrectly classifying a true defaulter leads to problematic credit debt. Management of delinquent credit accounts is very costly for a lender. When a scoring model incorrectly classifies a true defaulter/non-defaulter, the costs associated with a past due credit account are much higher than the opportunity costs of a foregone profit. This implies that in retail banking a lender is primarily interested in increasing the true positive rate in order to minimize the misclassification costs of the incorrectly predicted non-defaulters.

There are several techniques proposed in the literature which aim to incorporate misclassification costs in the assessment of the predictive accuracy. Metz (1978) proposed to measure the expected losses (costs) by summing up the probability weighted misclassification costs and benefits of the correct and false predictions. Given the probability of default $p(D)$ and the probability of non-default $p(ND)$ the expected losses can be calculated using the following formula

$$\begin{aligned}
Expected\ Loss = \ & C(D|D) \cdot p(D) \cdot TPR + C(ND|ND) \cdot p(ND) \cdot TNR + \\
& C(D|ND) \cdot p(ND) \cdot FPR + C(ND|D) \cdot p(D) \cdot (1 - TPR) \\
= \ & TPR \cdot p(D) \cdot \big(C(D|D) - C(ND|D)\big) + C(ND|ND) \cdot p(ND) + \\
& FPR \cdot p(ND) \cdot \big(C(D|ND) - C(ND|ND)\big) + C(ND|D) \cdot p(D),
\end{aligned}$$

where $C(ND|D)$ is the cost of a false negative classification, $C(D|ND)$ is the cost of a false positive classification. The cost of the correct classification of the true defaulter is $C(D|D)$ and the non-defaulter is $C(ND|ND)$, correspondingly.

Next, I apply the expected loss approach to compare the misclassification costs between different credit scoring models. For simplicity, I assume that the cost of the correct classification of a true positive (negative) outcome is zero. The cost of an incorrectly classified defaulter is assumed to be 10 times higher than the cost of a misclassified non-defaulter ($C(ND|D) = 100,\ C(D|ND) = 10$). Table 3.4 reports the expected losses a scorecard produces given three cut-off points for the accept/reject decision c_1=0.1, c_2=0.2 and c_3=0.3.

Cut-off points:	$c_1 = 0.1$	$c_2 = 0.2$	$c_3 = 0.3$
Scorecard 1	7.97	10.40	11.89
Scorecard 2	6.16	7.28	8.41
Scorecard 3	6.19	6.70	7.06
Scorecard 4	5.97	6.70	7.03
Scorecard 5	5.94	6.73	7.09

Table 3.4. *Misclassification costs produced by a credit scoring model given three different cut-off points for the accept/reject decision.*

The results in the table suggest that the multilevel scorecards outperform logistic scoring model by providing smaller misclassification costs.

Concluding the discussion about the application of a ROC curve and metrics derived from it in retail banking, I suggest that additionally to the ROC analysis it is important to compute and report alternative measures of accuracy and predictive performance. In particular, the partial area under the curve, misclassification rates and expected losses given a threshold are good complements to the regular ROC curve analysis. In addition, it is also important to report goodness-of-fit measures together with a ROC (AUC) curve metrics in order to avoid situations where a poorly fitted model shows a high discriminatory power because it overestimates all positive instances and produces a very high TPR (sensitivity is close to 100%).

3.2 Measures of fit and accuracy scores

This section assesses and compares the goodness-of-fit between the multilevel credit scoring models and the logistic regression scorecard. I compute and report several measures of the fit of an estimated statistical model which are commonly applied in econometrics. Following Akaike (1974) and Schwartz (1978) I calculate and report Akaike Information criterion (AIC) and Schwarz criterion or Bayesian Information criterion (BIC). AIC and BIC criteria are deviance-based measures of fit of an estimated model. Generally, these criteria are applied to select the model which provides the best fit among the range of the fitted models while keeping the model parsimonious at the same time.

Table 3.5 reports the AIC and BIC criteria for the multilevel credit scoring models and the logistic regression scorecard. The model with the smallest values of both AIC and BIC criteria provides the best fit.

Postestimation diagnostics	*AIC*	*BIC*
Scorecard 1	2991.3	3090.2
Scorecard 2	2957.1	3062.6
Scorecard 3	2927.1	3045.7
Scorecard 4	2909.2	3041.0
Scorecard 5	2884.5	3029.4

Table 3.5. *Postestimation diagnostic statistics: Akaike information criterion (AIC) and Bayesian information criterion (BIC).*

According to the information criteria the multilevel scorecards (scorecard 2-5) outperform the conventional logit scorecard. It is also true that among the multilevel models AIC and BIC values decrease with the degree of the model's complexity. Credit scorecards which include more microenvironment-specific effects and group-level characteristics show a superior classification performance.

A flexible version of a scoring model with multiple random-coefficients, microenvironment-level variables and interactions (scorecard 5) is preferred by the information criteria.

Next, I compute several scalar measures which aim to assess the predictive accuracy of the probability forecasts. Following Krämer and Güttler (2008) I use the predicted probabilities for the set of credit scoring models and apply a Brier score as well as logarithmic and spherical scores to check the accuracy of the forecasts.

The Brier score is the mean squared difference between the predicted probabilities and the observed binary outcomes (Brier (1950), Murphy (1973), Jolliffe and Stephenson (2003)). It is one of the oldest and most commonly used techniques for assessing the quality of the probability forecasts of a binary event (default/non-default).

The formula for the calculation of a Brier score is given in [3.1]. It calibrates the average squared deviation of the predicted probabilities $\widehat{p_i}$ from the actually observed outcomes θ_i. Lower values for the score indicate higher accuracy. The estimated Brier scores for the credit scorecards are reported in the second column in Table 3.6.

$$Brier\ Score = \frac{1}{N}\sum_1^N (\theta_i - \widehat{p_i}\,)^2\ ,\ \ where\ \ \ \theta_i = \begin{Bmatrix} 1,\ default \\ 0\ ,\ non-default \end{Bmatrix}. \qquad [3.1]$$

The logarithmic score is another measure of the forecasting accuracy of a model. The calculation of the score is shown in [3.2]. The logarithmic score values are always negative. For $\theta_i = 1$, $\ln(|\widehat{p_i} + \theta_i - 1|)$ is close to zero when $\widehat{p_i}$ approaches one; for $\theta_i = 0$ it is close to zero when $\widehat{p_i}$ is small. Accordingly, the scoring rule imposes that a model with the closest to zero logarithmic score shows the best performance. The third column in Table 3.6 presents the values of the logarithmic scores for the credit scoring models.

$$Logarithmic\ score = \frac{1}{N}\sum_{i=1}^N \ln(|\widehat{p_i} + \theta_i - 1|). \qquad [3.2]$$

A slightly modified version of the logarithmic score is a spherical score which was introduced by Roby (1965). The calculation of the score is shown in [3.3]. The logarithmic score approaches unity when the predicted probabilities are close to the observed outcomes. The values of the spherical scores for the credit scoring models are provided in the last column in Table 3.6.

$$Spherical\ score = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|\widehat{p_i}+\theta_i-1|}{\sqrt{\widehat{p}_i^2+(1-\widehat{p_i})^2}}\right). \qquad [3.3]$$

Predictive accuracy scores:	*Brier score*	*Logarithmic score*	*Spherical score*
Scorecard 1	0.08090	-0.301	0.910
Scorecard 2	0.06736	-0.235	0.926
Scorecard 3	0.06252	-0.208	0.932
Scorecard 4	0.05663	-0.187	0.938
Scorecard 5	0.05652	-0.186	0.939

Table 3.6. *Score measures of predictive accuracy for the logistic regression and the multilevel credit scoring models: the Brier scores, logarithmic scores and spherical scores.*

The results of the Brier scores confirm that the logistic scoring model produces the crudest forecasts yielding the highest per observation error. It is also true, that among the multilevel scorecards (scorecard 2-5), models with more microenvironment-specific effects provide a better calibration of the probabilities of default. The smallest error of the forecasts (0.05652) is produced by the flexible version of a credit scoring model (scorecard 5) which includes multiple area-specific coefficients, group-level variables and interactions.

Similar conclusion is made after comparing the logarithmic and spherical scores. The spherical scores are reported in the last column in the table. The best results of the logarithmic and spherical scores are given by the scorecard 5. It is also true that the score values increase with the degree of the model complexity.

To summarize the results of the predictive accuracy measures and the goodness-of-fit check, I conclude that the multilevel credit scoring models outperform the logistic regression scorecard. It is evident that the results of different postestimation diagnostics provide the same ranking to the credit scoring models discussed in the previous chapter. This confirms the main contribution of this thesis is to introduce a multilevel scorecard which improves the forecasting quality of a scoring model.

Multilevel credit scoring is more efficient because it allows specifying a two-level structure where borrowers are nested within microenvironments and modelling random-effects. Microenvironment-specific effects vary across groups and show the impact of the economic and demographic conditions in the living areas on the riskiness of borrowers. These area-specific effects are viewed as unobserved determinants of default. Accordingly, including them in the scoring model improves the predictive quality and provides better fit to the data.

Importantly, microenvironment-specific effects capture the information on unobserved determinants of credit worthiness of individuals which impact the probability of default additionally to the observed characteristics measured at the borrower-level or group-level. This implies that two identical borrowers with the same personal characteristics but different living area conditions (microenvironments) are going to have different forecasts of probabilities because they are exposed to different area-specific hazards.

In the next section I apply a graphical illustration of the fitted model results in order to analyse the quality of borrowers and microenvironment-specific effects in the living areas with different economic and demographic conditions.

3.3 Predictive quality comparison: bivariate probit versus multilevel scorecard

In this subsection I compare predictive quality of the multilevel scorecard to a credit scoring regression analyzed by W. Greene (1992). Both credit scorecards are fitted using the same data on credit histories of borrowers.

In the paper W. Greene introduces a credit scorecard which takes into account the problem of reject inference. He applies a sample selection bivariate probit to model the probability of default on a loan. In this specification probability of default $Pr(D = 1|x, C = 1)$ is conditional on the application status, where $C = 1$ means a borrower is granted a loan (accepted) and $C = 0$ means that a customer is rejected. Accordingly, the main aim of the paper is to show that unconditional scoring model will give a downward biased estimate of default probability for an individual selected at random from the population because a part of the applicants (below the defined threshold) are not accepted by a lender. This implies that the probability default model should condition specifically on the application status in order to be applicable to the population at large.

I use the multilevel scorecard with microenvironment-specific intercept from chapter 2 (as given in [2.5]) to compare the predictive accuracy. In the paper W.Greene does not assess classification quality of the scorecard by ROC curve and related metrics such as AUC, Gini coefficient, accuracy ratio and logarithmic score. Therefore, in order to calculate these accuracy measures I replicate the estimation of the probit credit scoring model following the steps described in the paper. I take the same set of explanatory variables and sampling weights to predict the probabilities. Then, I apply these predictions to compute different accuracy measures and perform a ROC curve analysis. The detailed description of the estimation procedure is given in Appendix I. In addition, Appendix I reports the coefficient estimates for the bivariate probit regression which models probability of default conditional on the acceptance status (whether application for a loan is accepted or rejected by a lender).

I start by comparing classification rates, sensitivity and specificity values for the probit scorecard and the multilevel scorecard which are computed given two cut-off points. In the paper W.Greene reports a classification table for the cut-off points $c_1 = 0.094$ and $c_2 = 0.12$. Therefore, in order to make classification rates comparable I accomplish similar calculations for the multilevel scorecard. Table 3.7 provides accuracy measures which are FPR, FNR, sensitivity, specificity and the correct classification rate.

Classification performance	$\boldsymbol{c_1 = 0.094}$		$\boldsymbol{c_2 = 0.12}$	
	Probit scorecard	*Multilevel scorecard*	*Probit scorecard*	*Multilevel scorecard*
Correctly classified, %	57.21	63.28	67.92	69.27
False D rate for true ND (FPR), %	45.02	39.08	31.98	31.36
False ND rate for true D (FNR), %	21.49	20.51	33.03	26.37
Specificity, %	54.98	60.92	68.02	68.64
Sensitivity, %	78.51	79.49	66.97	73.63

Table 3.7. *Predictive accuracy rates for the multilevel scorecard and the probit scoring model given two cut-off points, c_1 and c_2. D-defaulters, ND-non-defaulters. FPR - false positive rate, FNR – false negative rate.*

The results confirm that the multilevel credit scoring model outperforms the bivariate probit scorecard in both cases: given thresholds c_1 and c_2. Given $c_1 = 0.094$ multilevel scorecard correctly predicts 63.28% of the outcomes while the probit scorecard does only 57.21%. False negative rates are higher for the probit scoring model. Sensitivity and specificity rates are higher for the multilevel scorecard.

Next, I compare classification performance of the scorecards by applying a ROC curve analysis. Figure 3.3 shows the ROC curve and its 95% confidence interval for the bivariate probit scorecard. In the table below the graph I report the AUC, the Gini coefficient and the accuracy ratio. In addition, Table 3.8 compares the AUC values between the scoring models. I follow Hanley and McNeil (1984) and calculate the z-statistics in order to test if the difference in the areas ($\Delta AUC = AUC_{Multilevel} - AUC_{Probit}$) is statistically significant. The z-statistics tests the null hypothesis that the difference between two AUC values is zero.

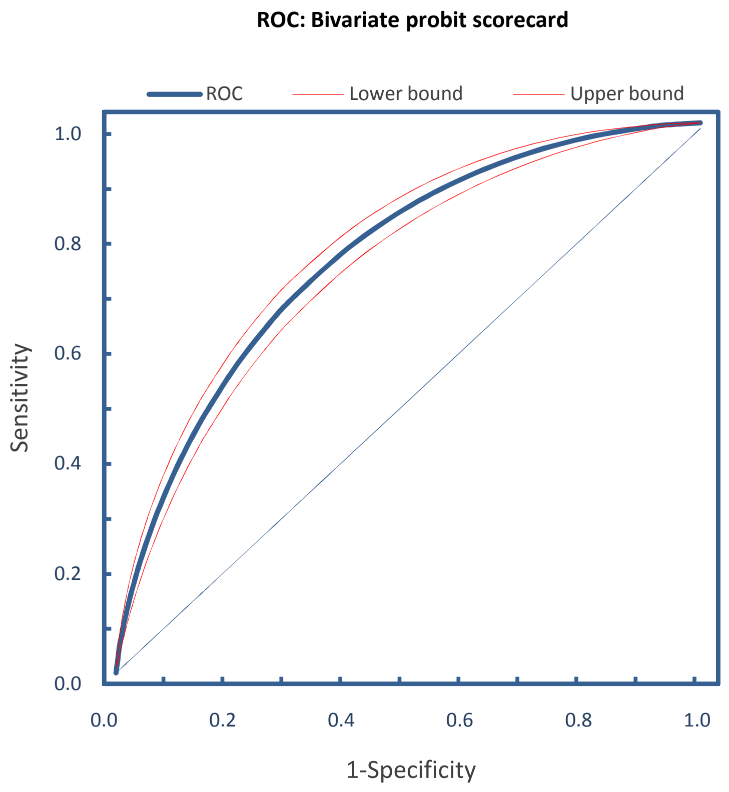


Figure 3.3. *ROC curve for the bivariate probit credit scoring model.*

Predictive accuracy measures	
Area under the ROC$_{\text{Probit}}$ (AUC_{Probit})	0.761
Standard error (DeLong)	0.010
95% confidence interval	[0.74 ; 0.78]
Gini coefficient	0.474
Accuracy ratio	0.522
Difference in AUC values:	
$\Delta AUC = AUC_{Multilevel} - AUC_{Probit}$	0.039
Z-statistics	-3.666
p-value	<0.001

Table 3.8. *Summary of the ROC curve metrics, the Gini coefficient, the accuracy ratio and the difference in the AUC values.*

The area under the ROC curve for the probit scorecard is 0.761 with standard error 0.01 (DeLong, 1988). The 95% confidence interval for the AUC_{Probit} does not overlap with the confidence interval for $AUC_{Multilevel}$. The difference in the AUC values between the multilevel and probit scorecards is statistically significant with a very small p-value. This confirms that the multilevel scoring model shows higher classification performance compared to the bivariate probit model.

In addition, I check accuracy of the forecasts by computing several scalar measures of classification quality. These measures are Brier score, logarithmic and spherical scores. Table 3.9 reports accuracy scores for the probit scoring model and the multilevel scorecard.

Accuracy scores:	*Brier score*	*Spherical score*	*Logarithmic score*
Probit Scorecard	0.0764	0.913	-0.269
Multilevel scorecard	0.0674	0.926	-0.235

Table 3.9. *Accuracy scores: comparison between the probit credit scoring model and the multilevel scorecard.*

The average error of the forecasts or Brier score is higher for the probit scoring model. This implies that the probit scorecard provides a lower classification quality compared to the multilevel scorecard. Similar conclusions can be made after comparing logarithmic and spherical scores.

In summary, it is evident that the credit scoring model with a multilevel structure outperforms the bivariate probit scorecard. The ROC curve metrics and the classification quality measures show higher predictive accuracy for the multilevel scoring model.

3.4 Graphical illustration of the fitted model results

3.4.1 Microenvironment-specific coefficients

This subsection aims to visualize the fitted model results. The credit scoring models introduced in chapter 2 include many microenvironment-specific effects at the second-level of the model hierarchy. The area-specific effects are captured by random-intercepts and random-coefficients in the scorecards. In order to emphasize the role of the microenvironment-specific effects I provide a graphical illustration of the fitted model varying-coefficients. In addition, I discuss and visualize the differences between area-specific effects within poor and rich areas.

Consider the credit scoring model 4 with two random-coefficients which is specified in [2.9]. Figure 3.3 illustrates the microenvironment-specific residuals $\hat{u}_{Enq,j}$ of the borrower-level variable $Enquiries_i$ (number of credit enquiries). I choose this variable for the graphical representation because the number of credit enquiries is a very powerful predictor which contains valuable information on the previous applications for a loan. The varying-coefficient of $Enquiries_i$ implies that the effect of credit enquiries differs across living areas with dissimilar economic and demographic conditions.

Figure 3.4 visualizes the microenvironment-specific effects of the variable $Enquiries_i$. In the second-level model for the coefficient $\beta_{j[i]}^{enq} = \gamma_{enq} + u_{j,enq}$, the residual $u_{Enq,j}$ explains the change in the probability over and above the population average value. The predicted area-specific effects $\hat{u}_{Enq,j}$ are illustrated by the blue points on the plot and the population average effect of enquiries is given by the straight red line. The line is parallel to the abscissa axis which implies the impact of enquiries on default is constant across borrowers. Including area-specific effects $u_{Enq,j}$ in the model for the varying-coefficient brings more flexibility in modeling. The microenvironment-specific residual reflects the economic and socio-demographic conditions in the residence area and explains the unobserved characteristics which impact the riskiness of a borrower who resides within a microenvironment j.

The abscissa axis on the graph shows the microenvironment ID. The highest values of the second-level residuals $\hat{u}_{Enq,j}$ are marked by the red triangles on the plot. These residuals correspond to low income areas with a high share of African-American residents and a low level of per capita real estate wealth.

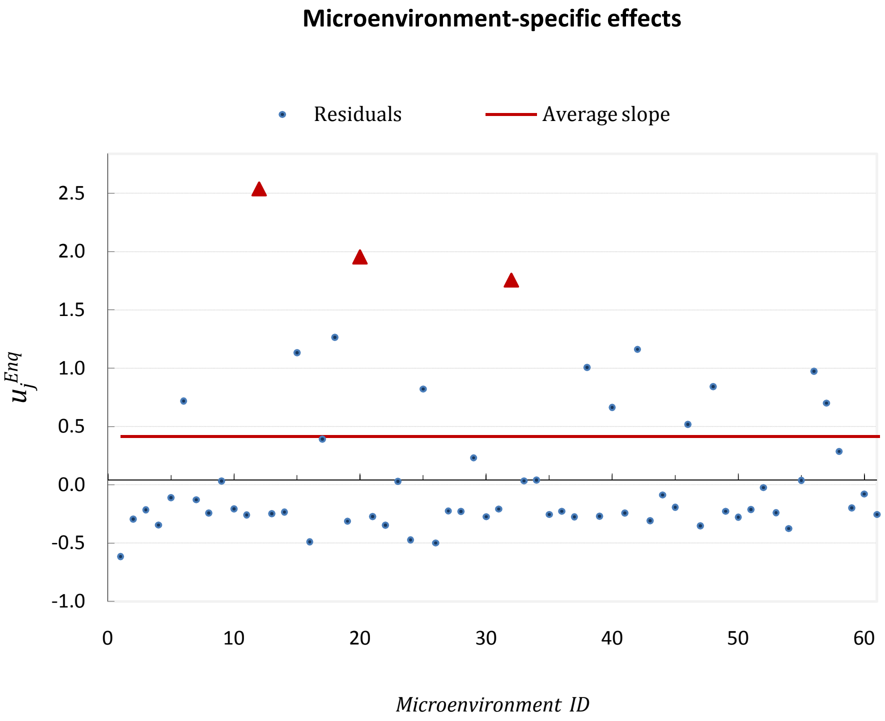


Figure 3.4. *Second-level residuals of $Enquiries_i$. Population average effect of enquiries is illustrated by the straight dotted line.*

If the fixed-effect coefficient is assigned to the variable $Enquiries_i$ then the impact of a unit change in the number of credit enquires is constant for all borrowers and predicts the change in the probability by $\pm 9.25\%$. This assumption may not hold given that nowadays retail bankers offer different credit opportunities under various conditions within different living areas. After monitoring and analysing the quality of borrowers a lender decides which kind of credit products to offer. Given a residence area of borrowers retail bankers may offer credit products with only fixed / flexible interest rates and with / without a revolving credit line.

The living conditions in a microenvironment may also determine the quality of the customers. Richer living areas contain more individuals with a good credit history and poor districts have a higher share of borrowers with a bad credit history. A customer has a good credit history if he frequently applies for different types of loans and pays back his credit obligations according to the scheduled repayment time. At the same time, a customer with a bad credit history also often applies for a loan in different places. However, in the majority of cases this borrower is rejected because of an unsatisfactory credit history which contains many derogatory reports and records on the past due accounts. Even if a bad credit history borrower is accepted for a loan he defaults with a higher probability.

For these two strictly dissimilar types of borrowers (a good credit history borrower and a bad credit history borrower), a lender would observe the same high number of enquiries. Consequently, if a fixed-effect coefficient is applied it leads to a situation when the impact of *Enquiries$_i$* on default is the same for a good and a bad borrower which is not realistic in practice. Assigning a varying-coefficient to the variable *Enquiries$_i$* helps to overcome this drawback. In this case the area-specific slopes are steeper in the poor living areas and flatter in the rich residence areas.

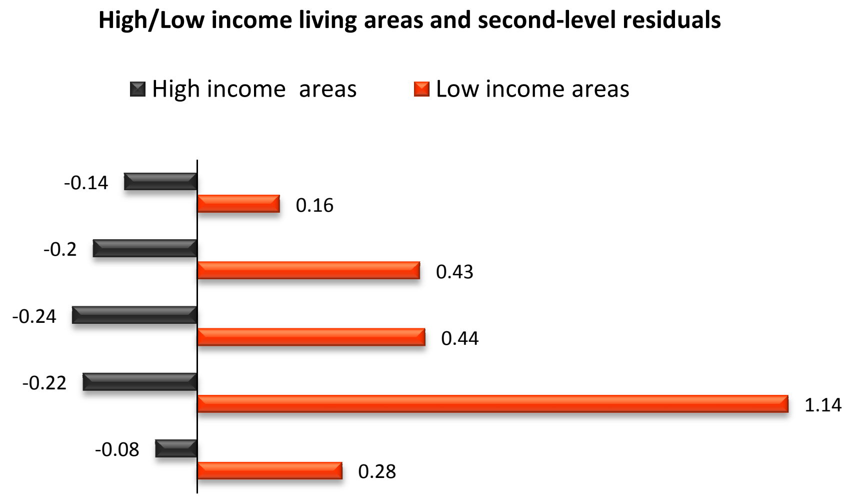


Figure 3.5. *Predicted microenvironment-specific effects for five lowest and five highest income areas.*

In order to visualize the last statement I graphically illustrate the impact of the number of credit enquiries on default within the low and high income microenvironments. Figure 3.5 illustrates the microenvironment-specific effects ($u_{j,Enq}$) predicted for the five lowest (red charts) and five highest income regions (grey charts). The abscissa axis on the graph shows the estimated residuals measured on the logit scale.

It is evident that the impact of the number of credit enquiries on probability is much more pronounced within the poorer microenvironments than within richer living areas.

The next figure visualizes the relationship between two varying-coefficients which are included in the multilevel credit scoring model in [2.9]. It is assumed that the area-specific coefficient of the variable $Enquiries_i$ and the coefficient of the variable $Past_{due,i}$ follow a multivariate normal distribution.

Figure 3.6 presents the pairwise residuals comparison plot for the varying-coefficients β_j^{Enq} and β_j^{Past} which are specified in [2.10]. The second-level residuals u_j^{Enq} are plotted on the abscissa axis and u_j^{Past} are given on the ordinate axis. It is evident from the plot that the correlation between microenvironment-specific effects is positive. This implies that the living areas with steep slopes of the number of credit enquiries are also going to have steeper slopes of the past due accounts. The upper-right red triangle corresponds to a low income area, with a high share of African-American residents and a low share of college graduates.

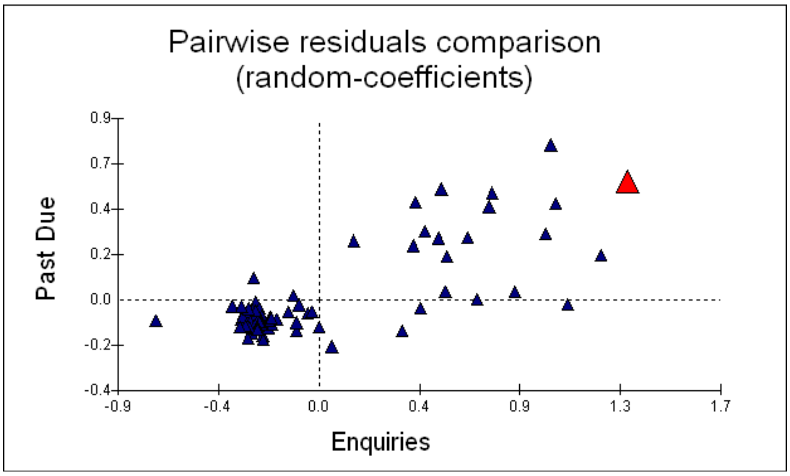


Figure 3.6. *A pairwise residuals comparison plot. Microenvironment-level residuals of the explanatory variable $Enquiries_i$ (number of credit enquiries) are plotted against second-level residuals of the variable $Past_{due_i}$ (number of credit delinquencies in the last 12 months). The highlighted in red residuals is for the lowest income area.*

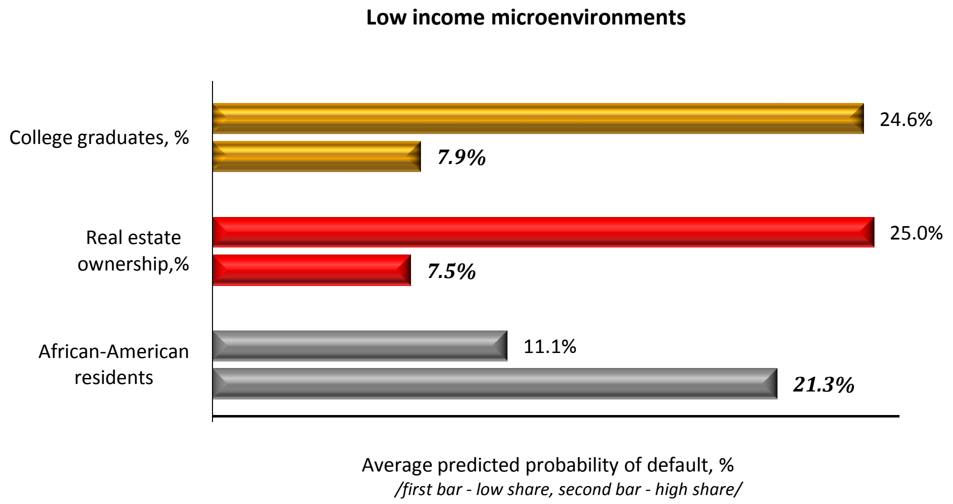
3.4.2 Predicted probabilities and living area economic conditions

In this subsection I show how to apply a graphical illustration of the fitted model predicted probabilities in the postestimation analysis and strategic planning in retail banking. Visualizing the probabilities not only allows easier interpretation of the results, it also helps to emphasize the role of the microenvironment-level characteristics and explore the impact of the economic and demographic conditions on default.

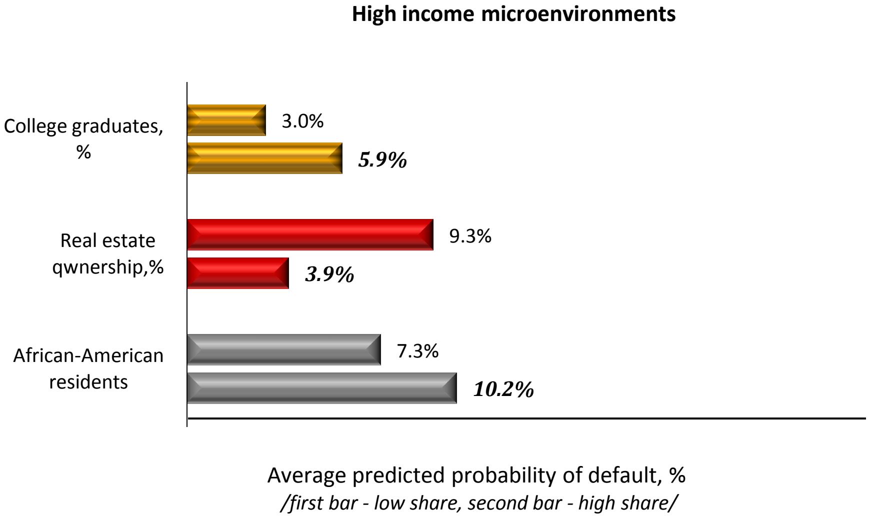
To compare the forecasts within the living areas with different economic and socio-demographic conditions, I calculate the average predicted probabilities of default within microenvironments. Figure 3.7 illustrates the results. The upper graph *a)* presents the probabilities of default for a low income microenvironment with a low / high share of college graduates in the market (orange bars), with a low/high share of African-American residents (grey bars) and with a low/high share of families who own a real estate property in the borrower's neighbourhood (red bars). Each bar on the graph illustrates the average riskiness of borrowers within a microenvironment with a particular combination of the living area conditions.

The comparison of the forecasts on the graph *a)* and *b)* reveals that the quality of borrowers is higher within the richer microenvironments compared to the poorer areas. Accordingly, the predicted probabilities of default in the high income areas are lower than in the low income regions. However, not only the regional level of income has an impact on the riskiness of customers. There are other microenvironment-level characteristics which should be considered. The forecasts on the graph *a)* show that within poor microenvironments the exposure to risk is higher in the areas with a higher share of African-American residents compared to the regions with a lower share of African-American residents (21.3% versus 11.1%). It is also true that within the low income regions the probability of default decreases if the level of the housing wealth or the share of college graduates in the market increase. Individuals within the areas where the majority of families own a real estate property are more financially stable. This leads to the average probability of default of 7.5%.

Controversially, the riskiness of borrowers increases up to 25% if a customer resides in a low income microenvironment with a low level of real estate wealth (the majority of families rent their accommodation). A high presence of college graduates in the area job market is negatively correlated with the probability of default. The average probability within low income regions with a high share of college graduates is 7.9%. This is 16.7% smaller than similar result for poor regions with a low share of college graduates. Similar conclusions can be made if the average probabilities of default are compared between different microenvironments but within the rich living areas. The probability of default is 10.2% in high income areas with a high share of African-American. This is 2.9% higher than the average riskiness of borrowers within rich regions with a low share of African-American residents. A house ownership in the area has negative impact on the riskiness. The probability of default within high income regions is 5.4% higher if the level of housing wealth within an area is low.



a). Average predicted probability of default for low income microenvironments with different composition of socio-demographic characteristics: with low/high share of college graduates in the market, low/high share of families with a real estate property and low/high share of African-American residents.



b). Average predicted probability of default for high income microenvironments with different composition of socio-demographic characteristics: with low/high share of college graduates in the market, low/high share of families with a real estate property and low/high share of African-American residents.

Figure 3.7. *Average predicted probabilities for microenvironments with different economic and socio-demographic conditions.*

In summary, the graphical illustration of the predicted probabilities and fitted model results not only shows the impact of economic and demographic conditions on default, it also reveals that exposure to risk within poor and rich regions depends not only on area income but also on the other living area characteristics such as real estate wealth, the share of African-American residents and the share of college graduates. Therefore, clustering of borrowers within microenvironments in the credit scoring model allows capturing the effect of the particular combination of living area conditions on default.

In addition, applying a graphical illustration of the predicted probabilities is very advantegeous for strategic planning in retail banking. It helps to detect areas where the exposure to the unobserved determinants of default is high. Given this information a lender can adjust his market strategy.

4 A cross-Classified Credit Scoring Model

In this chapter I introduce a new version of a multilevel credit scoring model which has a non-hierarchical structure. First, I describe the multilevel structure and show how to cluster borrowers within it. Second, I apply a non-hierarchical structure to the scorecards and estimate different versions of credit scoring models.

Importantly, the credit scorecards discussed in the previous chapter are specified with a hierarchical multilevel structure. The hierarchical nesting implies that all individual-level units (borrowers) are clustered within the second-level units (microenvironments). This chapter presents an extended version of the structure discussed previously which is more realistic in application to credit scoring. The new structure is a non-hierarchical one. It clusters individual-level units within the higher level classifications which are not nested one in the other. This kind of a multilevel structure is called cross-classified or non-nested.

The chapter is divided into three parts: structure, empirical analysis and predictive accuracy check. Section 4.1 introduces the structure and lists the characteristics which I apply to cluster borrowers within classifications at the second-level of the hierarchy. I present two specifications of credit scoring models with a cross-classified structure and provide empirical results for them in sections 4.2 and 4.3. The cross-classified scorecards differ by the composition of random-effects and explanatory variables measured at different levels of the hierarchy. The first version of a scorecard assigns a varying-intercept for each second-level classification. The second version elaborates the first and specifies group-level characteristics in the varying-intercept models at the classification-level of the hierarchy. Group-level information is presented by the explanatory variables defined within each of the second-level classifications. I apply a ROC curve analysis after estimation in order to check the predictive accuracy. In addition, I compute several other accuracy measures and show how to calculate an optimal cut-off point under particular conditions. The comparison of the goodness-of-fit measures and accuracy scores concludes the presentation of the empirical results for the fitted scoring models in section 4.3.

Importantly, credit scoring models with a cross-classified structure are computationally more complex than hierarchical scoring models. They contain several classifications which include random-effects and specify group-level characteristics at different levels. Maximum likelihood estimation is not an easy task in this case. Random-effects at the second-level have to be integrated out in the likelihood function which requires numerical integration techniques. Numerical approximation may fail to produce reasonable results when the number of random-effects is high. In order to overcome these computational problems, I apply Bayesian Markov chain Monte Carlo (MCMC) to fit the scorecards in this chapter.

In the case of multilevel modelling, Bayesian MCMC is a superior estimation approach. It is increasingly used as a method for dealing with problems for which there is no exact analytic solution and for which standard approximation techniques have difficulties. The basic principal of MCMC is to apply a Bayesian rule and carry out the necessary numerical integrations using simulations (Gelfand and Smith (1990)). The other motivation for the choice of this estimation approach is the flexibility of modelling random-effects. MCMC allows specifying different prior distributions for the group-specific effects and for the structural parameters (standard deviations, covariances). I provide a short summary of the estimation with Bayesian Markov chain Monte Carlo in chapter 5.

Before starting the next section, I briefly introduce the literature on non-hierarchical multilevel modelling. Although cross-classified models are computationally more complex than hierarchical multilevel models, the interest in using these structures in applied research is growing rapidly. The major advantage of cross-classified structures is that they better represent the complexity of real world situations where individuals may be subjects for multiple classifications. In particular, Zaccarin and Rivellini (2002) use multilevel cross-classified modelling in order to evaluate the effects of women's place of birth and women's current place of residence on the choice of bearing a second child by Italian woman in the mid-1990. In their structure the place of birth and current place of residence are the second-level classifications and women are nested into groups within each of the classifications. Goldstein and Fielding (2005) apply non-hierarchical multilevel modelling in the field of economics of education. They analyze students' examination results given that pupils are clustered within schools and at the same time within neighbourhood areas. The authors find that pupils' achievements are highly influenced by both school-specific and neighbourhood-specific characteristics.

4.1 Cross-classified structure of a scorecard

The credit scoring models presented in chapter 2 are specified with a hierarchical two-level structure. This implies that individual-level units (borrowers) are nested within the second-level units (microenvironments) which represent their living areas. Here I discuss other ways of clustering data for a credit scorecard. Alternatively, I could have defined a multilevel structure where borrowers are nested within clusters which describe their occupational activities or working experience. In this case, the structure would remain hierarchical. Applying this structure to the scorecard allows exploring the impact of unobserved occupation-specific effects on the probability of default. In general, there are many occupational hazards which influence the riskiness of individuals who are employed in different industries. Accounting for unobserved profession-specific characteristics improves the forecasting quality of a scorecard as more determinants of default are included.

In general, both types of a two-level structure (borrowers-within-microenvironments and borrowers-within-occupations) are relevant for more efficient credit scoring. Therefore in this chapter I combine these structures in one. The resulting multilevel structure is not hierarchical anymore because it nests borrowers within microenvironments and at the same time within their occupational activities. In this multilevel structure microenvironments and occupations are the second-level classifications which are not nested into each other.

The main advantage of a cross-classified structure over a hierarchical structure is that the former structure allows accounting not only for unobserved living area risks but also for occupation-specific determinants of default. In addition, the structure can incorporate group-level information which shows the impact of occupation-specific variables on the riskiness of borrowers. For instance, some changes within an industry may influence wages or employment which, in turn, impacts financial stability of individuals employed in these occupational fields.

Furthermore, I assume that there are infrastructure-specific determinants of credit worthiness which impact riskiness of borrowers additionally to the microenvironment-specific and occupation-specific effects. In general, the amount of credit burden and credit opportunities offered by lenders are highly correlated with the infrastructure of shopping facilities in the living areas of individuals. It is also true that good access to various department stores and shopping malls provokes spending and initiates borrowing. In order to satisfy the demand for credit resources lenders locate more branches and offices in areas with a highly developed infrastructure of shopping facilities. Accordingly, I specify the third classification – infrastructure and cluster borrowers within groups within different infrastructures.

Combining all three structures together produces a non-hierarchical multilevel structure with three classifications at the second-level: microenvironment, occupation and infrastructure. In this multilevel structure applicants for a loan are the individual-level units which are nested within groups and then within the second-level classifications. Separately, the structure within a classification is a hierarchical two-level. I cluster borrowers into groups according to the similarities in the particular characteristics of their occupations, living environments and infrastructure of shopping facilities.

It should be mentioned that it would be possible to specify other types of cross-classified structures which nest borrowers within different classifications and then within groups given a classification. However, in retail banking a decision about a particular structure for a credit scoring model should be guided by practical considerations within a lending institution. This dissertation focuses on a cross-classified structure which nests borrowers within occupations, microenvironments and infrastructures because, I suggest, that applying this structure helps to increase efficiency of credit worthiness assessment. The core idea here is that unobserved occupation, infrastructure and microenvironment-specific determinants of default have a noticeable impact on the probability and explain changes in the riskiness additionally to the observed characteristics on borrowers such as income, marital status and education.

4.1.1 Clustering within occupations

Clustering of borrowers within occupations allows exploring the impact of professional hazards on the probability of default. I start with an example in order to make the interpretation easier. Consider two individuals who apply for a bank loan, one is employed in military service and the other is in sales. According to some peculiarities of their professional activities they have different responsibilities, duties and working experience. These borrowers are subject to profession-specific hazards which differ across industries and occupational activities of individuals. On the one hand, a military man is exposed to multiple health-related hazards that originate in his working environment. On the other hand, a person employed in sales is influenced by other types of risks such as instability of wages or high labour fluidity in retail trade sector. Consequently, clustering of borrowers within occupations helps to account for unobserved occupation-specific hazards which are not similar in these two cases and which explain different triggering default factors.

I nest borrowers within occupations according to the similarities in the following characteristics of their professional activities: occupation, working experience and age. Table 4.1 provides a detailed list of the characteristics used in clustering. Each group within an occupation classification contains borrowers which are influenced by similar occupation-specific hazards. Alternatively, this group-effect can be viewed as an interaction effect of a particular profession with working experience and age.

Importantly, recognizing the impact of working experience and age on the riskiness implies that professional hazards have different impact on individuals with different experience and age given an occupation.

I model the exposure to the occupation-specific risks by including a random-intercept at the second-level of the hierarchy for the occupation classification. I define 70 groups within this classification.

4.1.2 Clustering within infrastructures

Individuals apply for the loan because they would like to smooth their consumption intertemporally. They use credit resources to make small purchases of durable goods, furniture, ordering vacation tours and for many other purposes including a car purchase. In living areas with a highly developed infrastructure of shopping facilities customers have access to a wider variety of goods and services which provokes spending and initiates borrowing. Therefore, I assume that there are unobserved infrastructure-specific determinants of default which should be included in a scoring model for a more efficient credit worthiness assessment.

I cluster borrowers within groups within the infrastructure classification according the similarities in the structure of shopping facilities in their neighbourhoods. Each cluster within infrastructures represents individuals who have similar access to the various shopping facilities and services in their residence areas. I measure access to shopping facilities by the percentage of retail store, dining, gas station, furniture, build materials and autohouse sales in the total sales in the local market. The determinants of clustering within infrastructures are given in the second column in Table 4.1.

The infrastructure classification has 50 clusters within which all borrowers are grouped. In a credit scorecard the infrastructure-specific effects are captured by a varying-intercept. Importantly, including unobserved infrastructure-specific characteristics explains that given personal information the riskiness of a borrower differs across living areas with good and bad access to various shopping facilities.

4.1.3 Clustering within microenvironments

Clustering of borrowers within microenvironments slightly differs from the one used in chapter 2. The main difference is that the information on the infrastructure of shopping facilities is not used in grouping within microenvironments. This is because now I define a separate classification, infrastructure, and apply these characteristics to nest borrowers within infrastructures. The other determinants of clustering within microenvironments remain the same as in chapter 2. Table 4.1 lists these determinants.

Occupation	*Infrastructure*	*Microenvironment*
Professional activity:	*Share in total sales*:	*Economic conditions*:
Management	Retail stores	Area Income
Military service	Autohouses	Housing wealth
Sales	Gasoline companies	Buying power index
Construction	Dinning & Catering	
High-skilled professionals	Medical & Drug stores	*Demographic conditions:*
Self-employed	Build materials	African-American (Hispanic) residents
Others	Furniture stores	Mean age
Working experience:	Apparel stores	Growth index
Less than 2 years		College graduates
3-5 years		
6-10 years		
More than 10		
Age :		
18-24		
25-30		
31-44		
45-60		
61-more		

Table 4.1. *Determinants of clustering within second-level classifications: occupations, microenvironments and infrastructures. Each cluster represents an interaction of the characteristics.*

I define 70 microenvironments within which individual applicants are nested according to the similarities in economic and socio-demographic conditions in their living areas. Each cluster within this classification represents a living environment of a borrower with a particular level of real estate wealth, per capita income, unemployment and with a particular demographic structure of residents (average age, share of African-American residents).

Microenvironment-specific effects are captured by a varying-intercept in a cross-classified scoring model. This intercept explains the exposure to the microenvironment-specific risks and hazards which trigger default on a loan.

4.1.4 Data and variables

I apply the same data on credit histories as in the previous chapters. The individual level data include personal information (income, marital status, dependents, etc.), Credit Reference Agency data (derogatory reports, enquiries, accounts past due, etc.) and living area descriptive data for the 5-digit area zip code in which a borrower resides (area income, demographic structure, house ownership, etc.). The full sample contains 9448 observations. I randomly split the sample into two parts: training and testing subsamples. The training dataset is applied to fit the scorecards. It contains 60% of the full sample. The testing sample is applied to check the classification accuracy of the out-of-sample predictions. It contains 3779 observations.

I apply a forward selection method to choose explanatory variables which are going to be included in the cross-classified scorecards. The variables are selected based on AIC criterion. Table 4.2 provides a short description of the selected characteristics. Importantly, this set of variables does not include the classification-level characteristics which are included in the scoring model in subsection 4.2.2. I combine market descriptive data with BEA data on regional economic accounts in order to construct the group-level variables.

Variable	*Description*
$Income$	Total annual income (including additional income) of an applicant for a loan, measured in thousands of dollars.
$Enquiries$	Number of credit enquiries in the credit profile of a borrower.
$Bank$	An indicator variable which takes the value of one if a borrower holds both bank savings and checking accounts.
Age	Age, in years.
$Trade$	Number of open and currently active trade accounts.
$Past_{due_{lines}}$	Total number of trade lines which are more than 30 day past due.
$Delinq$	Total number of 30-days delinquencies on credit obligations in the last 12 months.
$Credit_{exp}$	A dummy variable which equals one if a borrower has credit experience with a lender such as a personal loan or credit card (prior to the current application).
$Major$	Number of major derogatory reports in a credit profile of a borrower.
$Minor$	Number of minor derogatory reports in a credit profile of a borrower.
$Depend$	Number of dependents in a family.
$Prof$	A dummy variable which takes a value of one if a borrower is a high-skilled professional, and zero otherwise.
$Military$	A dummy variable which takes a value of one if an individual is employed in military service and zero otherwise.
Own_{rent}	A dummy variable which equals one if a borrower owns a real estate property (house, flat) and zero otherwise.
Rev_{credit}	Revolving credit balance (average over last 12 months).

Table 4.2. *Description of the explanatory variables used in the cross-classified credit scoring models.*

4.2 Empirical analysis

This section provides an empirical analysis for the credit scoring models with a cross-classified structure of the data. I introduce two versions of a scorecard which differ by the composition of random-effects and group-level variables. The first credit scoring model specifies a varying-intercept at the second-level for each cross-classification. The second version of a scoring model elaborates the first and additionally to the previous structure includes group-level characteristics in the second-level models for the varying-intercepts. The group-level variables capture the impact of occupation, microenvironment and infrastructure-specific characteristics on the probability of default.

I apply the training sample to fit the scoring models and the testing sample is used for the postestimation diagnostics. The credit scorecards with a cross-classified structure are complex and contain many random-effects. Therefore, I estimate them using a Bayesian MCMC approach. A ROC curve analysis concludes the presentation of the empirical results and provides a summary of different predictive accuracy measures. A pairwise comparison of the ROC curves and AUC values between the cross-classified models and the logistic scorecard is provided in section 4.3. In addition, I check the goodness-of-fit by applying DIC (Deviance information criteria) and evaluate the forecasting performance using different accuracy scores (logarithmic, spherical and brier score).

4.2.1 A cross-classified credit scorecard

The credit scoring model with a cross-classified structure is presented in [4.1]. The model assesses credit worthiness of borrowers by forecasting their probability of default. The dependent variable y_i is binary which takes a value of one if a borrower defaulted on his credit obligations and $y_i = 0$ if a borrower returned a loan without delinquencies. The individual-level explanatory variables are chosen using a forward selection method. The set of the selected explanatory variables includes 15 predictors. Importantly, the scorecard specified in [4.1] does not include group-level characteristics. This extension will be added in the next subsection.

In order to keep the notation transparent I do not apply multiple subscripts to indicate borrowers nested within classifications and within groups given a classification. Instead, I assign a subscript *j (for j=1,..,70)* to the groups within microenvironments, a subscript *k (for k=1,..,70)* to the groups within occupations and a subscript *l (for l=1,..,50)* to the groups within different infrastructures.

$$\begin{aligned} Pr\,[y_i = 1 \,|\, x_i\,, u_j,\, u_k\,, u_l\,] = Logit^{-1}\{ & \beta_0 + \beta_{inc}Income_i + \beta_{Enq}Enquiries_i + \beta_{bank}Bank_i \\ & + \beta_{age}Age_i + \beta_{tr}Trade_i + \beta_{past}Past_{due_i} + \beta_{del}Delinq_{i_i} \\ & + \beta_{Cr}Credit_{prev,i} + \beta_{maj}Major_i + \beta_{min}Minor_i + \beta_{pr}Prof_i \\ & + \beta_{dep}Dependents_i + \beta_{mil}Military_i + \beta_{rev}Rev_{credit_i} \\ & + \beta_{own}Own_i + u_{k[i]}^{occupation} + u_{j[i]}^{microenvrt} + u_{l[i]}^{infrastruct} \}. \quad [4.1] \end{aligned}$$

$$\begin{aligned} u_{j[i]}^{microenvnt} & \sim N\left(0, \sigma^2_{microenvrt}\right). \\ u_{j[i]}^{occupation} & \sim N\left(0, \sigma^2_{occup}\right). \\ u_{j[i]}^{infrastruct} & \sim N\left(0, \sigma^2_{infrast}\right). \quad [4.2] \end{aligned}$$

The random-effects in the scoring model are presented by the varying-intercepts within the second-level cross-classifications. I include the population average intercept β_0 in the scorecard; therefore, the varying-intercepts within classifications are constrained to have a zero mean. Similarly to the scorecards from the previous chapter, the classification-specific effects are presented by the second-level residuals. Residual $u_j^{microenvnt}$ describes the impact of the microenvironment-specific risks which vary across living areas with different economic and demographic conditions. Profession-specific hazards are captured by the term $u_k^{occupation}$ which varies across occupational activities of borrowers. The random-term $u_l^{infrastr}$ defines the infrastructure-specific effects.

Importantly, the classification-level residuals capture the information on unobserved determinants of default which influences riskiness of borrowers additionally to the individual-level characteristics such as income, marital status, etc. Given the explanatory variables it is assumed that the second-level random-effects follow a normal distribution with zero mean and variances $\sigma^2_{microenvt}$, σ^2_{occup} and $\sigma^2_{infrast}$ as shown in [4.2].

Credit scoring models with a cross-classified structure are more complex than the scorecards with a hierarchical structure in chapter 2. In addition, it is computationally difficult to fit them with maximum likelihood because the number of random-effects at the classification-level is high. There are 70 varying-intercepts within microenvironments, 70 varying-intercepts within occupations and 50 within infrastructures plus fixed-effects and variance parameters. In this case Bayesian Markov chain Monte Carlo is a superior estimation approach which allows more flexibility in random-effects modelling. I apply this approach to fit the scoring models in this chapter. According to the main Bayesian principle prior knowledge about random-effects distributions is updated by the data in order to obtain posterior distributions. Given posterior distributions it is straight forward to calculate mean and standard deviation of random-effects. I summarize the technical details of the estimation with Bayesian MCMC in chapter 5. In addition, this chapter discusses the alternative choices of prior distributions for random-effects and describes several tests which check the convergence of the algorithm.

Table 4.3 provides the estimation results for the cross-classified credit scorecard specified in [4.1]. The second part of the table reports the estimation results for the classification-level models. Standard deviations of the varying-intercepts are reported together with their 95% confidence intervals.

The results confirm that there is a negative impact of income, use of banking checking and savings accounts, number of trade accounts and previous experience with a lender on the probability of default. It is also true that ownership of real estate property decreases the riskiness.

Variable	*Estimate*	*S.E.*	*z-statistics*	*p-value*
Income	-0.031	0.006	-5.16	<0.001
Enquiries	0.231	0.030	7.70	<0.001
Bank	-0.359	0.132	-2.71	0.007
Age	-0.015	0.012	-1.25	0.211
Trade	-0.206	0.035	-5.88	<0.001
Past due trade lines	1.038	0.234	4.43	<0.001
Delinquencies in the last 12 months	0.100	0.062	1.61	0.107
Credit experience with a lender	-0.811	0.366	-2.21	0.027
Major derogatory reports	0.400	0.148	2.70	0.006
Minor derogatory reports	0.175	0.101	1.73	0.085
Dependents	0.193	0.055	3.55	<0.001
Professional	-0.780	0.269	-2.88	0.005
Military	0.109	0.349	0.31	0.775
Own/rent	-0.061	0.155	-0.52	0.603
Revolving credits	-0.017	0.012	3.29	<0.001
Constant	-1.937	0.354	5.47	<0.001
Second-level model	*Estimate*	*S.E.*	*95% confidence interval*	
Microenvironment				
Standard deviation	0.830	0.197	[0.475;	1.204]
Intercept, 80% credible interval	-	-	[-1.062;	1.062]
Occupation				
Standard deviation	0.630	0.180	[0.304;	0.984]
Intercept, 80% credible interval	-	-	[-0.806;	0.806]
Infrastructure				
Standard deviation	0.650	0.176	[0.332;	0.995]
Intercept, 80% credible interval	-	-	[-0.832;	0.832]

Table 4.3. *Estimation results for the cross-classified credit scoring model with random-effects of microenvironments, occupations and infrastructures. The standard deviations of the second-level residuals are reported with their 95% confidence intervals.*

At the same time, major and minor derogatory information has a significant and positive effect on the riskiness of individuals. A higher number of past due trade lines or delinquencies raises the probability of default. Probability of default is smaller for high-skilled professionals compared to unskilled workers.

It is evident that the estimated standard deviations of the second-level residuals are significantly larger than zero which confirms that the classification-specific intercepts vary across groups at the second-level of the hierarchy. Importantly, the 95% confidence intervals for these estimates do not include zero.

Given the normality assumption I also calculate 80% credible intervals for the varying-intercepts. The credible interval for random-effects within occupations implies that 80% of the realizations of the occupation-specific effects in population are going to lie within the interval [-0.806; 0.806]. Similarly, I calculate credible intervals for infrastructure-specific and microenvironment-specific effects which equal [-0.832; 0.832] and [-1.062; 1.062], correspondingly.

In order to check the discriminatory power of the credit scoring model with a cross-classified structure I apply a ROC curve analysis and calculate several other accuracy measures. Figure 4.1 illustrates the ROC curve for the cross-classified scorecard I. The upper and lower bounds on the graph represent the 95% pointwise confidence interval for the curve which are calculated according to Hilgers (1991). Table 4.4 reports several accuracy measures derived from the ROC curve.

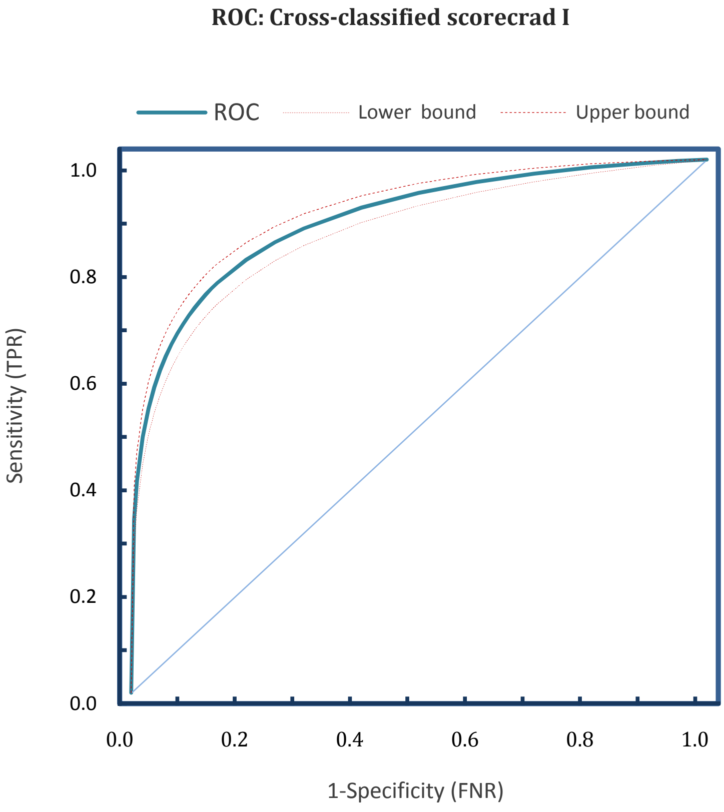


Figure 4.1. *ROC curve for the cross-classified score-card I.*

The area under the curve is 0.879 which is higher than the AUC value for the flexible credit scoring model in chapter 2 (scorecard 5). The Gini coefficient and the accuracy ratio are also increased. I test the difference in the AUC values between the cross-classified and a hierarchical scoring model by calculating the z-statistics as described in the previous chapter. The p-value of this difference is low and the 95% confidence intervals for the AUC measures do not overlap. This implies that specifying a cross-classified structure improves the discriminatory power of the scorecard.

ROC curve metrics	
Area under the ROC curve (AUC_{cross_1})	0.879
Standard error of AUC (bootstrap normal method)	0.009
95% confidence interval	[0.861; 0.897]
Gini coefficient	0.758
Accuracy ratio	0.866
Difference in AUC:	
$\Delta AUC = AUC_{cross_I} - AUC_5$	0.054
z-statistics of ΔAUC	5.809
p-value of ΔAUC	<0.001

Table 4.4. *The ROC curve metrics and the comparison of the AUC values between the cross-classified scorecard I and scorecard 5.*

On the next step I incorporate misclassification costs in the ROC curve analysis in order to compare the classification performance of the scorecards given asymmetric costs. In addition, I apply several alternative methods to compute an optimal cut-off point. Figure 4.2 visualizes the performance criteria for different thresholds within the range [0; 1]. The graph illustrates the sensitivity and specificity curves, the correct classification rate and the Cohen’s kappa curve.

Cohen's kappa coefficient is a statistical measure of inter-classifier agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than a simple percent agreement calculation since κ takes into account the agreement occurring by chance.

Cohen's kappa measures the agreement between two classifiers by calculating the kappa coefficient k as follows:

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

where $Pr(a)$ is the relative observed agreement among classifiers, and $Pr(e)$ is the hypothetical probability of a chance agreement (for details see Cohen (1960), Smeeton (1985)). In application to credit scoring $Pr(a)$ is the correct classification rate given a cut-off point and $Pr(e)$ is the sum of the joint probabilities

$$Pr(e) = P_{Obs}(D) \cdot P_{Pred}(D) + P_{Obs}(ND) \cdot P_{Pred}(ND),$$

where $P_{Obs}(D)$ and $P_{Obs}(ND)$ are the observed probabilities of default and non-default ; $P_{Pred}(D)$ and $P_{Pred}(ND)$ are the predicted probabilities of default and non-default. If the prediction models are in complete agreement then κ = 1. If there is no agreement among the models' predictions then κ ≤ 0.

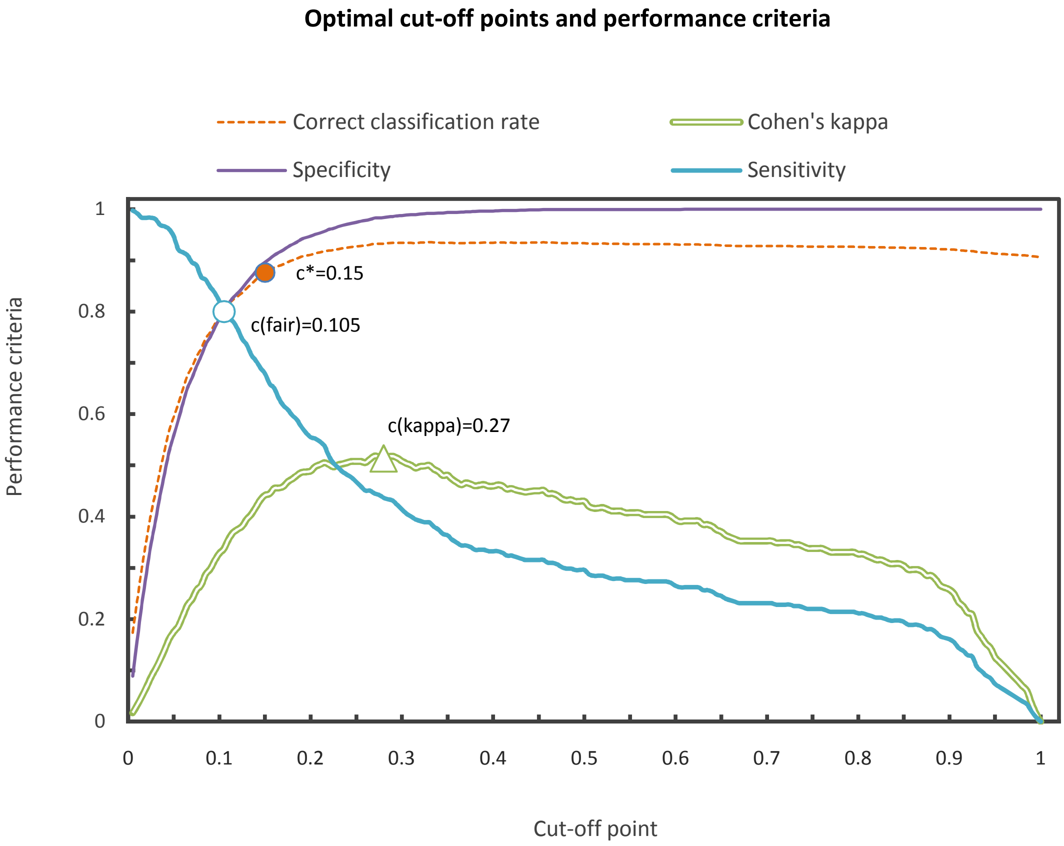


Figure 4.2. *Classification performance criteria: sensitivity, specificity, correct classification rate and Cohen's kappa curve.*

On the figure the Cohen's kappa curve is illustrated by the green double line. It is evident that the best agreement between observed and predicted outcomes is achieved at the cut-off point $c_{kappa} = 0.27$.

Table 4.5 illustrates several measures of the classification performance given the optimal threshold c_{kappa} on the kappa-curve and c_1 on the sensitivity curve. Given $c_{kappa} = 0.27$ the cross-classified scoring model correctly forecasts 158 of 355 true defaulters which yields a rather low sensitivity (44.5%). However, true negative rate and the rate of correct classifications are high. In addition, I evaluate the discriminatory power of a scorecard at the threshold $c_1 = 0.5$ in order to compare the discriminatory power of the model at the kappa-optimal point c_{kappa} and at the cut-off c_1. Setting a cut-off point at 0.5 indicates a very liberal way of accepting/rejecting applicants for a loan in retail banking. Given $c_1 = 0.5$ the scorecard produces very high specificity (99.9%) which in turn leads a high overall accuracy rate of 93.3%. However, the rate of true positive classifications is low and equals 29.6%. The classification performance of a scorecard at c_{kappa} is superior to the performance at c_1 as shown by the kappa-coefficient (0.5181 versus 0.4297). In application to retail banking, the cut-off point c_1 can be viewed as the upper bound for a threshold for accept/reject decision.

Cohen's kappa $c_{kappa} = 0.27$				*Cut-off point $c_1 = 0.5$*		
	Classified			**Classified**		
True	ND	D	Total	ND	D	Total
Non-default	3365	59	3424	3422	2	3424
Default	197	158	355	250	105	355
Total	3562	217	3779	3672	107	3779
Sensitivity, %			44.5			29.6
Specificity, %			98.3			99.9
Correct classification rate, %			93.2			93.3
Cohen's kappa			0.518			0.429

Table 4.5. *Classification table for the c_{kappa} and for the threshold c_1.*

The dashed orange curve on the graph presents the correct classification rate over all thresholds within the interval [0; 1]. I assume that the misclassification cost of an incorrectly predicted non-defaulter is five times higher than the cost of a falsely predicted defaulter. Then following Zweig and Campbell (1993) I calculate an optimal threshold for a accept/reject decision given asymmetric misclassification costs. The optimal cut-off point is indicated by the coloured circle on the dashed curve of correct classification rate. Table 4.6 presents the classification table for the optimal cut-off point c^*. Given $c^* = 0.1525$ the cross-classified scoring model properly predicts 3080 out of 3424 true negative outcomes which yields a rather high specificity (90%) and the correct classification rate (87.8%). The rate of true positive instances (sensitivity) is higher than at the threshold c_{kappa} and equals 66.5%.

An alternative to the optimal cut-off point c^* is a fair threshold which is illustrated by the empty circle on the sensitivity curve. This threshold is found on the intersection between sensitivity and specificity curves. Any thresholds above c_{fair} produce a higher true negative rate but a smaller true positive rate. Controversially, thresholds below c_{fair} produce a higher true positive rate but a smaller true negative rate. Table 4.6 presents the classification table for the optimal and fair cut-off points. The results confirm that the threshold $c_{fair} = 0.105$ is more conservative than the cut-off point c^*. Given $c_{fair} = 0.105$ the scorecard produces a higher rate of accurately predicted defaulters than at the optimal threshold c^*. However, specificity is more than 10% smaller at c_{fair} compared to c^*. The overall rate of correct classifications is also smaller at the fair cut-off point than at the optimal threshold. In retail banking, the fair cut-off point c_{fair} can be applied to set up a lower bound for an optimal threshold for an accept/reject decision.

Optimal cut-off point c=0.1525*				*Fair cut-off point: $c_{fair} = 0.105$*		
	Classified			**Classified**		
True	ND	D	Total	ND	D	Total
Non-default	3080	344	3424	2737	687	3424
Default	119	236	355	71	284	355
Total	3199	580	3779	2808	971	3779
Sensitivity,%			66.5			80.0
Specificity,%			90.0			79.9
Correct classification rate,%			87.8			80.0
Cohen's kappa			0.4395			0.3372

Table 4.6. *Classification table for the optimal and fair cut-off points. The performance criteria given the c*=0.1525 and $c_{fair} = 0.105$.*

In summary, I calculate and discuss four alternative choices for an optimal threshold for a accept/reject decision which range from a conservative c_{fair} (moderate c^* or c_{kappa}) to an liberal c_1. It is evident that given a cut-off point the classification performance of the scorecards varies considerably between these alternatives.

In retail banking a lender can apply these methods to set up an optimal cut-off point which is applied in order to discriminate the population of borrowers into two classes: accepted and rejected applicants. Importantly, the choice of an optimal threshold should be guided by practical considerations within a financial institution. In general, lenders define a threshold for an application credit scoring based on their risk attitudes. A risk-averted creditor prefers to minimize losses given default. Therefore, he chooses the fair c_{fair} or optimal c^* cut-off points which provide him a high sensitivity at the cost of low specificity. A profit-maximizing lender chooses a threshold for a decision-making from the range between c_{kappa} and $c = 0.5$. This guarantees him a high rate of correct classifications and a high true negative rate.

4.2.2 Classification-level characteristics in the cross- classified credit scorecard

In this subsection I introduce a new version of a cross-classified scoring model. This scorecard extends the previous model by including group-level characteristics at a higher level of the model hierarchy. Group-level variables are included in the varying-intercept models within microenvironment, occupation and infrastructure classifications. Accounting for classification-specific characteristics improves the estimation and increases the accuracy of random-effects predictions. In addition, it allows exploring the impact of group-level information on the probability of default. I combine living area descriptive data with aggregated individual-level data in order to define group-level explanatory variables.

The credit scoring model with group-level variables and varying-intercepts is specified in [4.3]. I apply the same set of the explanatory variables as in the credit scorecard in [4.1]. Given group-level characteristics random-intercepts within cross-classifications are modelled by themselves at the second-level of the hierarchy. The varying-intercept models for microenvironment, occupation and infrastructure classifications are presented in [4.4]-[4.6].

$$\begin{aligned} Pr\left[y_i = 1 \middle| x_i, u_j, u_k, u_l\right] = \; & Logit^{-1}\{\ \beta_0 + \beta_{inc}Income_i + \beta_{Enq}Enquiries_i + \beta_{bank}Bank_i \\ & + \ \beta_{age}Age_i + \beta_{tr}Trade_i + \beta_{past}Past_{due_i} + \beta_{del}Delinq_{i_i} \\ & + \ \beta_{Cr}Credit_{prev} + \beta_{maj}Major_i \ + \beta_{min}Minor_i + \beta_{dep}Dependnts_i \\ & + \ \beta_{pr}Prof_i \ + \ \ \beta_{mil}Military_i + \ \beta_{own}Own_i + \beta_{rev}Rev_{credit_i} \\ & + \ \beta_{bur}Burden_i \ + \beta^{microenvrt}_{j[i]} + \beta^{Occupation}_{l[i]} + \beta^{Infrastr}_{k[i]}\ \}. \qquad [4.3] \end{aligned}$$

$$\begin{aligned} \beta^{microenvrt}_{j[i]} &= \gamma_{Inc}Area_{income_j} + \gamma_{Own}Own_{rent,j} + \gamma_{AA}African_{Am,j} + u^{microen}_{j[i]}. \quad [4.4] \\ \beta^{Occupation}_{l[i]} &= \gamma_{coll}\ College_l \ + \ \gamma_{age}\ Mean_{age_l} + \gamma_{ex}Exp_l + u^{occupation}_{l[i]}. \quad [4.5] \\ \beta^{Infrastructure}_{k[i]} &= \gamma_{bur}\ Credit_{bur_k} + \ \gamma_{tr}\ Trade_k + \gamma_{del}Delinquen_k + u^{infrastr}_{l[i]}. \quad [4.6] \\ u^{microenvrt}_{j[i]} &\sim N\left(0, \sigma^2_{microenvrt}\right). \\ u^{occupation}_{k[i]} &\sim N\left(0, \sigma^2_{occupation}\right). \\ u^{infrastruct}_{l[i]} &\sim N\left(0, \sigma^2_{infrastr}\right). \quad [4.7] \end{aligned}$$

The microenvironment-level model for the varying-intercept $\beta^{microenvrt}_{j[i]}$ contains information on the living area economic and demographic conditions and the area-specific residual $u^{microenvnr}_{j[i]}$. The microenvironment-level variables are $Area_{income,j}$ - per capita area income, $Own_{\%j}$ - the level of real estate wealth (percentage of families who own a house in a living area) and the share of African-American residents in the living area of a borrower ($African_{Am,j}$).

The occupation-specific intercept $\beta_{l[i]}^{Occupation}$ explains exposure to occupational hazards given professional activity, working experience and age. The varying-intercept model contains three occupation-level characteristics and a random-term. The group-level variables are the share of college graduates ($College_l$), average age of borrowers ($Mean_{age_l}$) and average working experience given an occupation ($Exper_l$). The occupation-specific residual $u_{l[i]}^{occupation}$ explains changes in the probability over and above the population average value.

The infrastructure-specific intercept $\beta_{k[i]}^{Infrastructure}$ captures the effect of shopping facilities in a living area on the probability of default. The varying-intercept model specifies three group-level variables which characterize borrowers' credit worthiness within an infrastructure. The variables are $Burden_{[k]}$ – average amount of the credit card burden per household member given an infrastructure, $Trade_k$ – average number of currently active trade accounts in the last 12 months and $Delinquen_k$ – average number of 30-days delinquencies on credit obligations in the last 12 months.

The model specification in [4.3] includes a population average intercept β_0. Therefore, the varying-intercept models for occupation, infrastructure and microenvironment classifications are constrained to have a zero mean.

It is assumed that given the borrower-level and classification-level variables the second-level residuals within microenvironments ($u_{j[i]}^{microenvrt}$), occupations ($u_{k[i]}^{occupation}$) and infrastructures ($u_{l[i]}^{infrastr}$) are independently normally distributed with mean 0 and variances $\sigma^2_{microenvrt}$, $\sigma^2_{occupation}$, $\sigma^2_{infrastr}$ as given in [4.7]. I apply Bayesian MCMC to estimate the cross-classified scoring model.

Table 4.7 presents the estimation results for the individual-level explanatory variables and for the classification-level characteristics in the varying-intercepts.

It is evident that the population average effects of the individual-level variables are similar to the estimates from the previous scorecard. Probability of default decreases if a borrower has previous experience with a lender, holds banking savings and checking accounts and owns a real estate property. Derogatory information in credit history has a significant positive impact on the riskiness of applicants for a loan.

Variable	*Estimate*	*S.E.*	*z-statistics*	*p-value*
Income	-0.031	0.007	-4.429	<0.001
Enquiries	0.235	0.037	6.351	<0.001
Bank	-0.401	0.159	-2.522	0.011
Age	-0.021	0.010	-2.100	0.035
Trade	-0.206	0.038	-5.421	<0.001
Past due trade lines	1.455	0.278	5.234	<0.001
Delinquencies in the last 12 months	0.057	0.080	0.713	0.475
Credit experience with a lender	-0.922	0.430	-2.144	0.032
Major derogatory reports	0.406	0.147	2.762	0.005
Minor derogatory reports	0.189	0.102	1.853	0.063
Dependents	0.231	0.066	3.500	<0.001
Professional	-0.769	0.260	-2.958	0.003
Military	0.121	0.305	0.397	0.691
Own/rent	-0.109	0.154	-0.708	0.478
Revolving credits	0.024	0.007	3.429	0.145
Constant	-2.920	1.031	-2.832	0.004
Microenvironment-specific intercept model	*Estimate*	*S.E*	*95% confidence interval*	
$Area_{income_j}$	-0.107	0.034	-0.056	-0.172
$Own_{rent,j}$	-0.106	0.024	-0.063	-0.150
$African - American_j$	0.129	0.035	0.065	0.205
SD microenvironment (intercept)	0.541	0.179	0.222	0.881
Occupation-specific intercept model	*Estimate*	*S.E*	*95% confidence interval*	
$College_l$	-0.137	0.048	-0.236	-0.041
$Mean_{age_l}$	0.108	0.036	0.053	0.189
Exp_l	-0.256	0.104	-0.460	-0.071
SD occupation (intercept)	0.347	0.172	0.037	0.673
Infrastructure-specific intercept model	*Estimate*	*S.E.*	*95% confidence interval*	
$Trade_k$	-1.082	0.209	-1.290	-0.511
$Burden_{[k]}$	0.305	0.033	0.250	0.378
$Delinquen_k$	0.398	0.112	0.345	0.392
SD infrastructure (intercept)	0.440	0.185	0.110	0.791

Table 4.7. *Estimation results for the cross-classified credit scoring model II. Estimated standard deviations of random-effects are reported together with their 95% confidence intervals.*

The coefficient estimates of the microenvironment-intercept model imply that a thousand increase in the living area income decreases the probability by 2.67%. A similar effect is found for the level of housing wealth in a residence area. The standard deviation of the random-term $u_{j[i]}^{microenvrt}$ is smaller than in the case of the credit scoring model without classification-level characteristics. This is intuitive as specifying group-level variables improves the estimation. The 80% credible interval for the microenvironment-specific effects is [-0.69; 0.69].

The estimation results for the infrastructure-level model show that the number of current active trade accounts has a negative impact on the probability. An increase in the amount of credit card burden (per household member) raises the probability by 7.6%. Additional delinquency leads to a 9.9% increase in the riskiness. The standard deviation of the infrastructure intercept is 0.44 on the logit scale. The 80% credible interval for the infrastructure-level residual equals [-0.56;0.56].

It should be mentioned that not all of the coefficients of the classification-level variables are precisely estimated which not surprising is given the training data sample is not large enough. However, I keep reporting them in the credit scoring model. I suggest, that economic significance of these variables is high and the information they incorporate is relevant for more efficient credit scoring. Observing a larger sample on credit histories of borrowers can resolve this problem and provide better inferences.

Similarly to the previous model, I assess the predictive accuracy of the cross-classified scorecard with classification-level variables by applying a ROC curve analysis after estimation and by calculating other accuracy measures derived from the curve. In addition, I evaluate and compare several alternative values for an optimal threshold for an accept/reject decision. Figure 4.3 illustrates the ROC curve for the cross-classified scorecard 2.

Table 4.8 reports the accuracy measures derived from the ROC curve. The area under the curve is 0.894. The difference in AUC values between the cross-classified scorecard II and scorecard 5 from chapter 2 is statistically significant with a very low p-value. Similarly, I compare the difference in AUC measures between the cross-classified scorecard I and scorecard II. The difference is small yielding a test result which is only significant at the 10% level.

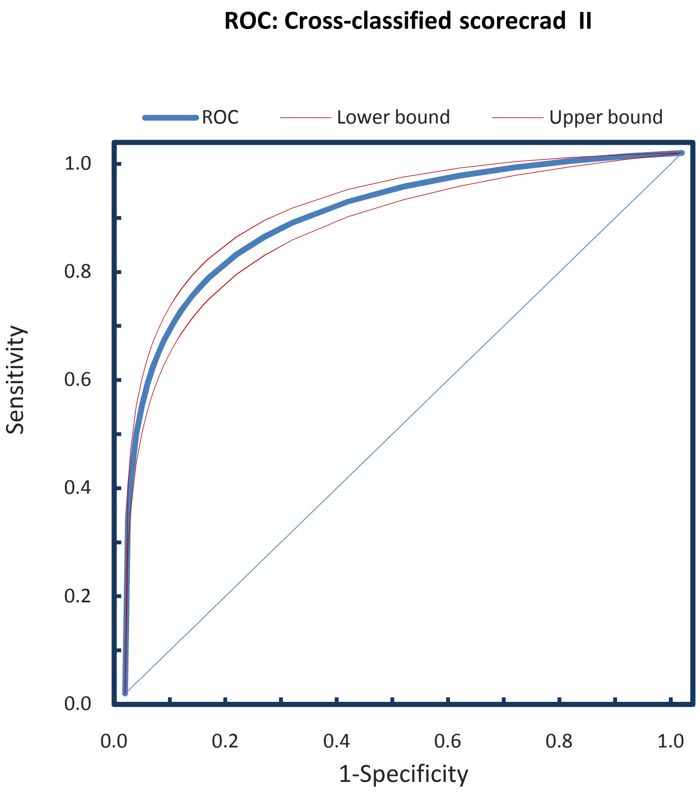


Figure 4.3. *ROC curve for the cross-classified scorecard with classification-level variables.*

Statistics	
Area under the ROC (AUC_{cross_1})	0.894
Standard error of AUC (bootstrap normal method)	0.009
95% confidence interval	[0.876; 0.911]
Gini coefficient	0.788
Accuracy ratio	0.900
Comparison of the areas under ROC curve	
1).	
$\Delta AUC = AUC_{cross_2} - AUC_5$	0.069
z-statistics of ΔAUC	7.560
p-value of ΔAUC	<0.001
2).	
$\Delta AUC = AUC_{cross^{II}} - AUC_{cross^{I}}$	0.015
z-statistics of ΔAUC	1.601
p-value of ΔAUC	0.101

Table 4.8. *ROC curve metrics and comparison of the AUC values between the cross-classified scorecard II and scorecard 5 from chapter 2.*

Next, I compare the classification performance of the scorecard at different cut-off points and compute an optimal threshold given an accuracy curve. Figure 4.4 illustrates sensitivity and specificity curves, correct classification and Cohen's kappa coefficient curves.

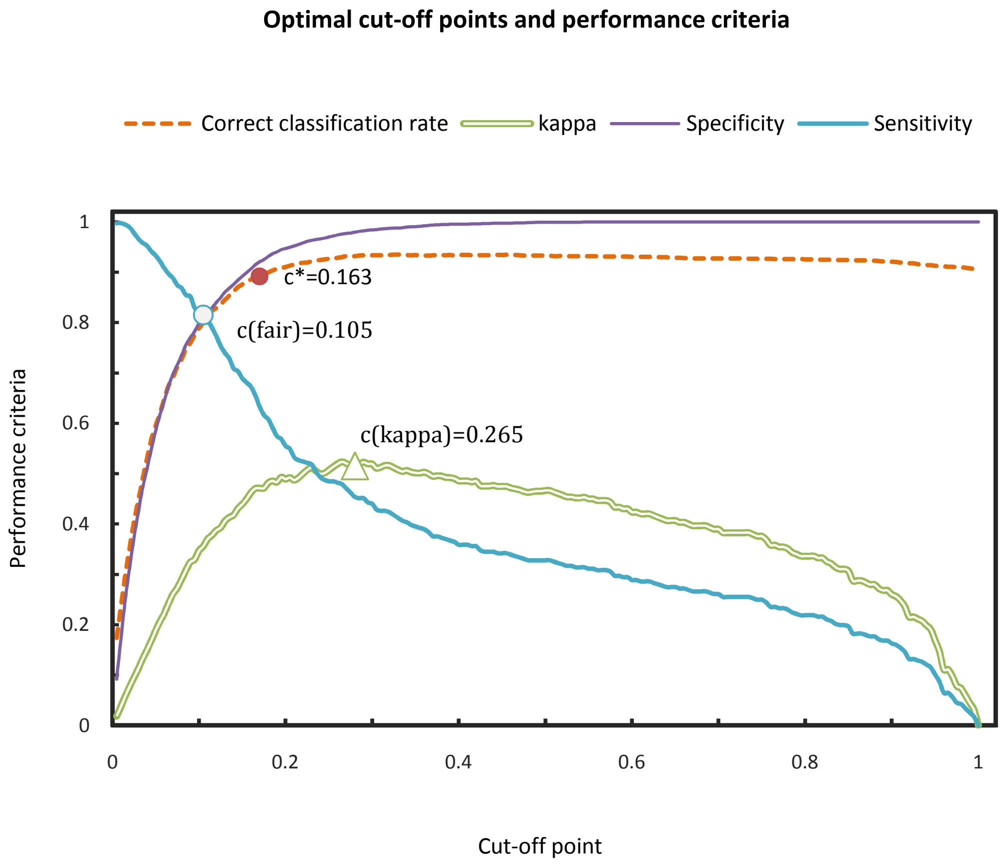


Figure 4.4. *Classification performance criteria for the cross-classified scorecard II: sensitivity, specificity, correct classification rate and Cohen's kappa coefficient.*

The highest Cohen's kappa coefficient is reached at $c_{kappa} = 0.265$ as indicated by the empty triangle on the graph.

Given c_{kappa} the cross-classified scorecard II produces a higher true positive rate than the scorecard I without group-level characteristics. However, the correct classification rates are practically the same. The Cohen's coefficient at the cut-off point c_{kappa} is higher in the current case which confirms that the cross-classified scorecard II provides a higher discriminatory power than the scorecard I. Table 4.9 compares the classification performance of the scorecards given c_{kappa} and $c_1 = 0.5$. The overall predictive accuracy (or the correct classification rate) is higher at the liberal cut-off point c_1 compared to the kappa-optimal c_{kappa}. However, given c_{kappa} the credit scoring model II provides a higher true positive rate than given the cut-off point c_1. Specificity is higher at *c_1=0.5*.

Cohen's kappa c_{kappa}=0.265				*Cut-off point c=0.5*		
	Classified			**Classified**		
True	ND	D	Total	ND	D	Total
Non-default	3342	82	3424	3422	2	3424
Default	184	171	355	238	117	355
Total	3526	253	3779	3660	119	3779
Sensitivity,%			48.2			33.0
Specificity, %			97.6			99.9
Correct classification rate,%			93.0			93.6
Cohen's kappa			0.5233			0.4664

Table 4.9. *Classification tables for the Cohen's kappa threshold and liberal cut-off point for probability of default.*

The optimal cut-off point for the correct classifications curve is illustrated by the coloured (red) circle on the graph. Table 4.10 shows that at the threshold *c*=0.163* the credit scorecard II shows the best classification performance yielding a misclassification error of 11.2%. Given *c** the true positive rate equals 67.0% and the true negative rate equals 91.0%. Compared to the scorecard without group-level characteristics, the overall accuracy rate and Cohen's kappa at *c** are increased.

The intersection of sensitivity and specificity curves illustrates the optimal threshold c_{fair}. It is indicated by the empty circle on the graph. Given c_{fair}=0.105 the cross-classified scoring model II provides a high classification rate of true positive outcomes (82%). However, true negative rate and correct classification rate are smaller than at the optimal cut-off point c*. Comparing Cohen's kappa coefficients reveals that the scorecard II outperforms scorecard I by showing better classification agreement between observed and forecasted outcomes.

In summary, I evaluate different predictive accuracy measures in order to compare the classification performance of the cross-classified scoring models. In addition, I apply several alternative methods to illustrate how to assess an optimal cut-off point for an accept/reject decision in retail banking. The results confirm that the cross-classified scorecard II outperforms the scorecard I by providing a higher discriminatory power.

Optimal cut-off point c=0.1630*				*Fair cut-off point:* $c_{fair} = 0.105$		
	Classified			**Classified**		
True	ND	D	Total	ND	D	Total
Non-default	3117	307	3424	2762	662	3424
Default	117	238	355	64	291	355
Total	3234	545	3779	2826	953	3779
Sensitivity,%			67.0			82.0
Specificity,%			91.0			80.7
Correct classification rate,%			88.8			80.8
Cohen's kappa			0.4669			0.3557

Table 4.10. *Classification tables for the optimal and fair cut-off points.*

4.3 Goodness-of-fit and accuracy scores

This section applies several postestimation diagnostic statistics in order to evaluate the goodness-of-fit of the estimated cross-classified scorecards. In addition, I apply several accuracy scores to check the forecasting performance. In particular, I compute and report deviance information criterion (DIC), logarithmic and spherical scores and Brier score. The logistic regression scorecard is used as a reference model for the between-models comparison.

In multilevel modelling, deviance information criterion (DIC) is applied in order to select the best performing model among the range of models estimated with Bayesian MCMC. In other words, the model with the smallest DIC is considered to be the model that would best predict a replicate dataset which has the same structure as the one currently observed. According to Spiegelhalter and Best (2002) DIC is calculated as follows:

$$DIC = \bar{D} + p_D = D(\bar{\theta}) + 2p_D,$$

where $\bar{D}$ is the expected measure of deviance $\bar{D} = E^{\theta}[D(\theta)]$ of how well the model fits the data. The deviance is defined as $D(\theta) = -2\log(p(y|\theta))$, where y are the data, θ are the unknown parameters of the scorecard and $p(y|\theta)$ is the likelihood function.

The effective number of parameters of the model is given by $p_D = \bar{D} - D(\bar{\theta})$, where $\bar{\theta}$ is the expectation of θ. $\bar{D}$ is the posterior mean of deviance and $D(\bar{\theta})$ is the deviance of the posterior means. Models are penalized by both the value of $\bar{D}$, which favors a good fit, but also (in common with AIC and BIC) by the effective number of parameters p_D. Since $\bar{D}$ will decrease as the number of parameters in a model increases, the p_D term compensates for this effect by favouring models with a smaller number of parameters.

I apply DIC in order to assess the fit and compare the cross-classified scorecards and the logistic regression scoring model. Table 4.11 reports DIC, mean deviance and effective number of parameters.

	$\bar{D}$	$D(\bar{\theta})$	p_D	DIC
Logistic scorecard	2080	2064	15.88	2095.9
Scorecard I	1567	1430	137.1	1704.1
Scorecard II	1479	1337	141.9	1621.9

Table 4.11. *Deviance information criterion, mean deviance and effective number of parameters.*

The results confirm that as the scoring models get more complicated, the mean deviance $\bar{D}$ decreases (measure of fit) which makes sense. More elaborated structures provide a better fit to the data. The largest jump in the DIC values is found between the logistic scorecard and the scorecard I. This illustrates the impact of a cross-classified structure application to a credit scoring model. The scorecard II with three varying-intercepts and group-level characteristics provides the best fit.

Next, I compute several accuracy scores in order to compare the forecasting quality between the cross-classified scoring models and the logistic regression scorecard. Table 4.12 reports the logarithmic, spherical and Brier scores.

Accuracy scores:	*Logarithmic*	*Spherical*	*Brier*
Logistic scorecard	-0.2685	0.9168	0.0771
Scorecard I	-0.1962	0.9415	0.0543
Scorecard II	-0.1922	0.9429	0.0532

Table 4.12. *Accuracy scores: comparison of brier score, logarithmic and spherical scores between cross-classified credit scorecards and logistic regression scoring model.*

In summary, it is evident that the accuracy scores provide the same ranking to the credit scorecards. The largest jump in the scores is found between the cross-classified scoring models and the logistic regression scorecard. The accuracy scores for the cross-classified scorecards are similar. However, it is evident that the cross-classified scorecard II provides a higher classification performance than the scorecard I.

5 Estimation techniques

In this chapter I provide some technical details on the estimation approaches which are applied to fit the multilevel scorecards. Since estimation techniques are not the main topic of this dissertation, I only give a basic overview of the methods used in the previous chapters.

This thesis discusses two types of multilevel scoring models which cluster borrowers within hierarchical and cross-classified structures. For each structural type, I present several variations of scorecards which differ by the degree of complexity and combine random-effects at different levels. Accordingly, I apply maximum likelihood to fit the hierarchical two-level models. Section 5.1 provides an overview of maximum likelihood estimation. The cross-classified scorecards are much more complex than the hierarchical scoring models. It is computationally more difficult to fit them by maximum likelihood. Therefore, I estimate the non-hierarchical credit scorecards with Bayesian Markov chain Monte Carlo. Section 5.2 discusses the main advantages of this estimation approach and explains the choices of prior distributions for random-effects and main model parameters. In addition, I report several diagnostics to check the convergence of the Monte Carlo algorithm.

5.1 Maximum likelihood estimation

The credit scorecards in chapter 2 are extensions of generalized linear models which are specified with a hierarchical two-level structure. In order to estimate these models with maximum likelihood, random-effects at the second-level have to be integrated out in the likelihood. This requires application of numerical methods. Following Rabe-Hesketh and Skrondal (2002) I apply adaptive Gauss-Hermite quadrature to approximate the marginal likelihood by numerical integration.

As an example, I take a simple two-level logistic regression scorecard as given in [2.5] in order to illustrate the estimation with maximum likelihood and explain the main assumptions. The other hierarchical models can be estimated in a similar way. In reduced form the credit scoring model with random-intercept and a single predictor x_i can be written as follows

$$\begin{aligned} Pr(y_i = 1|x, u) &= Logit^{-1}(\ \beta_{j[i]} + \beta_1 x_i\ ), \\ \beta_{j[i]} &= \beta_0 + u_{j,0}, \\ for\ microenvironments\ \ j &= 1, .., N. \end{aligned} \qquad [5.1]$$

For a fixed microenvironment j, the marginal likelihood for the multilevel scorecard in [5.1] is the joint probability of all observed responses y_i given the observed explanatory variable x_i. Importantly, the dependent variables y_i are conditionally independent given the second-level residual $u_{j,0}$ and the predictor variable x_i. Therefore, the conditional density function $f(y_j|x_i, u_{j,0})$ for microenvironment j, given the $x_i, u_{j,0}$, is the product of the probabilities of individual responses as shown in [5.2]. The number of level-one units within a level-two microenvironment is given by n_j.

$$f(y_j|x_i, u_{j,0}) = \prod_{i=1}^{n_j} \frac{\exp(\beta_0 + \beta_1 x_i + u_{j,0})^{y_i}}{1 + \exp(\beta_0 + \beta_1 x_i + u_{j,0})}. \qquad [5.2]$$

The last term in [5.2] is given by

$$\frac{\exp(\beta_0 + \beta_1 x_i + u_{j,0})^{y_i}}{1 + \exp(\beta_0 + \beta_1 x_i + u_{j,0})} = \begin{cases} \dfrac{\exp(\beta_0 + \beta_1 x_i + u_{j,0})}{1 + \exp(\beta_0 + \beta_1 x_i + u_{j,0})}, & if\ y_i = 1 \\ \dfrac{1}{1 + \exp(\beta_0 + \beta_1 x_i + u_{j,0})}, & if\ y_i = 0. \end{cases}$$

The random-intercept $u_{j,0}$ is assumed to follow a normal distribution with mean 0 and variance σ_u^2. The unconditional density $f(y_j\,|x_i)$ for microenvironment j is the product of the conditional density and the distribution function for the random-effect $g(u_{j,0})$.

$$f(y_j|x_i) = \int f(y_j|x_i\,,u_j)\; g(u_{j,0})du_{j,0}. \qquad [5.3]$$

The integral in [5.3] does not have a closed form solution which requires application of numerical approximation techniques.

Assuming that microenvironments are independent, the marginal likelihood of all responses for all microenvironments can be written as the product of unconditional densities as shown in [5.4]. The marginal likelihood is a function of the parameters β_0, β_1 and σ_u^2. The maximum likelihood estimates of β_0, β_1 and σ_u^2 are the values that jointly maximize $L(\beta_0, \beta_1, \sigma_u^2)$.

$$L(\beta_0, \beta_1, \sigma_u^2) = \prod_{j=1}^{N} \int f(y_j|x_i\,,u_j)\; g(u_{j,0})\,du_{j,0}. \qquad [5.4]$$

The Gauss-Hermite quadrature assumes that the integral over the random-effect in [5.3] can be approximated by the sum of R terms with e_r substituted for the $u_{j,0}$ and the normal density replaced by a weight w_r for the r-th term for $r = 1, \dots R$.

$$f(y_j|x_i) \approx \sum_{r=1}^{R} f(y_j|x_i, u_{j,0} = e_r)\, w_r. \qquad [5.5]$$

The approximation in [5.5] replaces the continuous density for the random-effect $u_{j,0}$ by a discrete distribution with R possible values of $u_{j,0}$. This means that increasing the number of integration points R helps to improve the approximation.

The adaptive quadrature is an extension of the standard Gauss-Hermite quadrature which allows accounting for situations where the number of observations within groups is large or intra-class correlations are very high. I refer to Rabe-Hesketh, Skrondal and Pickles (2002, 2005) for a more detailed description of the adaptive quadrature approach.

The adaptive quadrature improves approximation by applying some adjustments of the location e_r. Similar to the regular quadrature, in order to maximize the likelihood the adaptive quadrature starts with some initial values for the parameters and then updates the parameters until the likelihood is maximized.

In summary, Gaussian quadrature shows the best performance if second-level groups are small and the intra-class correlation is not too high. In the multilevel models with many random-effects one has to use a large number of quadrature points in order to get a good approximation. Adaptive quadrature works much better than regular quadrature. In particular, it is suitable for the case of a non-normal density of random-effects.

The alternatives to the Gauss-Hermite quadrature are the iterative generalized least squares (IGLS) or reweighted iterative generalized least squares (RIGLS) combined with marginal quasi-likelihood methods (MQL) or with penalized quasi-likelihood (PQL). Both, the MQL and PQL procedures use a linearization method based on a Taylor series expansion which transforms a discrete response model to a continuous response model. In this thesis I apply these estimation approaches to the scorecards fitted in MLwiN in order to provide the graphical illustration of random-effects in chapter 3. I refer to Goldstein [2003] for a more detailed description of MQL (PQL) approaches.

5.2 Bayesian inference with MCMC

This section provides a basic summary of the estimation with Bayesian Markov chain Monte Carlo. The cross-classified scoring models in chapter 4 are more complex than hierarchical scorecards. They contain many random-effects and group-level variables at different levels of the hierarchy. Therefore, I apply

Bayesian MCMC to fit the non-hierarchical credit scorecards. More technical details about the estimation procedure can be found in Lunn, Thomas, Best and Spiegelhalter [2000].

There are two major reasons why I choose Bayesian MCMC for estimating cross-classified scoring models. First, the main advantage of this method over other estimation approaches is the flexibility of modelling complex structures which include random-effects at different levels. In addition, Bayesian MCMC is intuitive in the case of a multilevel scorecard because it allows incorporating uncertainty about microenvironment-specific effects. The main difference from the classical statistical theory implies that some prior knowledge about the unknown model parameters (random-variables, standard deviations) can be used. Each parameter in the model is assigned with a prior probability distribution. Prior distributions express ex-ante believes about the parameters before the knowledge on the observed data is added.

The second advantage of the Bayesian approach is computational efficiency. Bayesian MCMC performs better and produces more accurate results than maximum likelihood approaches in the case of non-hierarchical models.

The basic idea of the Bayesian approach imposes that combining prior knowledge and the observed data it is possible to make statistical inference about the posterior distribution of unknown parameters given the data. The posterior distribution is viewed as the target distribution from which the random-effects are drawn. In the case of multilevel credit scoring models the main interest lies in making inferences about the population values of random-effects.

In application to credit scoring posterior distributions of random-effects are calculated by combining historical credit history data on borrowers and some knowledge about their prior distributions. In mathematical terms, the Bayes theorem states that the posterior distribution $p(\theta|y)$ of scorecard parameters $\theta = [\beta, u, \sigma_u]$ given the observed data y can be written in the form presented in [5.7], where $p(\theta)$ is the prior distribution and $p(y|\theta)$ is the likelihood. This implies that the posterior distribution is proportional to the likelihood $p(y|\theta)$ multiplied by the prior distribution $p(\theta)$.

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)} \propto p(y|\theta)\, p(\theta). \qquad [5.7]$$

The general idea of the MCMC algorithm is to generate samples from the conditional posterior distribution of all unknown parameters in the model. Then these samples are used to calculate point and interval estimates of the parameters of interest (Metropolis and Ulam (1949)). WinBugs uses three different sampling algorithms to simulate a Markov chain with the correct stationary distribution. I apply Metropolis-Hasting (MH) sampling to fit the scorecards in chapter 4.

The MH algorithm generates values of θ, the parameter of interest, from a proposal distribution and corrects these values so that the draws are actually simulating from the posterior distribution $p(\theta|y)$. The proposal distribution is generally dependent on the last value of θ drawn but independent of all other previous values of θ (Markov property). The method works by generating new values at each time step from the current proposal distribution but only accepting the values if they meet a criterion. In this way the estimates of θ are improved at each time step and the Markov chain reaches its equilibrium or stationary distribution, which by construction is the posterior distribution of interest.

The MH sampling algorithm for an unknown parameter θ is as follows:

1. For each time step t sample a point from the current proposal distribution $p_t(\theta^*|\theta^{(t-1)})$.
2. Calculate the acceptance probability $\alpha_t = \min(1, r_t)$ given the posterior ratio r_t defined as

$$r_t = \frac{p(\theta^*|y)/p_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/p_t(\theta^{(t-1)}|\theta^*)}.$$

3. Accept the new value $\theta = \theta^*$ with probability α_t, otherwise let $\theta^{(t)} = \theta^{(t-1)}$.

The marginal distribution of θ approaches the conditional posterior of interest as the number of iterations increases. WinBugs utilizes a random-walk proposal distribution (normal distribution) centred at the current value of the parameter, $\theta^{(t-1)}$. Ideally, the MH algorithm accepts the candidate in 40% to 50% of the iterations.

5.2.1 Prior distributions

The specification of prior distributions is important in Bayesian statistics since it influences posterior inference. Generally, if there is a strong belief of random-effects distribution in the credit scoring models then it would be possible to determine particular (informative) prior distributions for them. In this case, random-effects are assigned to have very small variances for unknown random-effects which implies precise prior knowledge about their distributions.

The credit scoring models I introduce in this dissertation have never been explored in credit scoring. Accordingly, there is no prior knowledge available about parameters' distributions from previous studies or related work. Therefore, I choose and specify prior distributions based on the information from similar studies on multilevel modelling in sociology and health economics (Browne (2009), Bellanger and Zeynep (2008), Gelman and Hill (2007)).

In the case of the cross-classified scoring models, I assign normal and multivariate normal prior distributions for random-effects within occupations, microenvironments or infrastructures. These prior distributions are non-informative which means that they do not put any restrictions on posterior distributions.

Importantly, the precision of random-effects predictions crucially depends on the choice of proper prior distributions for the scorecard variance parameters. Therefore, in order to make better inferences I estimate the models by specifying two types of prior distributions for the standard deviations ($\sigma_{u_j}^{microenvironment}$, $\sigma_{u_l}^{occupation}$, $\sigma_{u_k}^{infrastructure}$) of the random-effects. The choices are noninformative and weakly informative prior distributions.

Given the two types of prior the credit scorecards are estimated one by one and then I compare the outcome results and random effects predictions.

A noninformative prior on a variance parameter σ means that a prior distribution for it is non-restrictive and allows the data to speak for themselves. There are quite a few authors who considered using noninformative prior distributions in their applied research including a proper uniform density on σ_u (Gelman (2004, 2006)) or inverse Gamma distribution ($p(\sigma_u) \sim inv.Gamma(0.001, 0.001)$) as described by Spiegelhalter et al. (1994, 2003). In the thesis I follow Gelman (2003, 2007) and use a uniform prior distribution on σ_u which has a finite integral near $\sigma_u = 0$. The uniform density on σ_u is equivalent $p(\sigma^2) \propto \sigma_u^{-1}$ giving an inverse-χ^2 density with -1 degrees of freedom. This density can also be interpreted as a limit of the half-t family on σ_u where the scale approaches ∞.

The main benefit of using uniform prior distributions for σ_u is that it implies that the posterior distribution is the same as the likelihood function. Accordingly, the standard deviations of the cross-classifications of credit scorecards in chapter 4 are specified to have independent uniform prior distributions

$$\begin{aligned} \sigma_{u_j}^{microenvironment} &\sim uniform\left(0, \frac{1}{\epsilon}\right), \\ \sigma_{u_l}^{occupation} &\sim uniform\left(0, \tfrac{1}{\epsilon}\right), \\ \sigma_{u_k}^{infrastructure} &\sim uniform\left(0, \tfrac{1}{\epsilon}\right), \end{aligned}$$

where $\frac{1}{\epsilon} = \tau$ is the precision which equals the inverse variance, $\tau = \frac{1}{\sigma^2}$ (for details Spiegelhalter et al. (1997)). The commonly used value for ϵ is 0.01.

The second choice of prior distribution for the standard deviation of random-effects is a weakly informative. This prior distribution is a reasonable alternative to the noninformative prior which implies that some prior knowledge is available. The main advantage of weakly informative priors over noninformative is that the former helps to restrict σ_u from very large values. In addition, assigning weakly informative priors helps to speed up the estimation as the algorithm reaches the convergence faster. I refer to Jakulin and Gelman (2008) for a more detailed description.

In the case of the cross-classified credit scorecards weakly informative prior distributions for the standard deviations $\sigma_u^{microenvironment}, \sigma_u^{occupation}$, $\sigma_u^{infrastructure}$ are assigned to a class of half-t distributions. These distributions are half-Cauchy with scale parameter 25.

In summary, I found that the iteration results for the variance parameters in the case of noninformative and weakly informative priors are similar. However, the MCMC algorithm converges much faster when weakly informative priors are specified. In chapter 4 I report the empirical results for the scorecards which are assigned with weakly informative priors.

5.2.2 Initial values

I fit the cross-classified scorecards in WinBugs (Lunn, Thomas, Best, and Spiegelhalter, D. (2000)). To start a simulation initial values for all stochastic nodes and parameters have to be defined. These values are the starting points for Markov chains which are required in order to start simulating samples from a target posterior distribution. In general, there are two choices for initial values. The first one is to generate starting points randomly. The second choice is to supply initial values. I apply the second alternative because randomly assigned initial values do not work well in the case of complex models with many random effects.

In order to get starting values for the cross-classified credit scoring models I independently estimate three multilevel models for each classification in Stata. These scorecards cluster borrowers within two-level structures within microenvironments, occupations and infrastructures. Then, I predict occupation, microenvironment and infrastructure-specific effects for each scorecard and apply these estimates and predictions as initial values for the chains.

I keep Bayesian MCMC running for a long time in order to obtain reliable results. The first 200 000 iterations are discarded from the estimation as a burn-in sample. Then, the scorecards are run for 500 000 additional iterations.

5.2.3 Convergence check

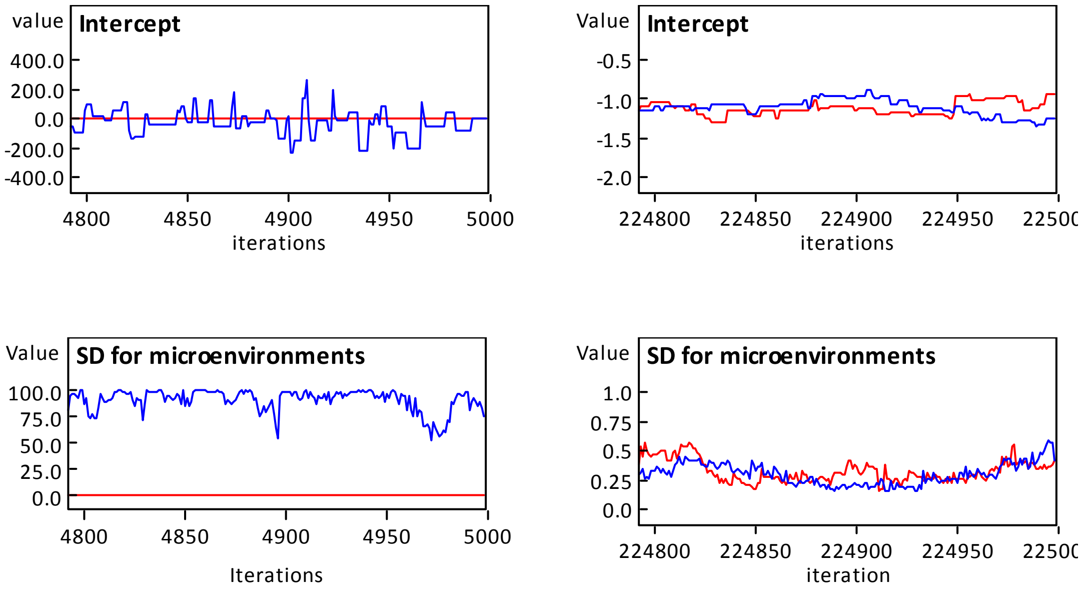
Convergence implies that the MCMC algorithm starting with some initial values for the chains has reached a common equilibrium distribution. The equilibrium distribution is the true posterior distribution of the random-effects. Accordingly, monitoring convergence is essential for obtaining accurate and reliable results. After the model has converged, samples from the posterior distributions are used to summarize the parameters' estimates.

There are many different approaches and rules applied in the literature to check if convergence is reached (Carlin and Cowles (1996)). In this dissertation I apply several convergence diagnostics methods. This subsection provides a short summary and graphically illustrates the results.

The simplest way to check convergence is to monitor the Monte Carlo error. Small values of this error indicate that the parameter of interest is calculated with certain precision. The MC error shows the variability of the estimate due to the simulation and it should be low. According to Geyer (1992) and Carlin and Luis (2000) there are two most common ways to estimate MC error: the batch mean method and the window estimator method. I compute the MC error by applying the batch mean method to the cross-classified credit scoring models.

The batch means method partitions the iteration output sample into K batches (usually K=30). Both the number of batches K and the sample size of each batch $m = \frac{Number\ of iterations}{K}$ must be sufficiently large in order to estimate the variance consistently (Carlin and Louis (2000)). To calculate the MC error of the posterior mean for each parameter I first calculate each batch mean and then the overall sample mean. The MC error is then obtained by finding the standard deviation of the batch means. The batch mean estimator of MC error is discussed in more detail by Hastings (1970), Geyer (1992), Roberts (1996), and Givens and Hoeting (2005).

A second way to check convergence is to examine the trace plots. The trace plots are the plots of iterations versus the generated values. If all values are within a zone without strong periodicities and tendencies, this implies that the convergence is reached. In addition, I run two chains in parallel in order to compare how different chains mix. The chains are assigned to have different initial values. The convergence is reached when the trace lines for different chains mix and cross.



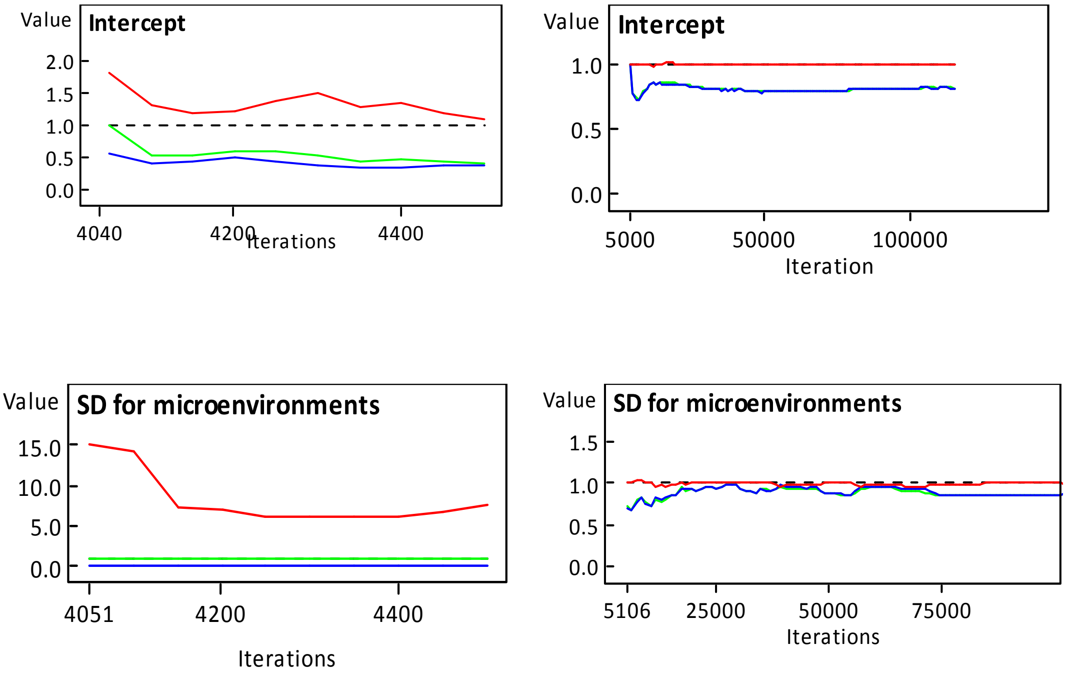
a). *Trace plots for the first 5000 of iterations for the intercept and standard deviation of microenvironment-specific effects. Convergence is not reached.*

b). *Trace plots for the following 220.000 iterations for the intercept and standard deviation of microenvironment-specific effects. Convergence is reached.*

Figure 5.1 *Diagnostics plots: trace plots.*

I choose two parameters from the credit scorecard specified in [5.1] to illustrate the convergence of the MCMC algorithm using the trace plots. Figure 5.1 illustrates the trace plots for the population average intercept and the standard deviation of microenvironment-specific effects. The left hand side plots present the results for the first 5 000 iterations when the chains are far away from convergence. The right hand side plots on Figure 5.1 b) illustrate the case when the algorithm has reached convergence. It is evident, that in this case the chains mix well. Similarly, I assess convergence for the other scorecards' parameters.

The third possibility to monitor convergence is to apply statistical diagnostics tests. The most popular in the literature are Gelman-Rubin test and Raftery-Lewis diagnostics. In this thesis I follow Brooks and Gelman (1998) and apply Gelman-Rubin diagnostics to monitor convergence. Figure 5.2 visualizes Gelman-Rubin diagnostics for the microenvironment-specific intercept and standard deviation of random-effects.



a). *Gelman-Rubin diagnostics for the intercept and standard deviation of random- effects. Convergence is not reached.*

b). *Gelman-Rubin diagnostics for the intercept and standard deviation of microenvironment-specific effects. Convergence is reached.*

Figure 5.2. *Gelman-Rubin diagnostics.*

The basic idea of this diagnostics is to generate multiple chains (m) starting at different initial values. Then, the convergence is assessed by comparing within-chain and between-chain variability over the second half of m chains. The within-chain variance is

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2,$$

where $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\theta_{ij} - \bar{\theta}_j\right)^2$ is the variance of the *j-th* chain and W is the mean of the variances of each chain.

The between chain variance is the variance of the chain means multiplied by n because each chain is based on n iterations:

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\overline{\theta_j} - \bar{\bar{\theta}}\right)^2,$$

where $\bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m} \overline{\theta_J}$ is the mean over all chains. The between and within chain variances are applied to compute the variance of the stationary distribution as a weighted average:

$$\widehat{Var}(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B.$$

Gelman-Rubin diagnostics $\hat{R} = \sqrt{\frac{\widehat{Var}(\theta)}{W}}$ is a potential scale reduction factor which should tend to 1 as convergence is approached. It is assumed that the convergence is reached if $1<R < 1.05$. On the graphs $\hat{R}$ is indicated by the red line.

Figure 5.2 a) shows the situation when the algorithm is far away from convergence. The red line for R does not fall within the interval [1; 1.05]. Figure 5.2 b) visualizes the situation when the convergence is reached.

In summary, this chapter provides a basic overview of the estimation approaches applied to fit the credit scoring models. I apply maximum likelihood to the hierarchical scorecards in chapter 2 and Bayesian MCMC to the cross-classified scoring models in chapter 4. The convergence of the Monte Carlo algorithm is checked by using several techniques including Gelman-Rubin diagnostics.

Conclusion

This dissertation introduces a new type of credit scoring model which specifies a multilevel structure to the data. It is shown that the multilevel scorecards outperform conventional scoring models and can be considered as improved alternatives to the standard scoring techniques. Similarly to the logistic or probit regression a multilevel scoring model assesses credit worthiness of applicants for a loan by forecasting their probability of default. In addition, this thesis proposes a new way of data clustering for a multilevel structure which is more intuitive and relevant for more efficient credit scoring. I introduce different specifications of the multilevel scorecards which are developed using hierarchical and non-hierarchical data structures. These scorecards vary by the degree of complexity and are designed to answer different questions in application credit scoring. The main goal in credit scoring is to define factors which influence riskiness of individuals who apply for a bank loan. In this case the multilevel structure is advantageous because it allows accounting for unobserved characteristics which impact credit riskiness of borrowers additionally to the observed characteristics such as income, marital status or credit history. Including unobserved determinants of default in a credit scoring model helps to increase predictive accuracy and improves a model's performance.

Hierarchical credit scorecards are assigned with a two-level structure. This structure treats borrowers as level-one units which are nested within level-two units – microenvironments. Each microenvironment represents a living area of a customer with a particular combination of socio-economic and demographic conditions. The empirical results confirm that microenvironment-specific effects are heterogeneous across residence areas with dissimilar economic conditions. These effects are random in the model. They capture the impact of area-specific determinants of credit riskiness additionally to the observed personal characteristics.

Importantly, clustering within microenvironments differs from a simple geographical grouping. The main advantage of the former structure is that microenvironments are allowed to include individuals from different cities or regions if their living area conditions are essentially the same. This implies that area-specific conditions influence probability of default but not a geographical location itself. Geographical grouping can be misleading if living areas are similar in terms of socio-economic conditions but have different locations.

The second type of multilevel structure this dissertation applies to a credit scoring model is a non-hierarchical structure. It nests applicants within different classifications according to the similarities in particular characteristics of their occupational activities, living area conditions and infrastructure of shopping facilities in their residence areas. Specifying a cross-classified structure to the credit history data allows exploring the impact of occupation-specific, infrastructure-specific and area-specific characteristics on the riskiness and significantly improves discriminatory power of the scorecards.

Empirical part of the thesis applies maximum likelihood and Bayesian Markov chain Monte Carlo to estimate various specifications of the credit scoring models. After estimation I use a ROC curve analysis in order to assess predictive accuracy of the scoring regressions and evaluate models' performance at the particular cut-off points for probability of default. In addition to the standard ROC curve metrics, several other measures of classification performance are calculated. These measures include a partial ROC area, Gini coefficient, accuracy ratio, correct classification rates and forecasting accuracy scores (Brier, logarithmic and spherical). A partial area under the ROC curve assesses discrimination quality of the scorecards over a region of the ROC curve between two cut-off points. I perform a cost-benefit analysis in order to account for the asymmetric costs associated with falsely predicted positive and negative outcomes.

Chapter 3 concludes the presentation estimation results for the multilevel scorecards with a two-level hierarchical structure. In addition, it compares the multilevel scoring regressions with the logistic scorecard and with the bivariate probit model discussed by W. Greene (1992). The comparison results confirm that the multilevel scorecards outperform conventional scoring techniques (logistic, probit) and produce more accurate forecasts of probability of default. I check goodness-of-fit of the estimated credit scorecards by applying various information criteria (AIC, BIC and DIC).

Complementary to the general accuracy metrics chapter 4 evaluates and compares classification performance between the cross-classified scorecards by evaluating discriminatory power at optimal threshold, fair cut-off point and the kappa-optimal threshold.

To emphasize the role of the microenvironment-specific, occupation-specific and infrastructure-specific effects I provide the graphical illustration of the fitted models results in chapter 3 and 4. In particular, visualizing second-level residuals for various microenvironments allows clarifying the differences between area-specific random-effects within poor and rich regions. It is investigated that socio-demographic characteristics of microenvironments such as area income or housing wealth have a significant impact on probability of default and credit worthiness of borrowers.

Bibliography

Aerts, M. (2002). *Topics in Modelling of Clustered Data*. Chapman and Hall / CRC Press.

Airy, G. (1879). *On the Algebraic and Numerical Theory of Errors of Observations and the Combination of Observations*. Third Ed. MacMillan, London: MacMillan.

Agresti, A. and Zheng, B. (2000). Summarizing the predictive power of a generalized linear model. *Statistics in Medicine,* 19, pp. 1771-1781.

Altman, E. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance,* 23 (4), pp. 589-611.

Altman, E. and Sabato, G. (2005). Effects of the new Basel Capital Accord on bank capital requirements for SMEs. *Journal of Financial Services Research,* 28 (1/3), pp. 15-42.

Anderson, R. (2008). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation.* Oxford University Press.

Anderson, D. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society* B, 47, pp. 203–210.

Andreeva, G., Ansell, J. and Crook, J. (2005). Modelling the purchase propensity: analysis of a revolving store card. *Journal of the Operational Research Society*, Vol. 56, pp. 1041-1050.

Avery, R. and Calem, K. (2004). Consumer credit scoring: do situational circumstances matter? *Journal of Banking and Finance*, Vol. 28, pp. 835-856.

Austin, M. (2007). Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, 200, pp. 1–19.

Baesens, B. and Gastel, T. (2005). Neural network survival analysis for personal loan data. *Journal of the Operational Research Society*, Vol. 56, pp. 1089-1098.

Baesens, B. and Gestel, V. (2005). Linear and nonlinear credit scoring by combining logistic regression and support vector machines. *Journal of Credit Risk*, 1(4).

Balakrishnan, N. (1991) *Handbook of the Logistic Distribution*. Marcel Dekker.

Banasik, J., Thomas, L. and Crook, J. (1999). Not if but when will borrower default. *Journal of the Operational Research Society*, Vol. 50, pp. 1185-1190.

Basel Committee on Banking Supervision (2006). International Convergence of Capital Measurement and Capital Standards. www.bis.org.

Beaver, W. (1967). Financial ratios predictors of failure. *Journal of Accounting Research*, 4,pp. 71-111.

Beauchamp, M., Bray, S., Fielding, A. and Eys, A. (2005). A Multilevel Investigation of the Relationship between Role Ambiguity and Role Efficacy in Sport. *Psychology of Sport and Exercise, 6*, pp. 289-302.

Beling, P. and Oliver, M.(2000). Optimal scoring cut-off policies and efficient frontiers. *Journal of the Operational Research Society*, Vol. 56, pp. 1016-1029.

Bellanger, M. and Zeynep, O.(2008). What can we learn from a cross-country comparison of the costs of child delivery? *Health Economics,* Vol. 17(S1), pp. 47-57.

Bellotti, T. and Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, N. 60, pp. 1699-1707.

Berger, A. and Frame, S. (2007). Small business credit scoring and credit availability. *Journal of Small Business Management,* 45 (1), pp. 5-22.

Blatchford, P., Goldstein, H. and Browne, W. (2002). A study of class size effects in English school reception year classes. *British Educational Research Journal,* Vol. 28, pp. 169-185.

Blöchlinger, A. and Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking and Finance*, 30, pp. 851-873.

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, pp. 1145-1159.

Breslow, N. and Clayton, D.(1993). Approximate inference in Generalized linear mixed models. *Journal of American Statistical Association*, 88, pp. 9-25.

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, 47, pp. 69-100.

Brown, C. and Davis, H. (2006) Receiver operating characteristic curves and related decision measures: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 80, pp. 24–38.

Brown, H. and Prescott, R. (2006). *Applied Mixed Models in Medicine*. Wiley.

Browne, W. (1998). *Applying MCMC methods to multilevel models*. Ph.D thesis, University of Bath.

Brumback, L., Pepe, M. and Alonzo, T. (2006). Using the ROC curve gauging treatment effect in clinical trials. *Statistics in Medicine*, 25, pp. 575-590.

Bryk, A. and Raudenbush, S. (1992). *Hierarchical linear models*, Newbury Park, California.

Burkholder, G. and Harlow, L. (2003). Teacher's corner: An illustration of a longitudinal cross-lagged design for larger structural equation models. *Structural Equation Modelling*, 10(3), pp. 465-486.

Cameron, C. and Trivedi, P. (2005). *Microeconometrics: Methods and Applications.* Cambridge University Press, NY.

Carey, K. (2000). A multilevel modelling approach to analysis of patient costs under managed care. *Health Economics*, Vol. 9(5), pp. 435-446.

Carlin, B. and Cowles, M. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of American Statistical Association,* Vol. 91.

Carlin, B. and Louis, T. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd Ed., London: Chapman and Hall.

Carsten, S., Wesseling, S., Schink, T. and Jung, K. (2003) Comparison of eight computer programs for receiver operating characteristic analysis. *Clinical Chemistry*, 49, pp. 433–439.

Cary, K. (2000). A multilevel modelling approach to analysis of patient costs under managed care. *Health Economics*, Vol. 9, pp. 435-446.

Castermans, G., Martens, Van Gestel, T., Hamers, B. and Baesens, B. (2007). An overview and framework for PD back-testing and benchmarking. *Proceedings of Credit Scoring and Credit Control X*, Edinburgh, Scotland (U.K.).

Chamberlain, G. (1984). *Panel Data.* Handbook of Econometrics, Volume II, ed. by Z. Griliches and M. Intriligator, 1247–1318. Amsterdam: North-Holland.

Chantala, K. and Suchindran, C. (2006) Adjusting for Unequal Selection Probability in Multilevel Models: A Comparison of Software Packages. *Proceedings of the American Statistical Association, Seattle, WA: American Statistical Association.*

Cherlin, A., Chase-Lansdale, P. and McRae C. (1998). Effects of parental divorce on mental health throughout the life course. *American Sociological Review,* 63, pp. 239–249.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* Vol.20, No.1, pp. 37–46.

Cohen, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

Coffin, M. and Sukhatme, S. (1997). Receiver operating characteristics studies and measurement errors. *Biometrics*, 53, pp. 823-837.

Cotter, D., JoAnn, D., Hermsen, J., Kowalewski, B. and Vanneman, R. (1997). All women benefit: the macro-level effect of occupational integration on gender earnings equality. *American Sociological Review,* 62, pp. 714–734.

Corcoran, C., Douglas, G., Pavey, S., Fielding, A., McLinden, M. and McCall, S. (2004). Network 1000: The changing needs and circumstances of visually-impaired people. *British Journal of Visual Impairment, 22,* pp. 93-100.

Crook, J. and Bellotti, T. (2009). Time varying and dynamic models for default risk in consumer loans. *Journal of the Royal statistical Society series A (Statistics in Society)*, accepted for publication, available online 2009.

Cowles, M. and Carlin, B. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association, N.* 91, pp. 883-904.

Dash, D. and Luis, S. (2009). Credit risk evaluation: the application of scorecard in financial services. *International Journal of Financial Services Management,* Vol.4, N. 1, pp. 38-47.

Deakin, E. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 10 (1), pp. 167-179.

Dey, S. and Mumy, G. (2005). Determinants of borrowing limits on credit cards. *Bank of Canada working paper, Number* 7.

Diez-Roux, A. (2000). Multilevel Analysis in Public Health Research. *Annual Reviews of Public Health,* Vol. 21, pp. 171-192.

Drummond, C. and Holte, R. (2000). Explicitly representing expected cost: an alternative to roc representation. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 198–207.

Duncan, S. and Strycker, L.(2002). A multilevel analysis of neighbourhood context and youth alcohol and drug problems. *Prevention Science*, 3, pp.125-134.

Duncan, S., Strycker, L., Duncan, T. and Okut, H. (2002). A multilevel contextual model of family conflict and deviance. *Journal of Psychopathology and Behavioural Assessment*, Vol. 24, No 3, pp. 169-175.

Edelman, D. and Crook, J. (2002). *Credit scoring and its applications*. SIAM: Society of industrial and applied mathematics, Philadelphia.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics,* Vol. 3, pp. 1-21.

Engelman, B., Hayden, E. and Tasche, D. (2003). Testing Rating Accuracy. *Risk*, N. 1, pp. 82-86.

Ferri, C., Flach, P., Hernández-Orallo, J. and Senad, A. (2005). Modifying ROC curves to incorporate predicted probabilities. *Proceedings of the 2nd workshop on ROC analysis in machine learning ROCML.* Bonn, Germany.

Fielding, A., Yang, M. and Goldstein, H. (2003). Multilevel ordinal models for examination grades. *Statistical Modelling,* Vol. 3, pp. 127-153.

Fielding, A. (2003). Ordered Category Responses and Random Effects in Multilevel and Other Complex Structure. In N. Duane and S. Reise (Eds.). *Multilevel Modelling: Methodological advances, issues and applications,* Chapter 9, pp. 181-208. Mahwah NJ, Erlbaum.

Fielding, A. (2004a). Scaling for Residual Variance Components of Ordered Category Responses in Generalised Linear Mixed Multilevel Models. *Quality and Quantity, The European Journal of Methodology, 38, 4, pp.* 425-433.

Fielding, A. (2004b). The Role of the Hausman Test and Whether Higher Level Effects Should be Treated as Random or Fixed. *Multilevel Modelling Newsletter,* 16, 2, pp. 3-9.

Fielding, A. and Hughes, N. (2004). Indicator Data and Targeting Groups and Units in` Implementing Preventative Measures. *National Evaluation of the Children's Fund, Online conference on understanding prevention: children, Families and Social Inclusion.* Published on the web: www.necf.org.uk/conferences/.

Firmstone, V., Bullock, A., Fielding, A., Frame, J., Gibson, C. and Hall, J. (2004). The Impact of Course Attendance on the Practice of Dentists. *British Dental Journal,* 196, *2,* pp. 773-777.

Fisher, R. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, Vol. 52, pp. 399–433.

Freese, J. and Scott Long, J. (2006). *Regression Models for Categorical Dependent Variables Using Stata.* College Station: Stata Press.

Gamoran, A. (1992). The variable effects of high school tracking. *American Sociological Review,* 57, pp. 812–828.

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, pp. 398-409.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

Gelman, A., Carlin, J., Rubin, S. and Donald, B. (2004). *Bayesian Data Analysis.* Second edition. Boca Raton, Florida: Chapman & Hall/CRC. pp. 182–184.

Gelman, A., Brown, C., Carlin, J. and Wolfe, R. (2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics*, 2, pp. 397-416.

Geyer, C.(1992).Practical Markov chain Monte Carlo. *Statistical Science*, 4, 473-482.

Givens, G. and Hoeting, J. (2005). *Computational Statistics.* Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ.

Goldberger, A. (1963). Best linear unbiased prediction in the generalised linear regression model. *American Statistical Association*, 57, pp. 369-375.

Goldstein, H. (2003). *Multilevel statistical models*. London: Arnold, Third edition.

Goldstein H., Rasbash J. and Browne, W. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, 1, pp.223-231.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, Series A, pp. 505-513.

Gonen, M. (2007). *Analyzing Receiver Operating Characteristic Curves Using SAS.* SAS Press, ISBN 978-1-59994-298-1.

Graham, S., Singer, J. and Willett, J. (2008). An Introduction to the Multilevel Model for Change. In P. Alasuutar, L. Bickman and J. Brannen (Eds.) *Handbook of Social Research Methods*. Newbury Park, CA: Sage, pp. 377-394.

Greene, W. (1992). A statistical model for credit scoring. *NYU working paper, EC-92-29.*

Green, W. (2003) *Econometric Analysis*, fifth edition, Prentice Hall, ISBN 0130661899.

Gunter, H., Rayner, S., Thomas, H., Fielding, A., Butt, G. and Lance, A. (2005). Teachers, Time and Work: Findings from the Evaluation of the Transforming the School Workforce Pathfinder Project. *School Leadership and Management, 5,* pp. 441-454.

Guang, G. and Hongxin, Z. (2000). Multilevel modelling for binary data. *Annual Review of Sociology*, 26, pp.441-462.

Gwet, K. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, Vol. 61, pp. 29-48.

Hamilton, R. and Khan, M. (2001). Revolving credit card holders: who are they and how can they be identified? *The Service Industries Journal*, Vol. 21, No. 3, pp. 37-48.

Hand, D. and Henley, W. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of Royal Statistical Society*, Vol. 160, pp. 523-541.

Hanley, J. and McNeil B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, pp. 29-36.

Hanley, J. and McNeil B. (1983). A method of comparing the areas under receiving operating characteristic curves derived from the same cases. *Radiology*, 148, pp. 839-843.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, pp. 97-109.

Heath, A., Yang, M. and Goldstein, H. (1996). Multilevel analysis of the changing relationship between class and party in Britain 1964-1992. *Quality and Quantity*, 30, pp. 389-404.

Henderson, C. (1953). Estimation of variance component with an exact confidence coefficient. *Mathematical Statistics*, 32, pp. 466-476.

Hosmer, D. and Lemeshow, S. (2000). *Applied logistic regression.* Wiley Series in Probability Statistics: New York.

Hsiao, C. (2007). *Analysis of Panel Data*. 2-d Ed. Cambridge University Press.

Huisman, M. and Snijders, T. (2003). Statistical analysis of longitudinal network data with changing composition. *Sociological Methods & Research*, 32, pp. 253-287.

Jakulin, A., Yu-Sung Su and Gelman, A. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, Vol. 2, N 4, pp. 1360-1383.

Keiley, M., Martin, N., Singer, J. and Willett, J. (2008). Discrete-time survival analysis: Predicting whether, and if so when, an event occurs. In S. Menard (Ed.) *Handbook of Longitudinal Research*, pp. 441-463. Newbury Park, CA: Sage.

Khudnitskaya, A. (2010). Microenvironment-specific Effects in the Application Credit Scoring Model. *China-USA Business Review*, Vol. 9, issue 7.

Khudnitskaya, A. (2009). Adverse selection in credit scoring through the prism of hierarchical multilevel modelling. *Proceedings book of Spring Meeting of Young Economists*, pp. 121-125, Beta: ISBN 978-605-377-049-7.

Koskinen, J. and Snijders, T. (2007). Bayesian inference for dynamic social network data. *Journal of Statistical Planning and Inference*, 137, pp. 3930-3938.

Krämer, W. (2009). *Wie schreibe ich eine Seminar- order Examensarbeit*? Campus Verlag, Frankfurt/Main.

Krämer, W. and Bücker, M.(2009). Statistischer Qualitätsvergleich von Kreditausfallprognosen. *Diskussionspapier 30/2009, Sonderforschungsbereich 823.* Available online: https://eldorado.tu-dortmund.de/handle/2003/26528.

Krämer, W. and Güttler, A. (2008). On comparing the accuracy of default predictions in the rating industry. *Empirical Economics*, Springer, Vol. 34(2), pp. 343-356.

Kreft, I., de Leeuw, J. (1998). *Introducing multilevel modelling*. London: Sage.

Lasko, T. and Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, 38(5), pp. 404–415.

Liang, X., Fuller, B. and Singer, J. (2000). Ethnic differences in child-care center selection: The influence of family structure, parental practices, and home language. *Early Childhood Research Quarterly*, 15(3), pp. 357-384.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics,* 38, pp. 963–374.

Lauritsen, J. (1998). The age-crime debate: assessing the limits of longitudinal self-report data. *Social Forces*, 77, pp. 127–155.

Liu, H., and Wu, T. (2003). Estimating the area under a receiver operating characteristic (ROC) curve for repeated measures design. *Journal of Statistical Software* 12, pp. 1-18.

Liu, K. and Lai, K. (2009). Dynamic credit scoring on consumer behaviour using fuzzy Markov model. *Fourth International Multi-Conference on Computing in the Global Information Technology.*

Lobo, J. and Real, R. (2006). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, Volume 17(2), pp. 145-151.

Longford, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika,* 74, pp. 817-827.

Lori, D. and Pepe, S. (2003). Partial AUC Estimation and Regression. *UW Biostatistics Working Paper Series.* 181, http://www.bepress.com/uwbiostat/paper181.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBugs - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, pp. 325-337.

Maas, M. and Snijders, T. (2003). The multilevel approach to repeated measures for complete and incomplete data. *Quality and Quantity*, 37, pp. 71-89.

Manca, A. and Rice, N. (2005). Assessing generalisability by location in trial-based cost-effectiveness analysis: the use of multilevel models. *Health Economics*, Vol. 14(5), pp. 471-485.

MacKay, D. (2003). An Example Inference Task: Clustering. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press. pp. 284–292.

MacQueen, J. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press. pp. 281–297.

Mason, W., Wong, G., and Entwistle, B. (1983). Contextual analysis through the multilevel linear model. *Sociological Methodology,* 13, pp. 72–103.

McCullagh, P., and Nelder, J.(1997).*Generalized linear models*, London: Chapman&Hall.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour. *Frontiers in Econometrics*, pp. 105-142. Academic Press, New York.

McLeay, S. and A. Omar. (2000). The sensitivity of prediction models to the nonnormality of bounded an unbounded financial ratios. *British Accounting Review*, 32, pp. 213-230.

Mousquès, J. and Renaud, T.(2008). A refutation of the practice style hypothesis: the case of antibiotics prescription by French general practitioners for acute rhinopharyngitis. *Working Papers DT18,* Institute for Research and Information in Health Economics.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18 (1), pp. 109-131.

Olsen , K. and Street, A. (2008). The analysis of efficiency among a small number of organizations: How inferences can be improved by exploiting patient-level data. *Health Economics*, Vol. 17(6), pp. 671-681.

Pepe, M. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford.

Rabe-Hesketh, S., Skrondal, A. (2008). *Multilevel and Longitudinal Modelling using STATA*. College Station, TX: *Stata* Press.

Pfefferman, D., Skinner, C., Holmes D., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the royal Statistics Society, B*, 60, pp. 123-140.

Polson, N. (1996). Convergence of Markov Chain Monte Carlo algorithms. In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (eds), *Proceedings of the Fifth Valencia International Conference on Bayesian Statistics*, pp. 297-323, Oxford University Press, Oxford.

Rabe-Hesketh, S., Skrondal A. and Pickles A. (2004). Generalized multilevel structural equation modelling. *Psychometrika*, 69, pp.167-190.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2001). Generalized multilevel parameterization of multivariate random effects models for categorical data. *Biometrics*, 57, pp.1256–1264.

Raeder, K., Siegmund, U., Grittner, U., Dassen, T. and Heinze, C. (2010). The use of fall prevention guidelines in German hospitals - a multilevel analysis. *Journal of Evaluation in Clinical Practice*, Vol. 16 Issue 3**,** pp. 464 – 469.

Raudenbush, S. and Bryk, A. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition*. Newbury Park, CA: Sage.

Robert, C. (1995). Convergence control methods for Markov chain Monte Carlo algorithms. *Statistical Science*, 10, pp. 231-253.

Roby, T. (1965). *Belief States: A Preliminary Empirical Study*. Decision Science Laboratory, L.G. Hascom Field.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society*, Series A 158, pp. 73-89.

Rodriguez, G. and Elo, I. (2003). Intra-class correlation in random-effects models for binary data. *The Stata Journal*, Vol. 3(1), pp. 32-46.

Sabato, G. (2008). *Managing credit risk for retail low-default portfolios. Credit Risk: Models, Derivatives and Management*. N. Wagner Edition. Chapman & Hall. CRC Financial Mathematics Series.

Sampson, R. (1991). Linking the micro- and macro-level dimensions of community social organization. *Social Forces*, 70, pp. 43–64.

Seltzer, M. (1993). Sensitivity analysis for fixed effects in the hierarchical model: a Gibbs sampling approach. *Journal of Educational Statistics*, 18, pp. 207-235.

Seltzer, M., Wong, H. and Bryk, A. (1996). Bayesian analysis in applications of hierarchical models: issues and methods. *Journal of educational and behavioural statistics*, 21, pp. 131-167.

Sim, J. and Wright, C. (2005). The Kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, Vol. 85, pp. 257—268.

Singer, J., Fuller, B., Keiley, M. and Wolf, A. (1998). Early Child Care Selection: Variation by Geographic Location, Maternal Characteristics, and Family Structure. *Developmental Psychology*, 34(5), pp. 1129-1144.

Skinner, C. (2005). The use of survey weights in multilevel modelling. Presented at the *Workshop on Latent Variable Models and Survey Data for Social Science research*. Montreal, Canada.

Skrondal, A. and Rabe-Hesketh, S. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal*, 2 (1), pp. 1-21.

Snijders, T. and Berkhof, J. (2008). Diagnostic checks for multilevel models. In J. de Leeuw (eds.), *Handbook of Multilevel Analysis*. Springer, pp. 141-175.

Snijders, T. (2006). Multi-level event history analysis for a sibling design: The choice of predictor variables. In F. Yammarino and F. Dansereau (eds.). *Research in Multi-level issues*. Vol. 5. *Multi-level issues in social systems*, pp. 243-251.

Snijders, T. and Baerveldt, C. (2003). A Multilevel Network Study of the Effects of Delinquent Behaviour on friendship evolution. *Journal of Mathematical Sociology*, 27, pp. 123-151.

Smeeton, N. (1985). Early History of the Kappa Statistic. *Biometrics,* Vol. 41, p.795.

Spiegelhalter, D., Best, N. and Carlin, B. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society,* Series B, 64(4), pp. 583-616.

Staw, B., Sandelands, L. and Dutton, J. (1981). Threat Rigidity Effects in Organizational Behaviour: A Multilevel Analysis. *Administrative Science Quarterly*, Vol. 26, No. 4, pp. 501-524.

Steele, F. and Jenkins, A. (2007). The effect of school resources on pupil attainment: a multilevel simultaneous equation modelling approach. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 170(3), pp. 801-824.

Strenio, J. Weisberg, H., and Bryk, A. (1983). Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics,* 39, pp. 71-86.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. J*ournal of the American Statistical Association*, 82, pp. 528-549.

Termansen, M., McClean, C., and Preston, C. (2006). The use of genetic algorithms and Bayesian classification to model species distributions. *Ecological Modelling*, 192, pp. 410–424.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, 22, pp. 1701-1762.

Tipett, L. (1931). *The Methods of Statistics*. 1-st Ed. London: 1illiams and Norgate.

Thomas, L. and Scherer, W. (2001). Time will tell: behavioural scoring and the dynamics of consumer credit assessment. *IMA Journal of Management Mathematics* 12, pp. 89-103.

Verbeke, G. and Molenberghs, G.(2006). *Models for Discrete Longitudinal Data.* Berlin: Springer.

Walter, S. (2005). The partial area under the summary ROC curve. *Journal of Statistics in Medicine,* 24(13), pp. 2025-2040.

Wong, G. and Mason, W. (1985). The hierarchical logistic regression model for multilevel analysis. Jo*urnal of American Statistical Association,* 80, pp. 513–523.

Xie, Y. and Hannum, E. (1996). Regional variation in earnings inequality in reform-era urban China. *American Journal of Sociology,* 101, pp. 950–992.

Zou, K. and Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 6, 115(5), pp. 654–657.

Appendix I

Default equation is a latent regression: $D_i^* = \beta' x_i + \varepsilon_i$, where dependent variable might be indentified with the 'propensity to default' and x_i are the explanatory variables. If D_i^* is sufficiently large relative to the attributes, that is, if the individual is in trouble enough, they default. Formaly,

$D_i = 1$ if $D_i^* \geq 0$ and 0 otherwise.

The probabilty of default given variables is

$P_i = Prob[D_i|x_i]$.

Assuming that ε_i is normally distributed with mean zero and variance 1, the default probahilty is

$P_i = Prob[D_i^* \geq 0\ |x_i] = Prob[\varepsilon_i \leq \beta' x_i|x_i] = \Phi(\beta' x_i)$,

where $\Phi(\beta' x_i)$ is the standard normal cumulative distribution function. The classification rule is:

Predict $D_i = 1$ if $\Phi(\beta' x_i) > P^*$,

where P^* is a threshold value chosen by the analist.

The quantity ultimately of interest here is the probabilty of default that would apply , if the individual were issued a credit, which is denoted by $Prob[D = 1\ |C = 1, x]$, where C=1 means an application is accepted and C=0 means it is rejected.

The default probability model that accoubts for the sample selection is constructed using bivariate probit regression. The structural equations are

Default equation: $D_i = \beta' x_i + \varepsilon_i$, $D_i = 1$ if $D_i^* \geq 0$, and 0 otherwise.

Acceptance equation: $C_i^* = \gamma' v_i + u_i$, $C_i = 1$ if $C_i^* > 0$, and 0 otherwise.

D_i and x_i are only observed if $C_i = 1$

C_i and v_i are observed for all applicants.

Selectivity: $[\varepsilon_i u_i] \sim N[\,0, 0, 1, \rho_{\varepsilon u}]$.

The vector of explanatory variables, v_i, are the factors used in the approval decision. The probabilty of interest is the probability of default given that a loan is accepted, which is

$$Prob[D_i = 1 | C_i = 1] = \frac{\Phi_2(\beta' x_i, \gamma' v_i, \rho)}{\Phi(\gamma' v_i)},$$

where Φ is the bivariate normal cumulative probabilty. If $\rho = 0$, the selection is of no consiquence, and the unconditional model of probabilty is appropriate.

Estimated acceptance equation joint with probahilty of default is given in Table A1.

Table A1. Probit model with sample selection

Number of obs = 13444:
Censored/Uncensored obs = 2945/ 10499
Log likelihood = -7312.57
Wald chi2(23) = 401.73
Prob > chi2 = 0.0000

Default equation (conditional)

	Coefficient	*Std. Err.*	*z*	*P>z*	*[95% Conf.interval]*	
Age	-0.0080	0.0033	-2.44	0.0150	-0.0144	-0.0016
Acadmos	0.0007	0.0004	1.86	0.0640	0.0000	0.0015
Adepcnt	0.0378	0.0269	1.40	0.1610	-0.0150	0.0906
Aempmos	0.0007	0.0004	1.76	0.0780	-0.0001	0.0014
Majordrg	-0.1451	0.0522	-2.78	0.0050	-0.2474	-0.0427
Minordrg	0.1105	0.0360	3.07	0.0020	0.0400	0.1810
Ownrent	-0.0167	0.0544	-0.31	0.7590	-0.1234	0.0900
Apadmos	0.0005	0.0003	1.95	0.0520	0.0000	0.0010
Amamind	-0.0071	0.0921	-0.08	0.9380	-0.1877	0.1735
Income	-0.0093	0.0025	-3.68	<0.001	-0.0143	-0.0043
Selfempl7	-0.0766	0.1073	-0.71	0.4760	-0.2869	0.1338
Tradacct	0.0160	0.0052	3.08	0.0020	0.0058	0.0262

Incper	0.0011	0.0028	0.39	0.696	-0.0044	0.0067
Exp_inc	-0.4021	0.2307	-1.74	0.0810	-0.8543	0.0502
Cptopnb	0.0105	0.0075	1.41	0.1590	-0.0041	0.0251
Cptopng	-0.1062	0.0171	-6.21	<0.0001	-0.1396	-0.0727
Cpt30c	0.0978	0.0885	1.11	0.2690	-0.0756	0.2712
Cptf30	0.0392	0.0219	1.79	0.0730	-0.0037	0.0820
Cptavrv	0.0045	0.0026	1.72	0.0860	-0.0006	0.0097
Cburden	0.0024	0.0011	2.19	0.0280	0.0003	0.0045
Constant	-1.4116	0.1252	-11.27	<0.0001	-1.6570	-1.1662
Acceptance equation						
Age	-0.0021	0.0030	-0.70	0.4850	-0.0081	0.0038
Acadmos	0.0018	0.0005	3.76	<0.0001	0.0009	0.0028
Adepcnt	-0.0393	0.0284	-1.38	0.1660	-0.0949	0.0164
Aempmos	-0.0002	0.0004	-0.54	0.5890	-0.0010	0.0006
Majordrg	-0.7427	0.0361	-20.55	<0.0001	-0.8135	-0.6718
Minordrg	-0.0104	0.0376	-0.28	0.7820	-0.0841	0.0633
Qwnrent	0.0497	0.0566	0.88	0.3790	-0.0612	0.1606
Apadmos	0.0001	0.0003	0.43	0.6680	-0.0004	0.0006
Amamind	0.1173	0.1120	1.05	0.2950	-0.1022	0.3369
Income	0.0103	0.0030	3.38	0.0010	0.0044	0.0163
Selfempl7	-0.4068	0.0945	-4.30	<0.0001	-0.5920	-0.2215
Tradacct	0.0994	0.0087	11.45	<0.0001	0.0824	0.1164
Incper	0.0019	0.0034	0.56	0.5770	-0.0047	0.0086
Cptopnb	-0.0287	0.0095	-3.01	0.0030	-0.0473	-0.0100
Cptopng	0.0378	0.0185	2.05	0.0400	0.0016	0.0740
Cpt30c	-0.3130	0.0839	-3.73	<0.0001	-0.4775	-0.1485
Cptf30	-0.0898	0.0188	-4.78	<0.0001	-0.1267	-0.0530
Cptavrv	0.0059	0.0033	1.79	0.0730	-0.0005	0.0123
Cburden	-0.0015	0.0007	-2.16	0.0310	-0.0028	-0.0001
Banksav	-0.4709	0.0907	-5.19	<0.0001	-0.6486	-0.2931
Bankboth	0.4658	0.0474	9.8200	<0.0001	0.3728	0.5587
Credmajr	0.3147	0.0489	6.4400	<0.0001	0.2189	0.4105
Acbinq	-0.1647	0.0109	-15.01	<0.0001	-0.1863	-0.1432
Constant	-1.1215	0.1198	-9.35	<0.0001	-1.3565	-0.8865
/athrho	0.5919271	0.07961	7.43	<0.0001	0.4358	0.7479
rho	0.5312802	0.05713			0.4102	0.6333

**Wald test of indep. eqns. (rho = 0): chi2(1) =   55.27   Prob > chi2 = 0.0000

APPENDIX II

	Default	Add.	Sav	Banking	Age	Other	Depepn	Profes	Milita	Cleric	Sales	Selfempl
		Income	acc	sav+chec		credit	dents	sional	ry	al		oyed
Default	1.000											
Additional Income	0.013	1.000										
Savings account	0.035	-0.033	1.000									
Banking savings and checking	-0.044	-0.015	-0.229	1.000								
Age	-0.059	-0.020	-0.043	0.006	1.000							
Other credit	-0.041	-0.002	-0.020	0.005	-0.007	1.000						
Dependents	-0.006	0.003	-0.036	0.005	0.257	-0.002	1.000					
Professional	-0.041	-0.010	-0.010	-0.008	-0.039	0.020	-0.073	1.000				
Military	0.043	0.060	-0.001	0.028	-0.067	-0.001	0.031	-0.053	1.000			
Clerical	0.037	0.013	0.038	0.022	-0.050	-0.027	-0.100	-0.111	-0.041	1.000		
Sales	-0.008	0.003	-0.018	-0.009	-0.063	0.002	-0.035	-0.113	-0.042	-0.089	1.000	
Selfemployed	-0.011	0.008	-0.020	-0.012	0.122	-0.013	0.055	-0.057	-0.033	-0.047	-0.032	1.000
Major DR	0.023	-0.013	-0.024	0.032	0.099	0.011	0.061	-0.004	-0.028	-0.030	0.015	0.020
Minor DR	0.036	-0.042	-0.029	-0.015	0.093	-0.019	0.078	0.009	-0.027	-0.029	0.011	0.005
Own/rent	-0.063	-0.029	-0.068	0.045	0.394	0.031	0.141	-0.045	-0.069	-0.077	0.002	0.087
Address/months	0.026	-0.073	0.040	0.000	0.000	-0.020	-0.111	-0.010	-0.027	0.009	0.004	0.012
Income	-0.113	-0.003	-0.050	-0.003	0.317	0.099	0.122	0.024	-0.073	-0.154	-0.001	0.146
Trade accounts	-0.069	0.032	-0.053	0.013	0.222	0.110	0.140	-0.025	-0.048	-0.039	0.018	0.033
Open active TA	-0.085	0.089	-0.058	-0.002	0.215	0.113	0.128	0.000	-0.014	-0.046	0.015	0.021
Trade lines	-0.130	0.038	-0.066	0.009	0.261	0.104	0.154	0.008	-0.040	-0.047	0.025	0.019

Delinquencies	0.077	0.006	-0.011	-0.019	0.020	0.003	0.017	0.004	0.007	-0.009	0.015	-0.013
Past due	0.068	-0.023	-0.005	-0.042	0.054	-0.025	0.045	0.018	-0.016	-0.008	0.004	-0.020
Average rev credit	0.019	-0.023	0.003	-0.045	0.093	0.072	0.080	0.005	-0.004	-0.022	-0.006	0.028
Credit burden	0.137	-0.004	0.036	-0.072	-0.111	-0.028	-0.022	-0.018	0.026	0.035	0.028	-0.023
BuyPower Index	-0.010	0.031	0.018	-0.001	-0.037	0.019	-0.097	0.030	-0.026	0.043	0.018	-0.032
Colleage graduates	-0.061	-0.003	-0.026	-0.013	-0.003	0.053	-0.092	0.086	-0.056	-0.034	0.044	-0.005
Med age	-0.013	-0.058	0.022	-0.007	0.020	0.017	-0.054	0.024	-0.106	-0.003	0.012	0.021
Med income	-0.079	0.002	-0.055	0.005	0.024	0.046	0.007	0.034	-0.076	-0.041	0.046	0.001
Housing wealth	-0.039	-0.035	-0.049	0.014	0.073	-0.026	0.121	-0.027	-0.054	-0.052	0.026	0.018
African-American	0.083	-0.061	0.080	0.001	0.032	-0.017	-0.047	0.010	0.011	0.073	-0.050	-0.030
Hispanic	0.060	0.150	0.042	0.001	-0.027	-0.023	-0.024	-0.020	0.000	0.068	-0.042	-0.006

	Major	Minor	Own/	Address/	Income	Trade	Active	Trade	Delin	Past	Aver	Credit
	DR	DR	rent	months		accoun	TA	lines	quenc.	due	rv.cr.	burden
Major DR	1.000											
Minor DR	0.147	1.000										
Own/rent	0.047	0.084	1.000									
Address/months	-0.015	-0.011	-0.031	1.000								
Income	0.111	0.066	0.119	-0.049	1.000							
Trade accounts	0.113	0.151	0.094	-0.049	0.105	1.000						
Open active TA	0.098	0.135	0.067	-0.054	0.135	0.150	1.000					
Trade lines	0.097	0.109	0.291	-0.044	0.157	0.078	0.087	1.000				
Delinquencies	0.090	0.121	0.009	0.004	0.010	0.062	0.076	0.045	1.000			
Past due	0.150	0.098	0.048	0.008	0.023	0.079	0.142	0.140	0.093	1.000		
Average rev credit	0.038	0.043	0.081	-0.003	0.119	0.034	0.046	0.066	0.032	0.026	1.000	

Credit burden	0.022	0.023	-0.135	-0.013	-0.102	-0.087	-0.098	-0.119	0.049	0.042	0.105	1.000
BuyPower Index	0.017	-0.007	-0.148	0.003	0.000	-0.056	-0.024	-0.021	-0.011	-0.003	-0.013	0.003
College graduates	0.011	-0.008	-0.085	-0.030	0.135	-0.019	-0.004	0.037	-0.016	-0.026	0.004	-0.042
Med age	0.002	-0.006	-0.034	0.033	0.044	-0.034	-0.019	-0.016	-0.009	-0.014	0.010	-0.006
Med income	0.012	0.015	0.082	-0.016	0.159	0.038	0.082	0.115	-0.015	-0.015	0.025	-0.073
Housing wealth	0.009	0.032	0.140	0.005	0.078	0.092	0.085	0.096	-0.001	0.013	0.021	-0.041
African-American	0.033	0.019	-0.077	0.042	-0.103	-0.040	-0.027	-0.051	0.041	0.054	-0.006	0.051
Hispanic	0.010	-0.031	-0.136	-0.013	-0.103	-0.032	-0.039	-0.059	0.001	-0.003	-0.016	0.033

	BuyPower Index	Coll grad	Med age	Med income	Hous. wealth	Afr-Amer	His-panic
BuyPower Index	1.000						
College graduates	0.108	1.000					
Med age	0.122	0.155	1.000				
Med income	0.154	0.110	0.152	1.000			
Housing wealth	-0.132	0.017	0.141	0.102	1.000		
African-American	-0.014	-0.153	-0.122	-0.144	-0.102	1.000	
Hispanic	0.127	-0.115	-0.091	-0.159	-0.113	0.090	1.000